Foundations and Trends® in
Communications and Information Theory
Vol. 9, Nos. 2–3 (2012) 113–381

DOI: 10.1561/0100000069

now
the essence of knowledge

# Optimal Resource Allocation in Coordinated Multi-Cell Systems

By Emil Björnson and Eduard Jorswieck

## Contents

**1 Introduction** **115**

1.1 Introduction to Multi-Antenna Communications 116
1.2 System Model: Single-Cell Downlink 119
1.3 Extending Single-Cell Downlink to Multi-Cell Downlink 126
1.4 Multi-Cell Performance Measures and Resource Allocation 139
1.5 Basic Properties of Optimal Resource Allocation 153
1.6 Subjective Solutions to Resource Allocation 161
1.7 Numerical Examples 168
1.8 Summary and Outline 170

**2 Optimal Single-Objective Resource Allocation** **172**

2.1 Introduction to Single-Objective Optimization Theory 173
2.2 Convex Optimization for Resource Allocation 183
2.3 Monotonic Optimization for Resource Allocation 210
2.4 Numerical Illustrations of Computational Complexity 230
2.5 Summary 234

**3 Structure of Optimal Resource Allocation** **236**

3.1 Limiting the Search-Space 237

3.2 Efficient Beamforming Parametrizations 241
3.3 Necessary and Sufficient Pareto Boundary Parametrization 250
3.4 Heuristic Coordinated Beamforming 255
3.5 General Guidelines for Solving Multi-Objective Resource Allocation Problems 271
3.6 Summary 276

**4 Extensions and Generalizations 278**

4.1 Robustness to Channel Uncertainty 279
4.2 Distributed Resource Allocation 294
4.3 Transceiver Impairments 311
4.4 Multi-Cast Transmission 323
4.5 Multi-Carrier Systems 325
4.6 Multi-Antenna Users 328
4.7 Design of Dynamic Cooperation Clusters 332
4.8 Cognitive Radio Systems 341
4.9 Physical Layer Security 346

**Acknowledgments 353**

**Notations and Acronyms 355**

**References 360**

Foundations and Trends® in
Communications and Information Theory
Vol. 9, Nos. 2–3 (2012) 113–381

DOI: 10.1561/0100000069

now
the essence of knowledge

# Optimal Resource Allocation in Coordinated Multi-Cell Systems

**Emil Björnson[1,2] and Eduard Jorswieck[3]**

[1] *KTH Royal Institute of Technology, ACCESS Linnaeus Center, Signal Processing Laboratory, KTH Royal Institute of Technology, Osquldas väg 10, SE-100 44 Stockholm, Sweden*
[2] *Alcatel-Lucent Chair on Flexible Radio, Supélec, Plateau du Moulon, 3 rue Joliot-Curie, 91192, Gif-sur-Yvette cedex, France, emilbjo@kth.se*
[3] *Dresden University of Technology, Communications Theory, Communications Laboratory, Dresden University of Technology, Dresden 01062, Germany, eduard.jorswieck@tu-dresden.de*

## Abstract

The use of multiple antennas at base stations is a key component in the design of cellular communication systems that can meet high-capacity demands in the downlink. Under ideal conditions, the gain of employing multiple antennas is well-recognized: the data throughput increases linearly with the number of transmit antennas if the spatial dimension is utilized to serve many users in parallel. The practical performance of multi-cell systems is, however, limited by a variety of nonidealities, such as insufficient channel knowledge, high computational complexity, heterogeneous user conditions, limited backhaul capacity, transceiver impairments, and the constrained level of coordination between base stations.

This tutorial presents a general framework for modeling different multi-cell scenarios, including clustered joint transmission, coordinated beamforming, interference channels, cognitive radio, and spectrum sharing between operators. The framework enables joint analysis and insights that are both scenario independent and dependent.

The performance of multi-cell systems depends on the resource allocation; that is, how the time, power, frequency, and spatial resources are divided among users. A comprehensive characterization of resource allocation problem categories is provided, along with the signal processing algorithms that solve them. The inherent difficulties are revealed: (a) the overwhelming spatial degrees-of-freedom created by the multitude of transmit antennas; and (b) the fundamental trade-off between maximizing aggregate system throughput and maintaining user fairness. The tutorial provides a pragmatic foundation for resource allocation where the system utility metric can be selected to achieve practical feasibility. The structure of optimal resource allocation is also derived, in terms of beamforming parameterizations and optimal operating points.

This tutorial provides a solid ground and understanding for optimization of practical multi-cell systems, including the impact of the nonidealities mentioned above. The Matlab code is available online for some of the examples and algorithms in this tutorial.

# 1

# Introduction

This section describes a general framework for modeling different types of multi-cell systems and measuring their performance — both in terms of system utility and individual user performance. The framework is based on the concept of dynamic cooperation clusters, which enables unified analysis of everything from interference channels and cognitive radio to cellular networks with global joint transmission. The concept of resource allocation is defined as allocating transmit power among users and spatial directions, while satisfying a set of power constraints that have physical, regulatory, and economic implications. A major complication in resource allocation is the inter-user interference that arises and limits the performance when multiple users are served in parallel. Resource allocation is particularly complex when multiple antennas are employed at each base station. However, the throughput, user satisfaction, and revenue of multi-cell systems can be greatly improved if we understand the nature of multi-cell resource allocation and how to exploit the spatial domain to obtain high spectral efficiencies.

Mathematically, resource allocation corresponds to the selection of a signal correlation matrix for each user. This enables computation of the corresponding signal-to-interference-and-noise ratio (SINR) of

each user. For a given resource allocation, this section describes different ways of measuring the performance experienced by each user and the inherent conflict between maximizing the performance of different users. The performance region and channel gain regions are defined to illustrate this conflict. These regions provide a bridge between user performance and system utility. Resource allocation is then naturally formulated as a multi-objective optimization problem and the boundary of the performance region represents all efficient solutions.

This section formulates the general optimization problem, discusses the different solution strategies taken in later sections, and derives some basic properties of the optimal solution and the performance region. A detailed outline of this tutorial is given at the end of this section. Mathematical proofs are provided throughout the tutorial to facilitate a thorough understanding of multi-cell resource allocation.

## 1.1 Introduction to Multi-Antenna Communications

The purpose of communication is to transfer data between devices through a physical medium called the *channel*. This tutorial focuses on wireless communications, where the data is sent as electromagnetic radio waves propagating through the environment between the devices (e.g., air, building, trees, etc.). The wireless channel distorts the emitted signal, adds interference from other radio signals emitted in the same frequency band, and adds thermal background noise. As the radio frequency spectrum is a global resource used for many things (e.g., cellular/computer networks, radio/television broadcasting, satellite services, and military applications) it is very crowded and spectrum licenses are very expensive, at least in frequency bands suitable for long-range applications. Therefore, wireless communication systems should be designed to use their assigned frequency resources as efficiently as possible, for example, in terms of achieving high *spectral efficiency* (bits/s/Hz) for the system as a whole. This becomes particularly important as cellular networks are transitioning from low-rate voice/messaging services to high-rate low-latency data services. The overall efficiency and user satisfaction can be improved by dynamic allocation and management of the available resources, and service

providers can even share spectrum to further improve their joint spectral efficiency.

The spectral efficiency of a single link (from one transmitter to one receiver) is fundamentally limited by the available transmit power [236], but the spectral efficiency can potentially be improved by allowing many devices to communicate in parallel and thereby contribute to the total spectral efficiency. This approach will however create inter-user interference that could degrade the performance if not properly controlled. As the power of electromagnetic radio waves attenuates with the propagation distance, the traditional way of handling interference is to only allow simultaneous use of the same resource (e.g., frequency band) by spatially well-separated devices. As the radio waves from a single transmit antenna follow a fixed radiation pattern, this calls for division of the landscape into cells and cell sectors. By applying fixed frequency reuse patterns such that adjacent sectors are not utilizing the same resources, interference can be greatly avoided. This near-orthogonal approach to resource allocation is, however, known to be inefficient compared to letting transmitted signals interfere in a controlled way [227].

In contrast to classical resource allocation with single-antenna transmitters [197, 267, 316], modern multi-antenna techniques enable resource allocation with precise spatial separation of users. By steering the data signals toward intended users, it is possible to increase the received signal power (called an array gain) and at the same time limit the interference caused to other non-intended users. The steering is tightly coupled with the concept of *beamforming* in classic array signal processing; that is, transmitting a signal from multiple antennas using different relative amplitudes and phases such that the components add up constructively in desired directions and destructively in undesired directions. Herein, steering basically means to form beams in the directions of users with line-of-sight propagation and to make multipath components add up coherently in the geographical area around non-line-of-sight users. The beamforming resolution depends on the propagation environment and typically improves with the number of transmit antennas [220]. The ability to steer signals toward intended users ideally enables global utilization of all spectral resources, thus

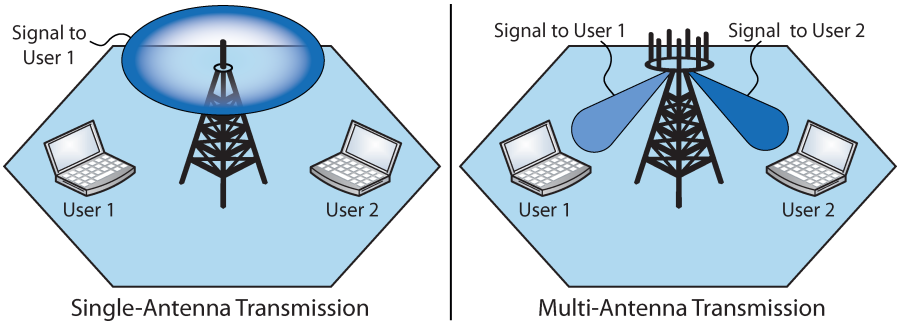



Fig. 1.1 Illustration of the difference between single-antenna and multi-antenna transmission. With a single antenna, the signal propagates according to a fixed antenna pattern (e.g., equally strong in all directions) and can create severe interference in directions where the intended user is not located. For example, interference is caused to User 2 when User 1 is served. With multiple antennas, the signal can be steered toward the intended user which enables simultaneous transmission to multiple spatially separated users with controlled inter-user interference.

removing the need for cell sectoring and fixed frequency reuse patterns; see Figure 1.1. This translates into a much higher spectral efficiency but also more complex implementation constraints — as described later in this section.

The seminal works of [74, 187, 261] provide a mathematical motivation behind multi-antenna communications; the spectral efficiency increases linearly with the number of antennas (if the receiver knows the channel and has at least as many antennas as the transmitter). The initial works considered *point-to-point* communication between two multi-antenna devices — a scenario that is fairly well-understood today [89, 165, 196, 269]. Encouraging results for the *single-cell downlink* where one multi-antenna device transmits to multiple user devices (also known as the broadcast channel) were initially derived in [46, 283]. The information-theoretic capacity region is now fully characterized, even under general conditions [295]. The optimal spectral efficiency is achieved by *nonlinear interference pre-cancelation* techniques, such as dirty paper coding [56]. The single-cell scenario is more challenging than point-to-point since the transmitter needs to know the channel directions of the intended users to perform nonlinear interference precancelation or any sensible linear transmission [84]. Thus, sufficient overhead signaling needs to be allocated for estimation and feedback of channel

information [15, 44, 113]. On the other hand, high spectral efficiency can be achieved in single-cell scenarios while having low-cost single-antenna user devices and non-ideal channel conditions (e.g., high antenna correlation, keyhole-like propagation, and line-of-sight propagation) [84] — this is not possible in point-to-point communication.

The multi-cell downlink has attracted much attention, since the system-wide spectral efficiency can be further improved if the frequency reuse patterns are replaced by cooperation between transmitters. Ideally, this could make the whole network act as one large virtual cell that utilizes all available resources [81]. Such a setup actually exploits the existence of inter-cell interference, by allowing joint transmission from multiple cells to each and every user. Unlike the single-cell scenario, the optimal transmit strategy is unknown even for seemingly simple multi-cell scenarios, such as the *interference channel* where each transmitter serves a single unique user while interference is coordinated across all cells [69, 101, 157, 235]. Part of the explanation is that interference pre-cancelation, which is optimal in the single-cell case, cannot be applied between transmitters in the interference channel. Among the schemes that are suboptimal in the capacity-sense, *linear* transmission is practically appealing due to its low complexity, asymptotic optimality (in certain cases), and robustness to channel uncertainty. The best linear transmission scheme is generally difficult to obtain [157, 168], even in those single-cell scenarios where the capacity region is fully characterized. Recent works have however derived strong parameterizations [16, 180, 235, 325] and these will be described in Section 3.

This tutorial provides theoretical and conceptual insights on the optimization of general multi-cell systems with linear transmission. To this end, the tutorial first introduces a mathematical system model for the single-cell downlink. This model serves as the foundation when moving to the multi-cell downlink, which has many conceptual similarities but also important differences that should be properly addressed.

## 1.2 System Model: Single-Cell Downlink

Consider a single-cell scenario where a base station with $N$ antennas communicates with $K_r$ user devices, as illustrated in Figure 1.2. The

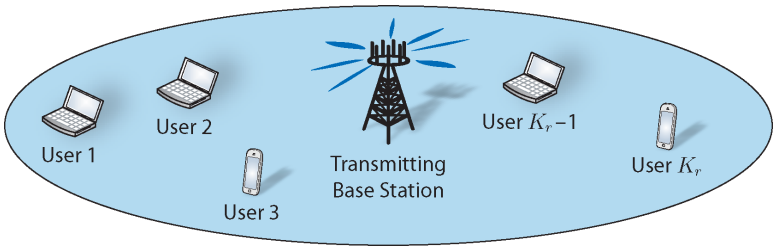



Fig. 1.2 Illustration of the downlink multi-user system in Section 1.2. A base station with $N$ antennas serves $K_r$ users.

$k$th user is denoted $\text{MS}_k$ (the abbreviation stands for mobile station) and is assumed to have a single effective antenna[1]; the case with multiple antennas per user is considered in Section 4.6. This scenario can be viewed as the superposition of several multiple-input single-output (MISO) links, thus it is also known as the *MISO broadcast channel* or *multi-user MISO communication* [46]. It is also frequently described as multi-user MIMO (multiple-input multiple-output) (cf. [84]), referring to that there are $K_r$ receive antennas in total, but we avoid this terminology as it creates confusion.

The channel to $\text{MS}_k$ is assumed to be flat-fading[2] and represented in the complex baseband by the dimensionless vector $\mathbf{h}_k \in \mathbb{C}^N$. The complex-valued element $[\mathbf{h}_k]_n$ describes the channel from the $n$th transmit antenna; its magnitude represents the gain (or rather the attenuation) of the channel, while its argument describes the phase-shift created by the channel. We assume that the channel vector is quasi-static; that is, constant for the duration of many transmission symbols, known as the *coherence time*. The collection of all channel vectors $\{\mathbf{h}_k\}_{k=1}^{K_r}$ is known as the *channel state information (CSI)* and is assumed perfectly known at the base station. We also assume that the transceiver hardware is ideal, without other impairments than can

[1] This means that $\text{MS}_k$ is equipped with either a single antenna or $M_k > 1$ antennas that are combined into a single effective antenna (e.g., using receive combining or antenna selection). There are several reasons for making this assumptions: it enables noniterative transmission design, put less hardware constraints on the user devices, requires less channel knowledge at the transmitter, and is close-to-optimal under realistic conditions [15, 28, 268].

[2] Flat-fading means that the frequency response is flat, which translates into a memoryless channel where the current output signal only depends on the current input signal.

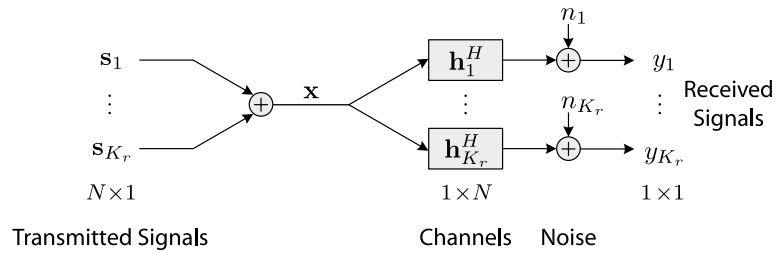



Fig. 1.3 Block diagram of the basic system model for downlink single-cell communications. $K_r$ single-antenna users are served by $N$ antennas.

be included in the channel vector and background noise. These assumptions are idealistic, but simplify the conceptual presentation in this and subsequent sections. It is generally impossible to find perfect models of reality, or as famously noted in [34]: "Remember that all system models are wrong." Therefore, the goal is to formulate a model that enables analysis and at the same time is accurate enough to provide valuable insights. Relaxations to more realistic conditions and assumptions are provided in Section 4.

Under these assumptions, the symbol-sampled complex-baseband received signal at $\mathrm{MS}_k$ is $y_k \in \mathbb{C}$ and is given by the linear input–output model

$$y_k = \mathbf{h}_k^H \mathbf{x} + n_k, \tag{1.1}$$

where $n_k \in \mathbb{C}$ is the combined vector of additive noise and interference from surrounding systems. It is modeled as circularly symmetric complex Gaussian distributed, $n_k \sim \mathcal{CN}(0, \sigma^2)$, where $\sigma^2$ is the noise power. This input–output model is illustrated in Figure 1.3. In a multi-carrier system, for example, based on orthogonal frequency-division multiplexing (OFDM), the input–output model (1.1) could describe one of the subcarriers. For brevity, we concentrate on a single subcarrier in Sections 1–3, while the multi-carrier case is discussed in Section 4.5.

The transmitted signal $\mathbf{x} \in \mathbb{C}^N$ contains data signals intended for each of the users and is given by

$$\mathbf{x} = \sum_{k=1}^{K_r} \mathbf{s}_k, \tag{1.2}$$

where $\mathbf{s}_k \in \mathbb{C}^N$ is the signal intended for $\mathrm{MS}_k$. These stochastic data signals are modeled as zero-mean with *signal correlation matrices*

$$\mathbf{S}_k = \mathbb{E}\{\mathbf{s}_k \mathbf{s}_k^H\} \in \mathbb{C}^{N \times N}. \tag{1.3}$$

This transmission approach is known as linear *multi-stream beamforming* ($\mathrm{rank}(\mathbf{S}_k)$ is the number of streams) and the signal correlation matrices are important design parameters which will be used to optimize the performance/utility of the system.

---

**Definition 1.1.** Each selection of the signal correlation matrices $\mathbf{S}_1, \ldots, \mathbf{S}_{K_r}$ is called a *transmit strategy*. The average transmit power allocated to $\mathrm{MS}_k$ is $\mathrm{tr}(\mathbf{S}_k)$.

---

The only transmit strategies of interest are those that satisfy the power constraints of the system, which are defined next.

### 1.2.1 Power Constraints

The power resources available for transmission need to be limited somehow to model the inherent restrictions of practical systems. The average transmit power $\mathrm{tr}(\mathbf{S}_k)$ and noise power $\sigma^2$ are normally measured in milliwatt [mW], with dBm as the corresponding unit in decibels. We assume that there are $L$ linear power constraints, which are defined as

$$\sum_{k=1}^{K_r} \mathrm{tr}(\mathbf{Q}_{lk} \mathbf{S}_k) \leq q_l \quad l = 1, \ldots, L, \tag{1.4}$$

where $\mathbf{Q}_{lk} \in \mathbb{C}^{N \times N}$ are Hermitian positive semi-definite weighting matrices and the limits $q_l \geq 0$ for all $l, k$. If $\mathbf{Q}_{lk}$ is normalized and dimensionless, then $q_l$ is measured in mW and serves as an upper bound on the allowed transmit power in the subspace spanned by $\mathbf{Q}_{lk}$. To ensure that the power is constrained in all spatial directions, these matrices satisfy $\sum_{l=1}^{L} \mathbf{Q}_{lk} \succ \mathbf{0}_N$ for every $k$. These constraints are given in advance and are based on, for example,

- physical limitations
  (e.g., to protect the dynamic range of power amplifiers);

- regulatory constraints
  (e.g., to limit the radiated power in certain directions);
- interference constraints
  (e.g., to control interference caused to certain users);
- economic decisions
  (e.g., to manage the long-term cost and revenue of running a base station).

Two simple examples are a total power constraint (i.e., $L = 1$ and $\mathbf{Q}_{1k} = \mathbf{I}_N$ for all $k$) and per-antenna constraints (i.e., $L = N$ and $\mathbf{Q}_{lk}$ is only nonzero at the $l$th diagonal element). While these examples can be viewed as two extremes, practical systems are typically limited in both respects.

The matrices $\mathbf{Q}_{lk}$ might be the same for all users, but can also be used to define subspaces where the transmit power should be kept below a certain threshold when transmitting to a specific user (or subset of users). The motivation is, for example, not to disturb neighboring systems and the corresponding constraints are called soft-shaping [107, 230], because the shape of the transmission is only affected if the power without the constraint would have exceeded the threshold $q_l$. For example, if the inter-user interference caused to $\mathrm{MS}_k$ should not exceed $q_l$, then we can set $\mathbf{Q}_{li} = \mathbf{h}_k \mathbf{h}_k^H$ for all $i \neq k$ and $\mathbf{Q}_{lk} = \mathbf{0}_N$. This is relevant both to model so-called zero-forcing transmission (i.e., with zero inter-user interference) and in the area of cognitive radio, where a secondary system is allowed to use licensed spectrum if the interference caused to the system of the licensee is limited.

The $L$ linear sum power constraints introduced in (1.4) can be also decomposed into per-user power constraints as

$$\mathrm{tr}(\mathbf{Q}_{lk}\mathbf{S}_k) \leq q_{lk} \quad k = 1, \ldots, K_r,\ l = 1, \ldots, L, \tag{1.5}$$

for some limits $q_{lk} \geq 0$ for all $l, k$. In order to fulfill (1.4), the per-user power limits need to satisfy the conditions

$$\sum_{k=1}^{K_r} q_{lk} \leq q_l \quad l = 1, \ldots, L. \tag{1.6}$$

This equivalent representation of the $L$ linear sum power constraints is useful to derive structural results on the optimal transmit strategies.

Selecting the limits $q_{lk}$ is part of the performance optimization and basically corresponds to the per-user power allocation.

### 1.2.2 Resource Allocation

The signal correlation matrices are important parameters that shape the transmission and ultimately decide what is received at the different users. Having defined the input–output model in (1.1) and the power constraints in (1.4), we are ready to give a first brief definition of the resource allocation problem considered in this tutorial.

---

**Definition 1.2.** Selecting a transmit strategy $\mathbf{S}_1, \ldots, \mathbf{S}_{K_r}$ in compliance with the power constraints is called *resource allocation*.

---

The selection should be based on some criterion on user satisfaction, which will be properly defined later in Section 1.4. Observe that resource allocation implicitly includes selecting which users to transmit to, the spatial directivity of the signals to selected users, and the power allocation. In principle, $\mathrm{tr}(\mathbf{S}_k)$ describes the power allocated for transmission to $\mathrm{MS}_k$, while the eigenvectors and eigenvalues of $\mathbf{S}_k$ describe the spatial distribution of this power. The rank of $\mathbf{S}_k$ equals the number of simultaneous data streams that are multiplexed to $\mathrm{MS}_k$. The general case when multiple users are served simultaneously is called *spatial division multiple access* (SDMA) [217], while the special case when only one user is allocated nonzero power at a time is known as *time division multiple access* (TDMA). The $N$ transmit antennas can be viewed as having $N$ spatial degrees-of-freedom in the resource allocation, which can be utilized for sending a total of $N$ simultaneous data streams in a controlled manner. The spectral efficiency is not always maximized by sending the maximum number of streams, since this might create much inter-user interference and can be very sensitive to CSI uncertainty — TDMA is the better choice in the absence of CSI [84].

SDMA is the main focus of this tutorial and we assume that there is an infinite queue of data to be sent to each user; thus, all users are available for transmission and are not upper-limited on how high

performance they can achieve. The data is delivered to the base station through a *backhaul network*, which also will be used for base station coordination when we extend the single-cell model into a multi-cell model in Section 1.3.

---

**Remark 1.1 (Basic Channel Modeling).** The analysis in this tutorial is applicable under any channel conditions, noise power, and power constraints. Some intuition on typical system conditions (used in numerical simulations) might however aid the understanding.

The channel vector is often modeled as complex Gaussian, $\mathbf{h}_k \sim \mathcal{CN}(\bar{\mathbf{h}}_k, \mathbf{R}_k)$, where the mean value $\bar{\mathbf{h}}_k \in \mathbb{C}^N$ describes the line-of-sight propagation (if it exists) and the covariance matrix $\mathbf{R}_k \in \mathbb{C}^{N \times N}$ characterizes the varying nature of the channel. This model is called *Rician fading* or *Rayleigh fading* (if $\bar{\mathbf{h}}_k = \mathbf{0}$), since the magnitude of each channel element is Rice or Rayleigh distributed, respectively. Although simple, this model makes sense in rich multipath scenarios (e.g., based on the Lindeberg Central limit theorem [309]) and has been validated by measurements [54, 132, 288, 294, 306]. The spatial directivity is specified by the off-diagonal elements in $\mathbf{R}_k$ and the exponential correlation model in [162] provides a simple parametrization. The channel attenuation depends strongly on the distance between the transmitter and the receiver; this is modeled as $-128.1 - 37.6\log_{10}(d)$ dB in 3GPP Long Term Evolution (LTE) [1], where $d$ is the separation in kilometers. Accordingly, $\frac{\mathrm{tr}(\mathbf{R}_k)}{N}$ lies in the range of $-70$ dB to $-140$ dB in cellular systems. Further reduction are introduced by signal penetration losses, while antenna gains improve the conditions.

The noise power $\sigma^2$ can be modeled as $-174 + 10\log_{10}(b) + n_f$ dBm, where $b$ is the bandwidth in Hertz and $n_f$ is the noise figure caused by hardware components. For example, the noise power is $-127$ dBm for a 15 kHz subcarrier with a 5 dB noise figure. Furthermore, the transmit power (per flat-fading subcarrier) is typically in the range of 0–20 dBm. As the received signal power and the noise power are both very small quantities, normalization is often beneficial in numerical computations.

---

## 1.3 Extending Single-Cell Downlink to Multi-Cell Downlink

In traditional multi-cell systems, each user belongs to one cell at a time and resource allocation is performed unilaterally by its base station. This is enabled by having frequency reuse patterns such that cell sectors utilizing the same resources cause negligible interference to each other. The single-cell system model, defined in the previous section, can therefore be applied directly onto each cell sector — at least if the negligible interference from distant cell sectors is seen as part of the additive background noise. Accordingly, the base station can make autonomous resource allocation decisions and be sure that no uncoordinated interference appears within the cell.

A different story emerges in multi-cell multi-antenna scenarios where all base stations are simultaneously using the same frequency resources (to maximize the system-wide spectral efficiency). The counterpart of SDMA in multi-cell systems have been given many names, including *co-processing* [233], *cooperative processing* [321], *network MIMO* [279], *coordinated multi-point (CoMP)* [202], and *multi-cell processing* [81]. It is based on the same idea of exploiting the spatial dimensions for serving multiple users in parallel while controlling the interference. Network MIMO is particularly important for users that experience channel gains on the same order of magnitude from multiple base stations (e.g., cell edge users). The initial works in [125, 233, 321] assumed perfect co-processing at the base stations and modeled the whole network as one large multi-user MISO system where the transmit antennas happen to be distributed over a large area; all users were served by *joint transmission* from all base stations and the multi-cell characteristics were essentially reduced to just constraining the transmit power per antenna array or antenna, instead of the total transmit power (as traditionally assumed for single-cell systems). The optimal spectral efficiency under these ideal conditions can be obtained from the single-cell literature, in particular [295]. Although mathematically convenient, this approach leads to several implicit assumptions that are hard to justify in practice. First, global CSI and data sharing is required, which puts huge demands on the channel estimation, feedback links, and backhaul networks [122, 174, 175, 200, 247, 312, 313].

Second, coherent joint transmission (including joint interference cancelation) requires very accurate synchronization[3] between base stations [18, 262, 318] and increases the delay spread [322], potentially turning flat-fading channels into frequency-selective. Third, the complexity of centralized resource allocation algorithms is infeasible in terms of computations, delays, and scalability [21]. On the other hand, the early works on the multi-cell downlink provide (unattainable) upper bounds on the practical multi-cell performance.

Various alternative models have been proposed to capture multi-cell-specific characteristics. The CSI requirements were reduced in [191, 114, 246] by using the so-called Wyner model from [299] where interference only comes from immediate neighboring cells; see Example 1.1 for details. This enables relatively simple analysis, but the results can also be oversimplified [300]. Another approach is to divide base stations into *static disjoint cooperation clusters* as in Figure 1.4 [106, 174, 323]. Each cluster is basically operated as a single-cell system.

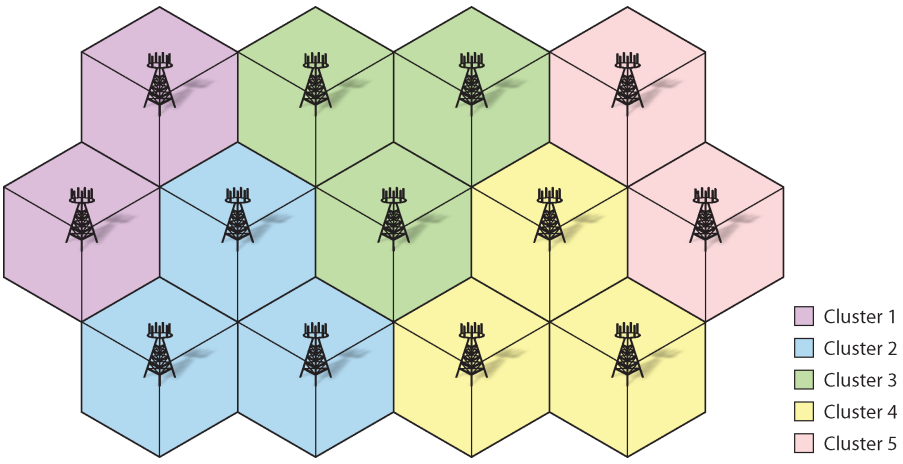



Fig. 1.4 Schematic illustration of static disjoint cooperation clusters.

[3] Synchronization is very important to enable signal contributions from different base stations to cancel out at nonintended users. Precise phase-synchronization can potentially be achieved and maintained by sending a common reference signal to the base stations from a master oscillator [8, 177], using reference clocks that are phase-locked to the GPS [124], or by estimating and feeding back the offset at the users [318].

If the clusters are sufficiently small (e.g., cell sectors connected to the same eNodeB in an LTE system), this approach enables practical channel acquisition, coordination, and synchronization within each cluster. Networks with static clusters unfortunately provide poor spectral efficiency when the user distribution is heterogeneous [173] and suffer from out-of-cluster interference [77]. The impact of these drawbacks can be reduced by having different static disjoint cooperation clusters on different frequency subcarriers [176], by increasing the cluster size and serve each user by a subset of its base stations [33], by having frequency reuse patterns in the cluster edge areas [146], and by changing the disjoint clusters over time [173, 199]. These approaches can however be viewed as treating the symptoms rather than the actual problem, namely the formation of clusters based on a base station-perspective. Steps toward more dynamic and flexible multi-cell coordination were taken in [18, 77, 109, 128, 129, 263] by creating clusters from a user-centric perspective. This means that the set of base stations that serve or reduce interference to a given user is based on the particular needs of this user. Consequently, each base station has its own unique set of users to coordinate interference toward and serves a subset of these users with data. Each base station coordinates its resource allocation decisions with exactly those base stations that affect the same users. This is very different from the disjointness mentioned above, because each base station basically cooperates with all of its neighbors and forms different cooperation clusters when serving different users. The geographical location of a user has a large impact on the clustering [109], but the desirable cooperation and coordination also change with time, for example, based on user activity levels, mobility of users, and macroscopic conditions such as congestion in certain areas. This tutorial considers *dynamic cooperation clusters* of this user-centric type and the framework includes the scenarios described above as special cases.

A seemingly different multi-cell setup arises in the area of *cognitive radio* [90, 102, 230]. Frequency spectrum is traditionally licensed to companies or agencies, which are given exclusive rights for utilization. Therefore, the licensee can unilaterally manage the transmissions and guarantee the service quality for its users. However, a major part of the licensed spectrum is under-utilized today, thus providing the

opportunity for improvements in spectral efficiency [55]. The cognitive radio paradigm is based on having secondary systems that are allowed to use the spectrum if they are not disrupting the primary system (which owns the license). Three ways for the secondary system to achieve this are: interweave (detect and transmit when primary system is inactive), underlay (steer signals away from primary users to avoid interference), and overlay (compensate for the interference caused to primary users by participating in joint transmission of their intended signals). These cognitive radio scenarios can be modeled using the framework of this tutorial (see Section 4.8), and can naturally be extended for spectrum sharing between operators on equal terms.

### 1.3.1 Dynamic Cooperation Clusters

Next, we extend the downlink single-cell system model in Section 1.2 to a multi-cell scenario with $K_t$ base stations. The $j$th base station is denoted $\mathrm{BS}_j$ and is equipped with $N_j$ antennas. The antenna array can have any structure and we assume that $N_j$ is a fixed parameter.[4] Observe that the total number of transmit antennas is still denoted $N = \sum_{j=1}^{K_t} N_j$. Based on the discussion in the previous section and on [18], our general multi-cell system model will embrace the following observations:

- Each user is jointly served by a subset of all base stations;
- Some base stations and users are very far apart, making it impractical to estimate and separate the interference on these channels from the background noise.

Based on these observations, we make the following definition.

[4] The hardware design of antenna arrays has important implications on channel properties such as spatial correlation, mutual antenna coupling, and aperture — all of which are affecting the spatial resolution of beamforming. Release 9 of the LTE standard supports $N_j = 8$ antennas [1], but current research investigates the potential of having much larger arrays (up to several hundred of antennas). We refer to [220] for a recent survey on the challenges and opportunities of having unconventionally large numbers of antennas.

**Definition 1.3.** *Dynamic cooperation clusters (DCC)* means that:

- $\text{BS}_j$ has channel estimates to users in $\mathcal{C}_j \subseteq \{1, \ldots, K_r\}$, while interference generated to users $i \notin \mathcal{C}_j$ is negligible and can be treated as part of the Gaussian background noise;
- $\text{BS}_j$ serves the users in $\mathcal{D}_j \subseteq \mathcal{C}_j$ with data.

This coordination framework is characterized by the sets $\mathcal{C}_j, \mathcal{D}_j \, \forall j$, which are illustrated in Figure 1.5. In this figure, the inner set $\mathcal{D}_j$ contains the users that $\text{BS}_j$ might serve with data. The larger outer set $\mathcal{C}_j$ contains all users that $\text{BS}_j$ should take into consideration and coordinate interference toward. The mnemonic rule is that $\mathcal{D}_j$ describes *data* from $\text{BS}_j$, while $\mathcal{C}_j$ describes *coordination* from $\text{BS}_j$. The membership of users in these sets changes dynamically during operation (e.g., based on individual user locations and the user density in different areas) and it should be noted that each base station may cooperate with different subsets of base stations for each of its users; in other words, the users can generally *not* be divided into disjoint groups served by disjoint groups of base stations.

How to select $\mathcal{C}_j, \mathcal{D}_j$ efficiently is a very important and complex problem [45]. On the one hand, joint transmission and interference coordination provide extra degrees-of-freedom to separate users spatially. This benefit comes, on the other hand, at the cost of spending

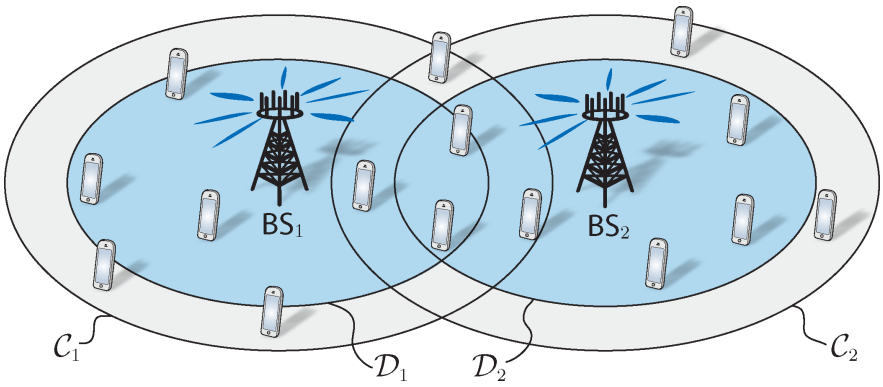



Fig. 1.5 Schematic intersection of two cells. $\text{BS}_j$ serves users in the inner circle ($\mathcal{D}_j$), while coordinating interference to users in the outer circle ($\mathcal{C}_j$). The interference caused to users outside both circles is negligible and included in the respective noise terms.

backhaul and overhead signaling on obtaining CSI, sharing data, and achieving base station synchronization. Increased expenditure is only well motivated if it leads to substantial improvements in spectral efficiency; joint transmission is more costly (it requires data sharing and tight synchronization) than interference coordination, thus we can generally expect $\mathcal{D}_j$ to be a much smaller set than $\mathcal{C}_j$. The clustering problem is discussed in Section 4.7, but for now we assume that the sets $\mathcal{C}_j, \mathcal{D}_j \, \forall j$ are given and known everywhere needed.

The reason for basing the tutorial on DCC is twofold. First, it enables joint analysis of different levels of multi-cell coordination (from the Wyner model or cognitive radio to global joint transmission). Second, it can resolve some of the issues that appear when the multi-cell downlink is viewed as a single-user system with a large distributed transmit antenna array and distributed power constraints. According to Definition 1.3, $\text{BS}_j$ only needs to know its own channel to users that receive non-negligible interference from it — a natural assumption since these are the users for which $\text{BS}_j$ can achieve reliable channel estimates.[5] In addition, only neighboring base stations need to be phase synchronized[6] and joint transmission only creates a small increase in delay-spread (which is easy to handle in OFDM systems by increasing the cyclic prefix [322]).

### 1.3.2 Extended System Model: Multi-Cell Downlink

In the multi-cell scenario, the channel from all base stations to $\text{MS}_k$ is denoted $\mathbf{h}_k = [\mathbf{h}_{1k}^T \dots \mathbf{h}_{K_t k}^T]^T \in \mathbb{C}^N$, where $\mathbf{h}_{jk} \in \mathbb{C}^{N_j}$ is the channel from $\text{BS}_j$. Based on the DCC in Definition 1.3, only certain channel elements of $\mathbf{h}_k$ will carry data and/or non-negligible interference. These can be selected by the diagonal matrices $\mathbf{D}_k \in \mathbb{C}^{N \times N}$ and $\mathbf{C}_k \in \mathbb{C}^{N \times N}$,

[5] There are two main system categories: Frequency division duplex (FDD) and Time division duplex (TDD). The main difference is that each frequency subcarrier in FDD is used for *either* downlink or uplink transmission, while each subcarrier in TDD switches between downlink and uplink transmission. TDD seems particularly useful for multi-cell coordination, because multiple base stations can exploit the same uplink pilot signal to estimate their respective channels (if channel reciprocity can be utilized [96]). The CSI acquisition is more demanding in FDD, since more resources are required for CSI feedback to the additional base stations (and possibly some extra backhaul signaling).

[6] Note that local phase synchronization does not imply global phase synchronization, because small deviations between neighboring base stations are acceptable but can grow into large deviation between distant base stations.

which are defined as

$$\mathbf{D}_k = \begin{bmatrix} \mathbf{D}_{1k} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{D}_{K_t k} \end{bmatrix} \quad \text{where } \mathbf{D}_{jk} = \begin{cases} \mathbf{I}_{N_j}, & \text{if } k \in \mathcal{D}_j, \\ \mathbf{0}_{N_j}, & \text{otherwise,} \end{cases} \tag{1.7}$$

$$\mathbf{C}_k = \begin{bmatrix} \mathbf{C}_{1k} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{C}_{K_t k} \end{bmatrix} \quad \text{where } \mathbf{C}_{jk} = \begin{cases} \mathbf{I}_{N_j}, & \text{if } k \in \mathcal{C}_j, \\ \mathbf{0}_{N_j}, & \text{otherwise.} \end{cases} \tag{1.8}$$

Thus, $\mathbf{h}_k^H \mathbf{D}_k$ is the channel that carries data to $\text{MS}_k$ and $\mathbf{h}_k^H \mathbf{C}_k$ is the channel that carries non-negligible interference.[7] It is necessary to have both $\mathbf{D}_k$ and $\mathbf{C}_k$, to make sure that only the correct base stations transmit to $\text{MS}_k$ when optimizing the resource allocation.

Extending the single-cell input–output model in (1.1), the symbol-sampled complex-baseband received signal at $\text{MS}_k$ is

$$y_k = \mathbf{h}_k^H \mathbf{C}_k \sum_{i=1}^{K_r} \mathbf{D}_i \mathbf{s}_i + n_k \tag{1.9}$$

and is illustrated in Figure 1.6.[8] The additive term $n_k \sim \mathcal{CN}(0, \sigma_k^2)$ is now assumed to model both noise and weak uncoordinated interference from all $\text{BS}_j$ with $k \notin \mathcal{C}_j$ (see Definition 1.3). This assumption limits the amount of CSI required to analyze the transmission and is reasonable if only users that would receive signals that are stronger than the background noise are included in $\mathcal{C}_j$. This might be satisfied if base stations coordinate interference to all cell edge users of adjacent cells (similar to the Wyner model [299]). The variance $\sigma_k^2$ is generally different among the users (representing how weak the uncoordinated interference is at

[7] The antennas that transmit to a certain user can, for simplicity, be thought of as being a single transmitter, although the antennas might belong to different base stations. The reality is however more complex, for example, due to base station-specific power constraints, separate channel acquisition, and distributed resource allocation; see Section 4.

[8] This tutorial considers transmission using linear beamforming over a single subcarrier and channel use. Higher spectral efficiency can potentially be achieved using nonlinear interference pre-subtraction at the base stations (e.g., dirty paper coding [56, 46, 283, 295]) or by extending the transmission over, for instance, a collection of channel realizations (e.g., interference alignment [41]). The truly optimal transmission scheme is unknown for general multi-cell systems, thus the linear beamforming considered in this tutorial should be viewed as a practically appealing transmission approach rather than the overall optimal strategy.

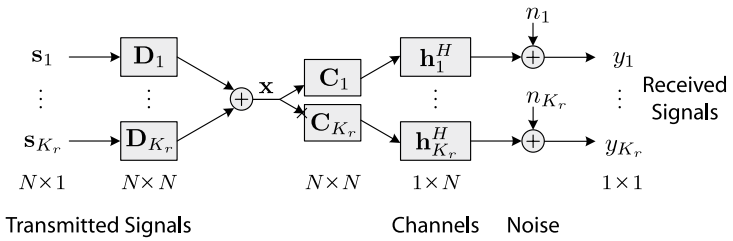



Fig. 1.6 Block diagram of the general system model for downlink multi-cell communications. $K_r$ single-antenna users are served by $N$ antennas.

a certain user) and is estimated and tracked using the received signals.[9] It is worth pointing out that $\sigma_k^2$ is implicitly coupled with the power constraints; if the system-wide power usage is increased, then the uncoordinated interference will also increase. This relationship has no particular impact on this tutorial since our power constraints are fixed, but is of paramount importance in any asymptotic analysis because multi-cell systems are fundamentally interference-limited in the high-SNR regime [164]. When nothing else is said, $\text{BS}_j$ is assumed to know the channels $\mathbf{h}_{jk}$ and variances $\sigma_k^2$ perfectly to all users $k \in \mathcal{C}_j$. The case with CSI uncertainty is considered in Section 4.

Just as in the single-cell scenario, the transmission is limited by the $L$ power constraints in (1.4). An important difference is that the actual transmitted signals are $\mathbf{D}_k\mathbf{s}_k$ (and not $\mathbf{s}_k$), thus each weighting matrix $\mathbf{Q}_{lk}$ should satisfy the additional condition that $\mathbf{Q}_{lk} - \mathbf{D}_k^H\mathbf{Q}_{lk}\mathbf{D}_k$ is diagonal for all $l, k$ (e.g., being zero). This technical assumption makes sure that power cannot be allocated to unallowed subspaces for the purpose of reducing the (measured) power in the subspaces used for transmission — which is only possible when $\mathbf{Q}_{lk}$ is nondiagonal.

It is frequently assumed in multi-cell scenarios (but not necessary) that each power constraint only affects the signals from one of the base stations; for example, per-transmitter power constraints is represented by having $L = K_t$ and the constraint affecting $\text{BS}_l$ is

$$\mathbf{Q}_{lk}^{\text{per-BS}} = \mathbf{D}_k^H \operatorname{diag}\big(\mathbf{0}_{N_1+\cdots+N_{l-1}}, \mathbf{1}_{N_l}, \mathbf{0}_{N_{l+1}+\cdots+N_{K_t}}\big)\mathbf{D}_k \quad \forall l. \tag{1.10}$$

[9] It is implicitly assumed that $n_k$ is an ergodic process, which is not necessarily satisfied if unknown communication systems with fast adaptive resource allocation strategies are creating the interference; a further discussion is available in [302].

The analysis in this tutorial is applicable to any feasible set of power constraints, when nothing else is stated.

### 1.3.3 Examples of Multi-Cell Scenarios

We conclude this section by illustrating that the proposed DCC can describe a variety of multi-cell scenarios. Different examples are given on the following pages.

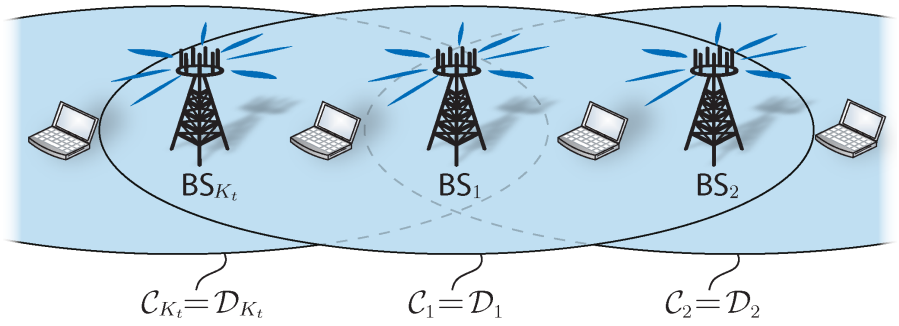



Fig. 1.7 Illustration of the multi-cell scenario called the *one-dimensional/linear Wyner model*. Users are jointly served by the closest base station and its two neighbors (in a cyclic manner), and only experience interference from these three base stations.

---

**Example 1.1 (Wyner model).** Based on an idea by A. Wyner [299], it can be assumed that users only receive signals from their own base station and the immediate neighboring base stations. This abstraction is supposed to capture the locality of interference. The one-dimensional (or linear) version of this model, where all devices are located on the boundary of a large circle, is illustrated in Figure 1.7. It is usually assumed that all users in the $j$th cell are jointly served by $\mathrm{BS}_{j-1}$, $\mathrm{BS}_j$, and $\mathrm{BS}_{j+1}$. This model was originally proposed for uplink transmission, but was used in [114, 191, 246] to analyze the ideal performance of joint downlink transmission.

Assume that there are $K_t$ base stations and $K_r$ users. If $\mathrm{MS}_k$ is geographically closest to $\mathrm{BS}_j$, then we have $\mathbf{D}_k = \mathbf{C}_k = \mathrm{diag}(\mathbf{0}_{N_1+\cdots+N_{j-2}}, \mathbf{I}_{N_{j-1}+N_j+N_{j+1}}, \mathbf{0}_{N_{j+2}+\cdots+N_{K_t}})$ since $\mathrm{MS}_k$ is served by $\mathrm{BS}_{j-1}$, $\mathrm{BS}_j$, and $\mathrm{BS}_{j+1}$ and only experiences interference from these base stations.

---

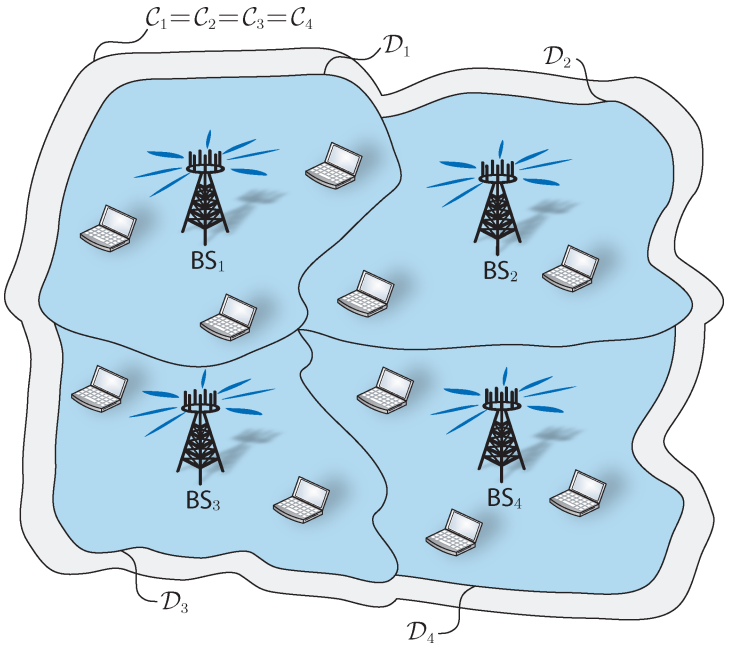



Fig. 1.8 Illustration of the multi-cell scenario of *coordinated beamforming*. Users are served by their own base station while interference is coordinated by joint resource allocation between all base stations.

---

**Example 1.2 (Coordinated Beamforming).** Coordinated beamforming means that each base station has a disjoint set of users to serve with data, but selects transmit strategies jointly with all other base stations to reduce inter-cell interference [59, 82, 139, 211]; see Figure 1.8. There is an arbitrary number of users in each cell. The special case with only one user per cell is called the *interference channel* [69, 101, 157, 235].

Assume that there are $K_t = 2$ base stations and $K_r$ users. Then, $\mathbf{D}_k = \mathrm{diag}(\mathbf{I}_{N_1}, \mathbf{0}_{N_2})$ for all MS$_k$ served by BS$_1$, while $\mathbf{D}_k = \mathrm{diag}(\mathbf{0}_{N_1}, \mathbf{I}_{N_2})$ for all MS$_k$ served by BS$_2$. In addition, $\mathbf{C}_1 = \mathbf{C}_2 = \mathbf{I}_N$ due to the global interference coordination.

---

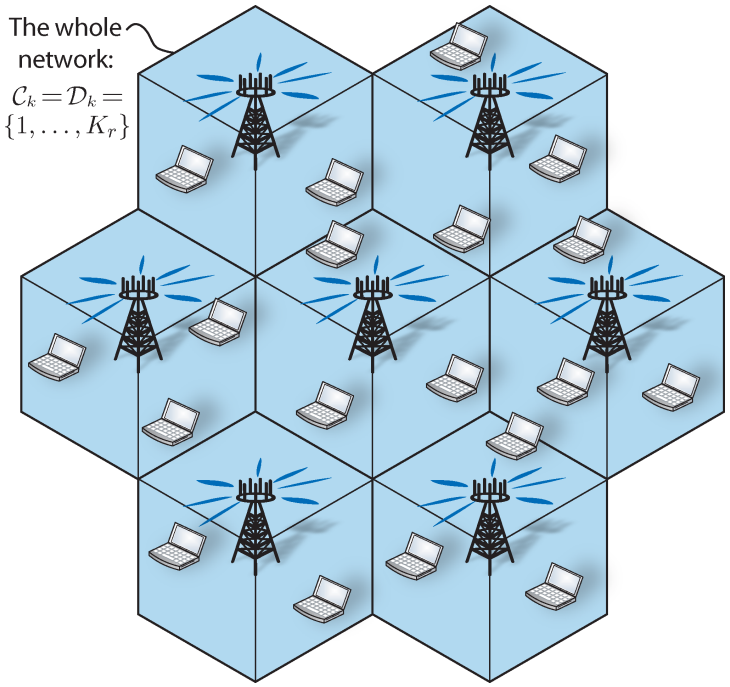



Fig. 1.9 Illustration of the *global joint transmission* scenario, where all cells and cell sectors are connected and perform joint transmission to all users in the whole network.

---

**Example 1.3 (Global Joint Transmission).** Ideally, all base stations can serve and coordinate interference to all users [125, 233, 321]. Even if the cellular network was originally built with many cells and cell sectors, this type of ideal/full CoMP turns the system into a single cell with distributed antenna arrays; see Figure 1.9. The main difference from the classic single-cell scenario might be the power constraints, which typically are defined per-antenna or per-transmitter.

This type of global joint transmission and interference coordination is represented by having $\mathbf{D}_k = \mathbf{C}_k = \mathbf{I}_N$ for all users $k$.

---

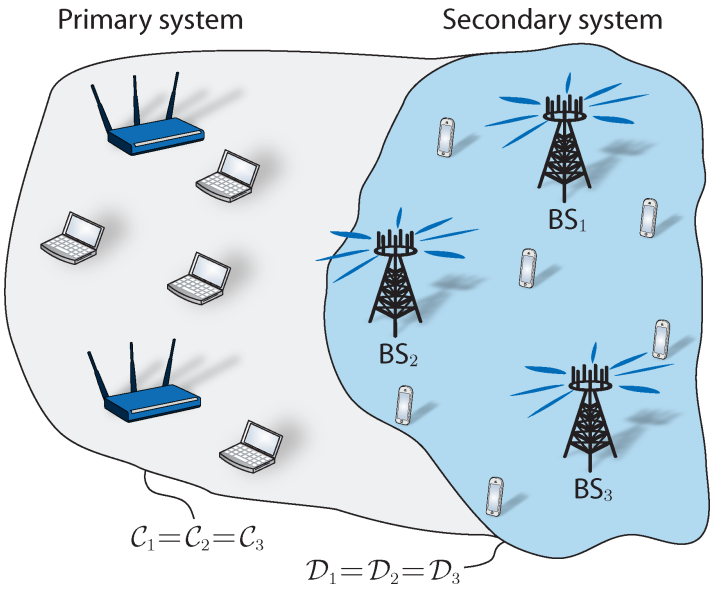



Fig. 1.10 Illustration of the scenario of *underlay cognitive radio*, where the secondary system is allowed to use frequency resources licensed by the primary system if the interference is kept below a threshold.

---

**Example 1.4 (Cognitive Radio).** *Underlay cognitive radio* is a scenario where a secondary system is allowed to use the licensed spectrum of a primary system if it causes mild interference on the primary system [90, 120, 230, 327]; see Figure 1.10. This scenario is particularly relevant when the primary system is not fully utilizing its spectrum.

Assume that users with indices in $\mathcal{K}_{\text{primary}} = \{1, \ldots, K_{\text{primary}}\}$ belong to the primary systems, while users in $\mathcal{K}_{\text{secondary}} = \{K_{\text{primary}} + 1, \ldots, K_r\}$ belong to the secondary system and are served by joint transmission. We then have $\mathbf{D}_k = \mathbf{0}_N$ for $k \in \mathcal{K}_{\text{primary}}$ and $\mathbf{D}_k = \mathbf{I}_N$ for $k \in \mathcal{K}_{\text{secondary}}$. We also have $\mathbf{C}_k = \mathbf{I}_N$ since interference is coordinated toward all users. Finally, we have $K_{\text{primary}}$ soft-shaping constraints of the form $\mathbf{Q}_{ki} = \mathbf{h}_i \mathbf{h}_i^H \; \forall k \in \mathcal{K}_{\text{secondary}}$ to limit the interference toward each primary user $i \in \mathcal{K}_{\text{primary}}$. The corresponding $q_i$ defines the maximal interference power that can be caused to user $i \in \mathcal{K}_{\text{primary}}$.

---

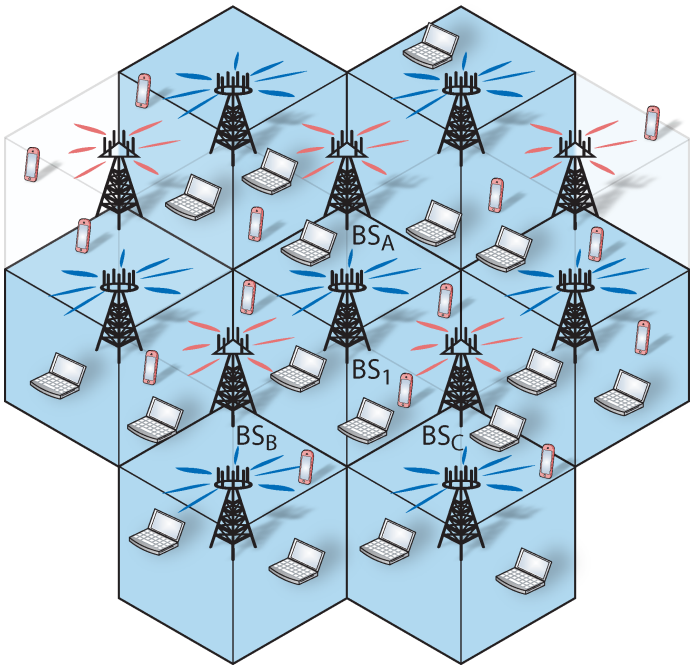



Fig. 1.11 Illustration of the scenario of *spectrum sharing* between two operators covering the same area, creating inter-operator interference.

---

**Example 1.5 (Spectrum Sharing Between two Operators).** *Spectrum sharing between operators* is a scenario where two operators agree to share some portion of their licensed frequency bands; see Figure 1.11 where Operator 1 has circular antenna arrays and serve laptops while Operator 2 has triangular arrays and serve smartphones.

Suppose $\mathrm{MS}_k$ is served by $\mathrm{BS}_1$ of Operator 1 with $\mathbf{D}_k = \mathrm{diag}(\mathbf{I}_{N_1}, \mathbf{0}_{N_2}, \ldots)$. The signal received at $\mathrm{MS}_k$ is a superposition of the signals from $\mathrm{BS}_1$ of Operator 1 and $\mathrm{BS_A}, \mathrm{BS_B}, \mathrm{BS_C}$ of Operator 2, thus $\mathbf{C}_k = \mathrm{diag}(\underbrace{\mathbf{I}_N}_{\text{BS 1}}, \mathbf{0}, \ldots, \mathbf{0}, \underbrace{\mathbf{I}_{N_\mathrm{A}}, \mathbf{I}_{N_\mathrm{B}}, \mathbf{I}_{N_\mathrm{C}}}_{\mathrm{BS_A}, \mathrm{BS_B}, \mathrm{BS_C}}, \mathbf{0}, \ldots)$. This model is easily extended to the case in which inter-cell interference from the same operator is also considered (by modifying the matrix $\mathbf{C}_k$ accordingly). Another extension is to apply full joint transmission within one operator, which could be modeled by $\mathbf{D}_k = \mathrm{diag}(\mathbf{I}_{N_1}, \mathbf{0}_{N_2}, \mathbf{I}_{N_3}, \mathbf{0}_{N_4}, \ldots)$.

---

## 1.4 Multi-Cell Performance Measures and Resource Allocation

In this section, we define a general way of measuring the performance in multi-cell systems. It is instructive to separate the performance into two parts: (1) the performance that each user experiences; and (2) the system utility which is a collection of simultaneously achievable user performances. These two parts are described and analyzed in the following subsections.

### 1.4.1 User Performance

To enable low-complexity and energy-efficient receivers, we assume *single user detection* meaning that a user is not attempting to decode and subtract interfering signals while decoding its own signals. This assumption is limiting in terms of spectral efficiency, except in the low-interference regime [4, 234], but requires less complex signal processing algorithms for reception. In principle, it also places the responsibility for interference control at the transmitter-side, where the computational resources are available. The corresponding SINR for $\mathrm{MS}_k$ is

$$\begin{aligned}\mathrm{SINR}_k(\mathbf{S}_1,\ldots,\mathbf{S}_{K_r}) &= \frac{\mathbf{h}_k^H\mathbf{C}_k\mathbf{D}_k\mathbf{S}_k\mathbf{D}_k^H\mathbf{C}_k^H\mathbf{h}_k}{\sigma_k^2+\mathbf{h}_k^H\mathbf{C}_k(\sum\limits_{i\neq k}\mathbf{D}_i\mathbf{S}_i\mathbf{D}_i^H)\mathbf{C}_k^H\mathbf{h}_k}\\ &= \frac{\mathbf{h}_k^H\mathbf{D}_k\mathbf{S}_k\mathbf{D}_k^H\mathbf{h}_k}{\sigma_k^2+\mathbf{h}_k^H\mathbf{C}_k(\sum\limits_{i\in\mathcal{I}_k}\mathbf{D}_i\mathbf{S}_i\mathbf{D}_i^H)\mathbf{C}_k^H\mathbf{h}_k},\end{aligned} \tag{1.11}$$

where the second equality follows from $\mathbf{C}_k\mathbf{D}_k=\mathbf{D}_k$ and $\mathbf{C}_k\mathbf{D}_i\neq\mathbf{0}$ only for users $i$ in

$$\mathcal{I}_k=\bigcup_{\{j\in\mathcal{J}:k\in\mathcal{C}_j\}}\mathcal{D}_j\setminus\{k\}. \tag{1.12}$$

This is the set of co-users being served by the same base stations that coordinate interference toward $\mathrm{MS}_k$. Observe that the SINR is a dimensionless quantity, thus it does not matter if the transmit and noise

powers are measured in milliwatt or watt. For brevity, we frequently write $\mathrm{SINR}_k$ instead of $\mathrm{SINR}_k(\mathbf{S}_1, \ldots, \mathbf{S}_{K_r})$ in this tutorial.

The signal-to-noise ratio (SNR) can be defined accordingly by removing the interference term in (1.11). We will however mostly use this term as an indication of the ideal signaling conditions to a given user: $q_j \frac{\|\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_k\|_2^2}{\sigma_k^2}$, where $q_j$ is the constraint that ultimately limits the transmit power. We show in Section 3.4 that the optimal transmission structure depends strongly on the SNR — roughly speaking, a low SNR is below 0 dB and a high SNR is above 20 dB.

Note that other channel gain based SINR expressions are possible. Consider the case in which $\mathrm{MS}_k$ receives two statistically independent data signals with correlation matrices $\mathbf{S}_k^{(1)}$ and $\mathbf{S}_k^{(2)}$, for example, from two different base stations. Then, the resulting SINR expression useful for information rate computation (after optimal receive processing with successive interference cancelation) is given by

$$\mathrm{SINR}_k^{\text{2-signals}}(\mathbf{S}_1, \ldots, \mathbf{S}_{K_r}) = \frac{\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_k (\mathbf{S}_k^{(1)} + \mathbf{S}_k^{(2)}) \mathbf{D}_k^H \mathbf{C}_k^H \mathbf{h}_k}{\sigma_k^2 + \mathbf{h}_k^H \mathbf{C}_k (\sum_{i \neq k} \mathbf{D}_i \mathbf{S}_i \mathbf{D}_i^H) \mathbf{C}_k^H \mathbf{h}_k}. \quad (1.13)$$

This expression is equivalent to (1.11) if all data signals are independent.[10] However, if $\mathbf{S}_k^{(2)}$ represents a multi-cast signal meant for multiple users, then (1.13) cannot be written as (1.11). Multi-cast signals can, for example, be used for overhead signaling to different groups of users [127, 245]. This type of multi-cast scenario is further described in Section 4.

Each user $k$ has its own quality measure represented by the user performance function $g_k : \mathbb{R}_+ \to \mathbb{R}_+$ of the SINR. This function describes the satisfaction of the user and generally depends on the service currently used (e.g., its throughput and delay constraints[11]) and on the priority given by the subscription profile.

[10] This is can be seen by defining $\mathbf{S}_k = \mathbf{S}_k^{(1)} + \mathbf{S}_k^{(2)}$.

[11] Voice traffic is an *inelastic* service as the user requires short delays and that a minimum information rate is constantly available (while higher rates unnecessary). On the contrary, Internet traffic is *elastic* as it can accept long delays and variations in the information rate, while the satisfaction is strictly increasing with the information rate.

**Definition 1.4 (User Performance Function).** The performance of $\mathrm{MS}_k$ is measured by an arbitrary continuous, differentiable, and *strictly monotonically increasing*[12] function $g_k(\mathrm{SINR}_k)$ of the SINR. This function satisfies $g_k(0) = 0$, for notational convenience.

With this definition, it is preferable for $\mathrm{MS}_k$ to have a large positive value on $g_k(\mathrm{SINR}_k)$ because it corresponds to good performance.[13] Ideally, the function $g_k(\cdot)$ should be selected to quantify the performance quality in a way comprehensible to the user and the system provider. It is certainly difficult to summarize and connect the user expectations and final service quality with a physical entity such as the SINR. Nevertheless, Definition 1.4 gives a reasonable structure since improving the signal quality should always increase the performance [196], or at least not degrade it [40].

Most of the analytical results in this tutorial only requires the structural properties in Definition 1.4 and are indifferent to the actual choice of user performance functions, therefore we will only explicitly specify $g_k(\cdot)$ when needed. Furthermore, the functions only need to satisfy the continuity and monotonicity properties in Definition 1.4 in the SINR ranges supported by the power constraints in (1.4). The assumption $g_k(0) = 0$ is nonlimiting and always fulfilled after a simple variable transformation. Here follow some common examples on performance measures that satisfy our definition.

**Example 1.6 (Information Rate).** The *achievable information rate* (or mutual information) is $g_k(\mathrm{SINR}_k) = \log_2(1 + \mathrm{SINR}_k)$ and describes the number of bits that can be conveyed to user $k$ (per channel use) with an arbitrarily low probability of decoding error [57]. The underlying

[12] A function $g_k : \mathbb{R} \to \mathbb{R}$ is *strictly monotonically increasing* if it for any $x, x' \in \mathbb{R}$ such that $x > x'$ also follows that $f(x) > f(x')$.

[13] If we would like to minimize some kind of error $\breve{g}_k(\mathrm{SINR}_k)$ that is strictly monotonically decreasing (e.g., mean square error or bit error rate), this can be reformulated into a maximization of the multiplicative inverse as $g_k(\mathrm{SINR}_k) = \frac{1}{\breve{g}_k(\mathrm{SINR}_k)} - \frac{1}{\breve{g}_k(0)}$ or maximization of the additive inverse as $g_k(\mathrm{SINR}_k) = \breve{g}_k(0) - \breve{g}_k(\mathrm{SINR}_k)$. Observe that both possibilities satisfy the condition of $g_k(0) = 0$ in Definition 1.4.

assumption is an infinite constellation $\mathbf{s}_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{S}_k)$, error-control coding over very many channel uses, and ideal decoding [65].

---

**Example 1.7 (Mean Square Error).** The sum *mean square error (MSE)* is $\mathrm{MSE}_k = \mathbb{E}\{\|\hat{\mathbf{s}}_k - \mathbf{s}_k\|_2^2\}$, where $\hat{\mathbf{s}}_k$ is an estimate of $\mathbf{s}_k$ obtained using the optimal Wiener filter [195] and noniterative reception. If $M$ data streams are intended for transmission to user $k$ (i.e., $\mathrm{rank}(\mathbf{S}_k) \leq M$), then $\mathrm{MSE}_k = M - \frac{\mathrm{SINR}_k}{1+\mathrm{SINR}_k}$. This error measure should be minimized, thus it is equivalent to maximizing $g_k(\mathrm{SINR}_k) = \frac{\mathrm{SINR}_k}{1+\mathrm{SINR}_k}$.

---

**Example 1.8 (Bit Error Rate).** The *bit error rate (BER)* for Gray coded transmission of a 16-QAM constellation is

$$P_{k,\text{16-QAM}} = \frac{3}{8}\mathrm{erfc}\left(\sqrt{\frac{1}{10}\mathrm{SINR}_k}\right) + \frac{1}{4}\mathrm{erfc}\left(\sqrt{\frac{9}{10}\mathrm{SINR}_k}\right) - \frac{1}{8}\mathrm{erfc}\left(\sqrt{\frac{5}{2}\mathrm{SINR}_k}\right), \tag{1.14}$$

where $\mathrm{erfc}(x) = \frac{2}{\sqrt{\pi}}\int_x^\infty e^{-t^2}dt$ is the complementary error function and $\mathrm{rank}(\mathbf{S}_k) \leq 1$ [73, 189]. This error measure should be minimized, thus it is equivalent to maximizing $g_k(\mathrm{SINR}_k) = 0.5 - P_{k,\text{16-QAM}}$.

---

In terms of merits and demerits, the information rate has a simple and marketable interpretation, but builds on idealized coding and signal processing assumptions. The MSE often gives simple expressions, but it can be argued that it is only vaguely connected to the user-experienced service quality. The BER is somewhat self-explanatory, but typically has complicated expressions (as seen from Example 1.8) and ignores channel coding which has a large impact on the effective error rate.

The actual throughput in modern communication systems, such as 3GPP LTE systems, can often be predicted as $\beta_1 \log_2(1 + \mathrm{SINR}_k/\beta_2)$, for some parameters $\beta_1 \in [0.5, 0.75]$ and $\beta_2 \in [1, 2]$ that reflect the

practical bandwidth and SNR efficiency, respectively [183]. This modified information rate expression is not perfect but is generally a good choice, because the parameters $\beta_1, \beta_2$ can be fitted to the output of a system-level simulator. However, there are certain practical situations in which adaptive coding and modulation is not possible (e.g., systems with very low-complexity devices) and BER/MSE measures are more appropriate.

The analysis and optimization procedure in this tutorial is applicable to any $g_k(\cdot)$ satisfying Definition 1.4; the particular choice will not affect the approach to achieve optimal resource allocation, but will certainly affect what is considered optimal.

Each transmitted data signal will in general affect all users and the impact is characterized by the channel gain region.

---

**Definition 1.5 (Channel Gain Region).** Consider the signal with correlation matrix $\mathbf{S}_k$. The received signal power at user $i$ is given by $x_{ki}(\mathbf{S}_k) = \mathbf{h}_i^H \mathbf{C}_i \mathbf{D}_k \mathbf{S}_k \mathbf{D}_k^H \mathbf{C}_i^H \mathbf{h}_i$. The *channel gain region* of this signal is defined as

$$\Omega_k = \left\{ (x_{k1}(\mathbf{S}_k), \ldots, x_{kK_r}(\mathbf{S}_k)) : \mathbf{S}_k \succeq \mathbf{0}_N, \mathrm{tr}(\mathbf{Q}_{lk}\mathbf{S}_k) \leq q_{lk} \quad \forall l \right\}. \tag{1.15}$$

---

The set $\Omega_k$ depends only on the signal correlation matrix $\mathbf{S}_k$ and on the per-user power constraints in (1.5). It describes the impact of the choice of $\mathbf{S}_k$ on the received channel gain at all users.

Note that the definition of the channel gain region in Definition 1.5 is different from the definition in [180] because of the feasible transmit strategies. Therefore, the next result which shows that $\Omega_k$ is compact and convex extends [180, Lemma 1].

---

**Definition 1.6.** A set $\mathcal{S} \subseteq \mathbb{R}^{K_r}$ is *compact* if it is closed and bounded. $\mathcal{S}$ is *convex* if $t\mathbf{r}_1 + (1-t)\mathbf{r}_2 \in \mathcal{S}$ whenever $\mathbf{r}_1, \mathbf{r}_2 \in \mathcal{S}$ and $t \in [0,1]$.

---

**Lemma 1.1.** The channel gain region $\Omega_k$ is compact and convex.

---

*Proof.* Define the vector with achievable channel gains as $\mathbf{x}_k(\mathbf{S}_k) = [x_{k1}(\mathbf{S}_k) \dots x_{kK_r}(\mathbf{S}_k)]^T$. The set of feasible signal correlation matrices is $\mathbb{S}_k = \left\{\mathbf{S}_k : \mathbf{S}_k \succeq \mathbf{0}_N, \mathrm{tr}(\mathbf{Q}_{lk}\mathbf{S}_k) \leq q_{lk} \quad \forall l\right\}$ and is compact and closed. Since $\Omega_k$ is achieved by a continuous mapping from the closed set $\mathbb{S}_k$, we can invoke [219, Theorem 4.14] to conclude that also $\Omega_k$ is a closed set.

It remains to show that $\Omega_k$ is convex: For any two points $\mathbf{x}_k(\mathbf{S}^{(1)}) \in \Omega_k$ and $\mathbf{x}_k(\mathbf{S}^{(2)}) \in \Omega_k$, we have to show that $\mathbf{x}_k(\mathbf{S}_z(t)) \in \Omega_k$ for $\mathbf{S}_z(t) = t\mathbf{S}^{(1)} + (1-t)\mathbf{S}^{(2)}$ and $t \in [0,1]$. It holds as

$$\begin{aligned} x_{ki}(\mathbf{S}_z(t)) &= \mathbf{h}_i^H \mathbf{C}_i \mathbf{D}_k \mathbf{S}_z(t) \mathbf{D}_k^H \mathbf{C}_i^H \mathbf{h}_i \\ &= \mathbf{h}_i^H \mathbf{C}_i \mathbf{D}_k \left( t\mathbf{S}^{(1)} + (1-t)\mathbf{S}^{(2)} \right) \mathbf{D}_k^H \mathbf{C}_i^H \mathbf{h}_i \\ &= t x_{ki}(\mathbf{S}^{(1)}) + (1-t) x_{ki}(\mathbf{S}^{(2)}). \end{aligned} \tag{1.16}$$

Furthermore, $\mathrm{tr}(\mathbf{Q}_{lk}\mathbf{S}_z(t)) \leq q_{lk}$ is satisfied because $\mathrm{tr}(\mathbf{Q}_{lk}\mathbf{S}_z(t)) = t\mathrm{tr}(\mathbf{Q}_{lk}\mathbf{S}^{(1)}) + (1-t)\mathrm{tr}(\mathbf{Q}_{lk}\mathbf{S}^{(2)}) \leq tq_{lk} + (1-t)q_{lk} = q_{lk}$. □

This lemma establishes the basic structure of the channel gain regions. The exact shape depends on the power constraints and the correlation between the channel vectors $\mathbf{C}_i^H \mathbf{h}_i$ of the users, as illustrated in Figure 1.12. If we consider a total power constraint, $\Omega_k$ resembles a triangle when the user channels are almost orthogonal (see Figure 1.12(a)), while it looks a line from the origin if the channels are almost parallel (see Figure 1.12(b)). Furthermore, the relative path losses $\|\mathbf{C}_i^H \mathbf{h}_i\|^2$ determine if the region looks thin or fat (see Figure 1.12(c)-(d)).

The relationship between individual user performance and channel gain regions is observed from the following SINR expression for $\mathrm{MS}_k$,

$$\mathrm{SINR}_k(x_{1k}(\mathbf{S}_1), \dots, x_{K_r k}(\mathbf{S}_{K_r})) = \frac{x_{kk}(\mathbf{S}_k)}{\sigma_k^2 + \sum\limits_{i \in \mathcal{I}_k} x_{ik}(\mathbf{S}_i)}. \tag{1.17}$$

From (1.17) the monotonicity of the SINR with respect to the different channel gains is easily observed. The SINR of $\mathrm{MS}_k$ is strictly monotonic increasing in $x_{kk}(\mathbf{S}_k)$ and strictly monotonic decreasing in $x_{ik}(\mathbf{S}_i)$ for all interfering links $i \in \mathcal{I}_k$. The conflict between the SINR expressions of different links becomes visible: increasing the own channel gain $x_{kk}$ might increase the channel gain $x_{ki}$ at some other user $i$ and thereby lower its SINR.

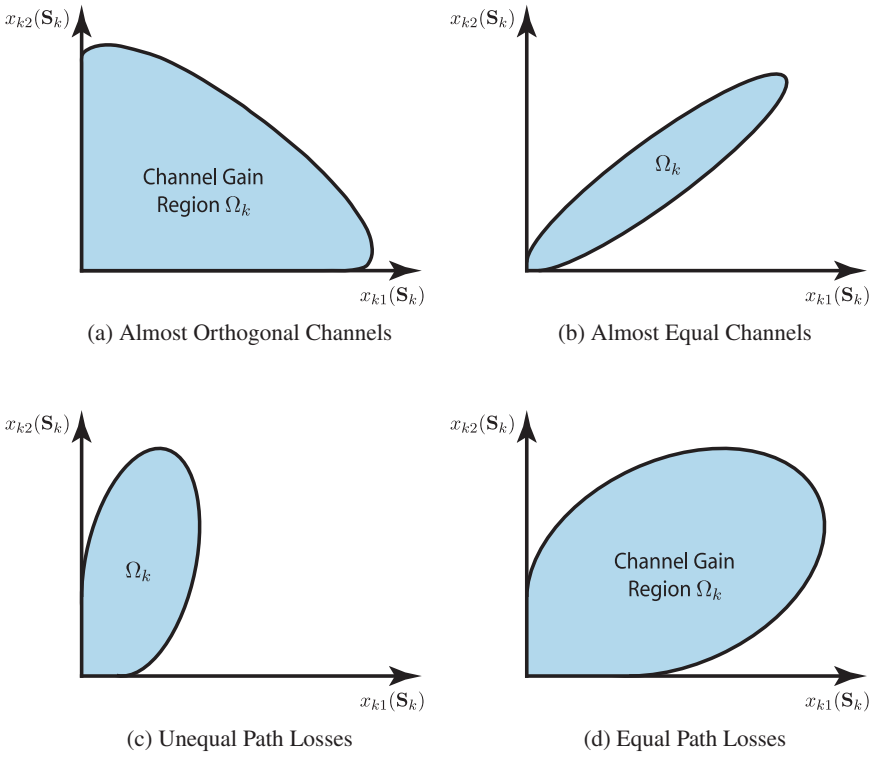



Fig. 1.12 Examples of channel gain regions with different shapes, but all being compact and convex. (a) and (b) illustrate the extremes of almost orthogonal and parallel channel vectors, respectively. (c) and (d) illustrate unequal and equal path losses $\|\mathbf{C}_i^H \mathbf{h}_i\|^2$, respectively.

The user performance function introduced in Definition 1.4 can also be expressed as a function of the channel gains,

$$g_k(\mathrm{SINR}_k) = g_k(x_{1k}(\mathbf{S}_1), \ldots, x_{K_r k}(\mathbf{S}_{K_r})). \tag{1.18}$$

By the monotonicity of the user performance function it follows that $g_k(\cdot)$ is also strictly monotonic increasing in $x_{kk}(\mathbf{S}_k)$ and strictly monotonic decreasing in $x_{ik}(\mathbf{S}_i)$ for all interfering links $i \in \mathcal{I}_k$.

### 1.4.2 Multi-Objective Resource Allocation

The channel gain regions highlight the inherent conflict and tradeoffs that appear when we want to maximize the performance of multiple users simultaneously. Each user has its own objective $g_k(\mathrm{SINR}_k)$ to be optimized, thus there are $K_r$ different objectives that typically are conflicting.

Optimization problems with multiple objectives appear naturally in many engineering fields to model tradeoffs between, for example, application performance, operational expenses, logistics, and environmental impacts. To analyze and obtain insights on such problems — without imposing any additional structure — it is common to formulate them mathematically as *multi-objective optimization problems* (MOPs). This tutorial will present and utilize some results and methods from the mathematical field of MOPs, but we refer to [38] for an in-depth survey.

Without loss of generality, our resource allocation problem is formulated as

$$\begin{aligned} \underset{\mathbf{S}_1 \succeq \mathbf{0}_N, \ldots, \mathbf{S}_{K_r} \succeq \mathbf{0}_N}{\text{maximize}} \quad & \{g_1(\mathrm{SINR}_1), \ldots, g_{K_r}(\mathrm{SINR}_{K_r})\} \\ \text{subject to} \quad & \sum_{k=1}^{K_r} \mathrm{tr}(\mathbf{Q}_{lk}\mathbf{S}_k) \leq q_l \quad \forall l. \end{aligned} \tag{1.19}$$

This MOP can be interpreted as searching for a transmit strategy $\mathbf{S}_1, \ldots, \mathbf{S}_{K_r}$ that satisfies the power constraints and maximizes the performance $g_k(\mathrm{SINR}_k)$ of all users [38]. Since the performance of different users are coupled by both power constraints and inter-user interference, there is generally not a *single* transmit strategy that simultaneously maximizes the performance of all users. For example, $\mathrm{SINR}_k$ in (1.11) improves if less interference is caused to $\mathrm{MS}_k$, but decreasing the interference at $\mathrm{MS}_k$ typically requires decreasing the useful signal power at other users and thereby degrading their SINRs. To study the conflicting objectives of a MOP it is instructive to consider the set of all feasible *operating points* $\mathbf{g} = [g_1 \ldots g_{K_r}]^T$ in (1.19) [38], which we call the performance region.[14]

---

**Definition 1.7.** The achievable *performance region* $\mathcal{R} \subseteq \mathbb{R}_+^{K_r}$ is

$$\mathcal{R} = \left\{ \big(g_1(\mathrm{SINR}_1), \ldots, g_{K_r}(\mathrm{SINR}_k)\big) \colon (\mathbf{S}_1, \ldots, \mathbf{S}_{K_r}) \in \mathbb{S} \right\} \tag{1.20}$$

where $\mathbb{S}$ is the set of feasible transmit strategies:

$$\mathbb{S} = \left\{ (\mathbf{S}_1, \ldots, \mathbf{S}_{K_r}) \colon \mathbf{S}_k \succeq \mathbf{0}_N, \quad \sum_{k=1}^{K_r} \mathrm{tr}(\mathbf{Q}_{lk}\mathbf{S}_k) \leq q_l \quad \forall l \right\}. \tag{1.21}$$

---

[14] The performance region can also be called the utility region or something that reflects the choice of user performance function (e.g., capacity region, rate region, or MSE region).

This region describes the performance that can be guaranteed to be simultaneously achievable by the users.[15] The $K_r$-dimensional performance region is nonempty as $\{\mathbf{0}_{K_r\times 1}\} \in \mathcal{R}$ and its shape depends strongly on the channel vectors, power constraints, and dynamic cooperation clusters. In general, $\mathcal{R}$ is not easily characterized and might be a nonconvex set, but we can prove that $\mathcal{R}$ is compact and normal [274].

---

**Definition 1.8.** A set $\mathcal{T}$ is called *normal on* $\mathcal{S} \subseteq \mathbb{R}^{K_r}$ if for any point $\mathbf{r} \in \mathcal{T}$, all $\mathbf{r}' \in \mathcal{S}$ with $\mathbf{r}' \leq \mathbf{r}$ also satisfy $\mathbf{r}' \in \mathcal{T}$ (componentwise inequality).

---

Normal sets are also known as comprehensive sets [39, 193].

---

**Lemma 1.2.** The achievable performance region $\mathcal{R}$ is compact and normal on $\mathbb{R}_+^{K_r}$.

---

*Proof.* To prove that $\mathcal{R}$ is a compact set, observe that the set of feasible transmit strategies $\mathbb{S}$ in (1.21) is compact. Next, observe that $g_k(\mathrm{SINR}_k)$ are continuous functions of $\mathbf{S}_1,\ldots,\mathbf{S}_{K_r}$ by definition. The compactness of $\mathcal{R}$ follows by invoking [219, Theorem 4.14], which says that the continuous mapping of a compact set is also a compact set. Since $\mathcal{R}$ is the image of a continuous mapping from $\mathbb{S}$, it is compact.

Proving that $\mathcal{R}$ is normal on $\mathbb{R}_+^{K_r}$ is a bit involved, although this property is quite intuitive. We outline the proof from [14, Lemma 5.1]. For any given $\mathbf{r} = (r_1,\ldots,r_{K_r}) \in \mathcal{R}$, we need to show that any $\mathbf{r}' = (r'_1,\ldots,r'_{K_r}) \in \mathbb{R}_+^{K_r}$ with $\mathbf{r}' \leq \mathbf{r}$ also belongs to $\mathcal{R}$. To this end, let $\mathbf{S}_1^*,\ldots,\mathbf{S}_{K_r}^*$ be a feasible transmit strategy that attains $\mathbf{r}$ and consider the alternative transmit strategy $p_1\mathbf{S}_1^*,\ldots,p_{K_r}\mathbf{S}_{K_r}^*$, where $p_1,\ldots,p_{K_r}$ is a set of power allocation coefficients that should belong to

$$\mathcal{A} = \left\{ (p_1,\ldots,p_{K_r})\colon \sum_{k=1}^{K_r} p_k \mathrm{tr}(\mathbf{Q}_{lk}\mathbf{S}_k^*) \leq q_l \quad \forall l \right\} \tag{1.22}$$

[15] Nonconvex performance regions can be increased by allowing for time-sharing between multiple operating points. This approach gives a region that equals the convex hull of $\mathcal{R}$, but the corresponding resource allocation problems are very complicated and not considered in this tutorial. The general framework for time-sharing in [39] can however be combined with the results in this tutorial. We also note that time-sharing can be viewed as part of the scheduling; see Section 4.7.

to make the strategy feasible. Obviously, the point $\mathbf{r}$ is achieved by selecting $(p_1^*,\ldots,p_{K_r}^*) = (1,\ldots,1)$. To prove that a given $\mathbf{r}' \leq \mathbf{r}$ also belongs to $\mathcal{R}$, we need to find $(p_1,\ldots,p_{K_r}) \in \mathcal{A}$ that gives this point. This corresponds to the conditions $\mathrm{SINR}_k = g_k^{-1}(r_k')\,\forall k$, which can be formulated as $K_r$ linear equations and solved using the approach in [205]. Finally, the existence of a $(p_1,\ldots,p_{K_r}) \in \mathcal{A}$ for any $\mathbf{r}' \leq \mathbf{r}$ can be proved using interference functions, see [227, Theorem 3.5]. □

This means that for any point $\mathbf{g} \in \mathcal{R}$, all points that give weaker performance than $\mathbf{g}$ are also in $\mathcal{R}$. This property is very natural and rational. In fact, if a region is not normal it looks very unnormal; see the illustrations in Figure 1.13 where only (b)–(f) are possible shapes for a performance region, while (a) is not a simply-connected set (i.e., contains holes) and has a strange boundary. Figure 1.13 also illustrates how the interference coupling and power constraints affect the region: (b) represents the degenerate case when the user have orthogonal channels and individual power constraints, while (c)–(f) describe a gradually increasing coupling between the users. Roughly speaking, $\mathcal{R}$ is convex when the users are weakly coupled and concave under strong coupling, while practical performance regions are hybrids of these extremes.

Apart from being compact, the performance region can also be upper bounded by a certain box.

---

**Definition 1.9.** A *box* is denoted $[\mathbf{a},\mathbf{b}]$, for some $\mathbf{a},\mathbf{b} \in \mathbb{R}^{K_r}$, and is the set of all $\mathbf{g} \in \mathbb{R}^{K_r}$ such that $\mathbf{a} \leq \mathbf{g} \leq \mathbf{b}$ (componentwise inequality).

---

**Lemma 1.3.** The performance region $\mathcal{R}$ satisfies $\mathcal{R} \subseteq [\mathbf{0},\mathbf{u}]$, where $\mathbf{u} = [u_1 \ldots u_{K_r}]^T$ is called the *utopia point*. The element $u_k$ is the optimum of the single-user optimization problem

$$\begin{aligned} &\underset{\mathbf{S}_k \succeq \mathbf{0}_N}{\text{maximize}} \;\; g_k\left(\frac{\mathbf{h}_k^H \mathbf{D}_k \mathbf{S}_k \mathbf{D}_k^H \mathbf{h}_k}{\sigma_k^2}\right) \\ &\text{subject to} \;\; \mathrm{tr}(\mathbf{Q}_{lk}\mathbf{S}_k) \leq q_l \quad \forall l. \end{aligned} \tag{1.23}$$

---

*Proof.* The single-user problem in (1.23) is achieved from the MOP in (1.19) by setting $\mathbf{S}_i = \mathbf{0}_N$ for all $i \neq k$. As inter-user interference

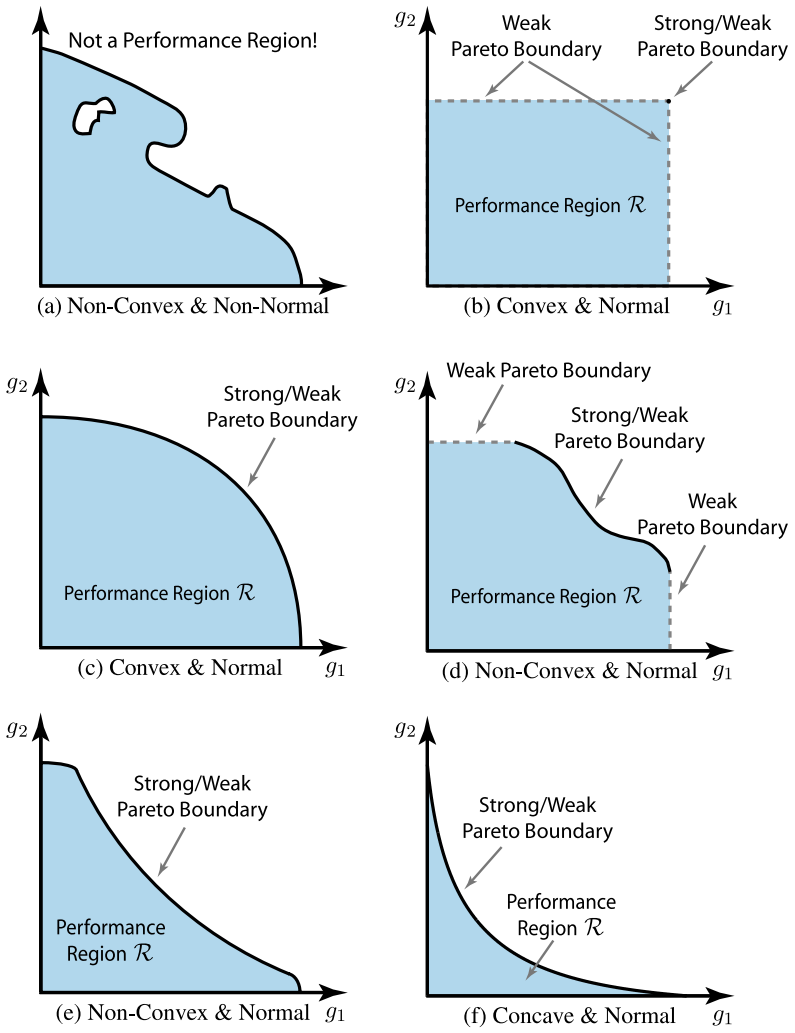



Fig. 1.13 Examples of compact regions with different shapes. Only (b)–(f) are normal and can thus be performance regions. The outer boundaries of (c), (e), (f) satisfy the conditions for both weak and strong Pareto optimality, while the horizontal and vertical parts of the outermost boundaries in (b) and (d) only satisfy weak Pareto optimality.

only can reduce $\mathrm{SINR}_k$, (1.23) provides an achievable upper bound on the performance of $\mathrm{MS}_k$ and it follows that $\mathcal{R} \subseteq [\mathbf{0}, \mathbf{u}]$. □

The utopia point $\mathbf{u}$ is the unique solution to (1.19) in degenerate scenarios (when the optimization decouples and all users can achieve

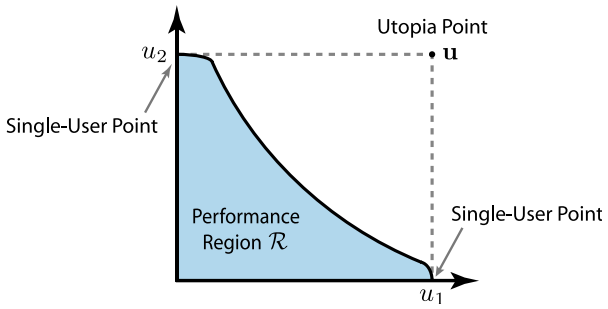



Fig. 1.14 Example of a performance region. The utopia point is shown, along with the single-user points achieved by solving (1.23).

maximal performance simultaneously, see Figure 1.13(b)). In general, $\mathbf{u} \notin \mathcal{R}$ and represents an unattainable upper bound on performance; see Figure 1.14. Since there is no total order of vectors in $\mathbb{R}_+^{K_r}$, we can only achieve a set of tentative vector solutions to (1.19) which are mutually unordered. These tentative solutions are all operating points in $\mathcal{R}$ that are not dominated by any other feasible point. These points are called *Pareto optimal* and are such that the performance cannot be improved for any user without deteriorating for at least one other user.

---

**Definition 1.10.** A point $\mathbf{y} \in \mathbb{R}_+^n$ is a *strong* Pareto optimal point of a compact normal set $\mathcal{T} \subseteq \mathbb{R}_+^n$, if $\mathbf{y} \in \mathcal{T}$ while $\{\mathbf{y}' \in \mathbb{R}_+^n : \mathbf{y}' \geq \mathbf{y}\} \cap \mathcal{T} \setminus \{\mathbf{y}\} = \emptyset$. The set of all strong Pareto optimal points is called the *strong Pareto boundary of* $\mathcal{T}$ and is denoted $\partial\mathcal{T}$.

In addition, a point $\mathbf{y} \in \mathbb{R}_+^n$ is a *weak* Pareto optimal point of a compact normal set $\mathcal{T} \subseteq \mathbb{R}_+^n$, if $\mathbf{y} \in \mathcal{T}$ while $\{\mathbf{y}' \in \mathbb{R}_+^n : \mathbf{y}' > \mathbf{y}\} \cap \mathcal{T} = \emptyset$. The set of all weak Pareto optimal points is called the *weak Pareto boundary of* $\mathcal{T}$ and is denoted $\partial^+\mathcal{T}$.

---

This definition distinguishes between (a) the strong Pareto boundary $\partial\mathcal{R}$ where the performance cannot be unilaterally improved for *any* user and (b) the weak Pareto boundary $\partial^+\mathcal{R}$ where we might be able to improve performance for some of the users but not simultaneously for *all* users. The strong Pareto boundary can be seen as the proper definition of the tentative solutions to a MOP, but we will see that the weak definition has better structural and analytical properties. The

strong Pareto boundary is always a subset of the weak Pareto boundary: $\partial\mathcal{R} \subseteq \partial^+\mathcal{R}$. The difference is visualized in Figure 1.13(b),(d), where the weak Pareto boundary contains the whole outermost boundary (including the vertical and horizontal parts) while the strong Pareto boundary only contains a subset of it. The single-user points $[0 \dots 0\, u_k\, 0 \dots 0]^T$ are always Pareto optimal, but might only satisfy the conditions for weak Pareto optimality.

Knowing that $\mathcal{R}$ is a normal, compact, and contained in $[\mathbf{0}, \mathbf{u}]$ simplifies the search for weak Pareto optimal points, particularly since these properties imply that $\mathcal{R}$ is simply-connected (i.e., contains no holes). We have the following result.

---

**Lemma 1.4.** The weak Pareto boundary $\partial^+\mathcal{R}$ of the performance region $\mathcal{R}$ is a compact and simply-connected set.

---

*Proof.* The compactness follows from that $\mathcal{R}$ is bounded and that the limit of any sequence of weak Pareto points must be contained in $\partial^+\mathcal{R}$ (easily shown by contradiction, see [40, Proposition A.3.4]). $\partial^+\mathcal{R}$ is simply-connected if there is a path in the set between any two points $\mathbf{r}_1, \mathbf{r}_2 \in \partial^+\mathcal{R}$. As $\mathcal{R}$ is normal there will always be a path between $\mathbf{r}_1$ and $\mathbf{r}_2$ that goes through the interior of $\mathcal{R}$, and every point on this path can be replaced by a dominating weak Pareto point to construct a Pareto optimal path; thus, $\partial^+\mathcal{R}$ is simply-connected. □

In comparison, the strong Pareto boundary $\partial\mathcal{R}$ need not be simply-connected, but can be a disconnected subset of the weak Pareto boundary. Therefore, it is easier to search for and characterize the weak Pareto boundary. This is mainly an academic limitation, because $\partial\mathcal{R} = \partial^+\mathcal{R}$ in most realistic scenarios. The explanation is that there are no truly orthogonal channels or resources in practice, thus there will always be some interference leakage that prevents unilateral improvements. As all properties of $\partial^+\mathcal{R}$ also hold for $\partial\mathcal{R}$, we sometimes refer to both as simply the *Pareto boundary*. We will later describe different algorithms for solving MOPs and as the Pareto boundary contains all tentative solutions, searching for Pareto optimal points is always an important part of such algorithms.

By the monotonicity of the user performance functions $g_k(\cdot)$ on the channel gains $x_{ki}(\mathbf{S}_k)$, there is a tight connection between the Pareto boundary of $\mathcal{R}$ and certain parts of the channel gain regions $\Omega_k$. Since the channel gain regions are not normal, we need to make a few definitions before specifying this relationship.

---

**Definition 1.11.** A vector $\mathbf{x}$ *dominates* a vector $\mathbf{y}$ *in direction* $\mathbf{e} \in \{-1,+1\}^n$, written as $\mathbf{x} \geq^{\mathbf{e}} \mathbf{y}$, if $x_i e_i \geq y_i e_i$ for all $i = 1,\ldots,n$ and there is at least one strict inequality.

---

Using this terminology, it is possible to describe the part of the boundary of a compact convex set we are interested in.

---

**Definition 1.12.** A point $\mathbf{y} \in \mathbb{R}_+^n$ is called an *upper boundary point* of a compact convex set $\mathcal{C} \subseteq \mathbb{R}_+^n$ *in direction* $\mathbf{e} \in \{-1,+1\}^n$ if $\mathbf{y} \in \mathcal{C}$ while the set $\{\mathbf{y}' \in \mathbb{R}_+^n : \mathbf{y}' \geq^{\mathbf{e}} \mathbf{y}\} \subseteq \mathbb{R}_+^n \setminus \mathcal{C}$. We denote the set of upper boundary points in direction $\mathbf{e}$ as $\partial^{\mathbf{e}}\mathcal{C}$.

---

An illustration of the definition is shown in Figure 1.15. The upper boundaries in the three directions $\mathbf{e}_1 = [+1\ +1]^T$, $\mathbf{e}_2 = [+1\ -1]^T$, and $\mathbf{e}_3 = [-1\ +1]^T$ are shown by the arrows. Note that the direction vector with all components equal to $-1$ is typically not of interest, as the

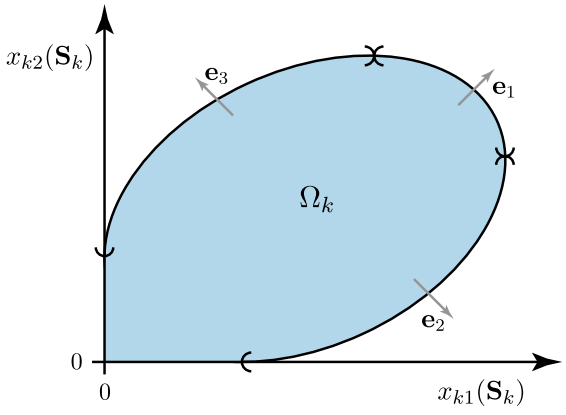



Fig. 1.15 Example of a channel gain region with upper boundary in direction $\mathbf{e}_1 = [+1\ +1]^T$, $\mathbf{e}_2 = [+1\ -1]^T$, and $\mathbf{e}_3 = [-1\ +1]^T$.

corresponding upper boundary is the origin. Also note that the upper boundary in direction $\mathbf{e}_1$ coincides with the usual Pareto boundary.

---

**Lemma 1.5.** Suppose the strong Pareto boundary of the performance region $\mathcal{R}$ is achieved by a transmit strategy $\mathbf{S}_1, \ldots, \mathbf{S}_{K_r}$. For each $k$, the matrix $\mathbf{S}_k$ also achieves the upper boundary of the channel gain region $\Omega_k$ in the direction $\mathbf{e}_k = [-1 \ldots -1 \ +1 \ -1 \ldots -1]^T$, where only the $k$th component is positive.

---

*Proof.* The proof works by contradiction. Assume that $\mathbf{S}_1, \ldots, \mathbf{S}_{K_r}$ achieve the strong Pareto boundary of $\mathcal{R}$ but there is a user $k$ that does not achieve the upper boundary of $\Omega_k$ in direction $\mathbf{e}_k$. Then, it is possible to shift the operating point $\mathbf{x}_k(\mathbf{S}_k)$ in $\Omega_k$ in the direction of the $k$th component without changing the other $K_r - 1$ components; that is, we can find $\mathbf{x}'_k \in \Omega_k$ with increased channel gain $x'_{kk} > x_{kk}$ for the intended user and the same channel gains $x'_{ki} = x_{ki}$ for all other users $i \neq k$. Since this new $\mathbf{x}'_k \in \Omega_k$ there exists a corresponding $\mathbf{S}'_k$ which achieves this point. Using the same set of signal correlation matrices for all other users but replacing $\mathbf{S}_k$ with $\mathbf{S}'_k$ leads to improved performance of user $k$ and unchanged performance for all other users. This is a contradiction to the assumption that $\mathbf{S}_1, \ldots, \mathbf{S}_{K_r}$ achieved the strong Pareto boundary of $\mathcal{R}$. □

The directions in Lemma 1.5 correspond to the monotonicity of the user performance functions on the channel gains. The performance function of user $k$ is monotonically increasing in $x_{kk}$ and monotonically decreasing in all other channel gains, therefore we want to maximize the channel gain $x_{kk}$ and minimize all other channel gains. This corresponds to a direction $\mathbf{e}_k = [-1 \ldots -1 \ +1 \ -1 \ldots -1]^T$ with $[\mathbf{e}_k]_k = 1$.

## 1.5 Basic Properties of Optimal Resource Allocation

Having defined the user performance functions and the concepts of performance region and channel gain regions, we have sufficient structure to derive two fundamental properties of the optimal multi-objective resource allocation:

- Sufficiency of single-stream beamforming;
- Conditions for full power usage.

These optimality properties are derived in this subsection. Taking these properties into account when solving (1.19) will greatly reduce the search space for optimal solutions. We will utilize the derived properties for simplified resource allocation in the remainder of this tutorial.

### 1.5.1 Sufficiency of Single-Stream Beamforming

The first property is the sufficiency of having signal correlation matrices $\mathbf{S}_k$ that are rank one. This might seem intuitive when each user only has a single (effective) receive antenna and is often assumed in resource allocation without discussion (see e.g., [59, 263, 264, 280, 308, 329]). In general, high-rank solutions might be necessary for optimality — it depends on the type of user performance functions and receive processing that is considered. In this tutorial, we assume single-user detection and $g_k(\cdot)$ of the type in Definition 1.4. We will show that it is sufficient (but not always necessary) to consider signal correlation matrices with rank one under these conditions. As the rank equals the number of data streams, this is called *single-stream beamforming*. First, we give a toy example from [18] showing that high-rank solutions sometimes can give the same performance (but never better) than the rank-one solutions.

---

**Example 1.9 (Rank of Optimal Strategy).** Consider a point-to-point system ($K_t = K_r = 1$) with $N = 2$ transmit antennas, the channel vector $\mathbf{h}_1 = [1\ 0]^T$, and per-antenna power constraints

$$\mathrm{tr}\left(\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \mathbf{S}_1\right) \leq 1, \quad \mathrm{tr}\left(\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \mathbf{S}_1\right) \leq 1. \tag{1.24}$$

The MOP in (1.19) reduces to a single-objective resource allocation problem which is solved optimally by both the rank-one matrix $\mathbf{S}_1 = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right]$ and by the rank-two matrix $\mathbf{S}_1 = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$.

---

To prove the sufficiency of rank-one signal correlation matrices, we will make use of some basic results in optimization theory (see Section 2.1 for an introduction to this topic). We start with a lemma.

**Lemma 1.6.** Consider the optimization problem

$$\begin{aligned} &\underset{\mathbf{V}\succeq 0}{\text{maximize}} \ \operatorname{tr}(\mathbf{AV}) \\ &\text{subject to } \operatorname{tr}(\mathbf{B}_m\mathbf{V}) \le b_m \quad m = 1,\dots,M, \end{aligned} \tag{1.25}$$

with an arbitrary Hermitian matrix $\mathbf{A}$, Hermitian matrices $\mathbf{B}_m \succeq 0$ that satisfy $\sum_{m=1}^{M} \mathbf{B}_m \succ \mathbf{0}$, and scalars $b_m \ge 0 \, \forall m$.

This problem is linear in $\mathbf{V}$ (and hence convex) and always has optimal solutions with $\operatorname{rank}(\mathbf{V}) \le 1$.

*Proof.* This is a linear optimization problem in $\mathbf{V}$ (see Section 2.1). The Lagrangian function is $\mathcal{L}(\mathbf{V}, \boldsymbol{\lambda}) = -\operatorname{tr}(\mathbf{AV}) + \sum_{m=1}^{M} \lambda_m (\operatorname{tr}(\mathbf{B}_m\mathbf{V}) - b_m)$ and the dual problem is

$$\begin{aligned} &\underset{\lambda_m \ge 0}{\text{minimize}} \ \sum_{m=1}^{M} \lambda_m b_m \\ &\text{subject to } \sum_{m=1}^{M} \lambda_m \mathbf{B}_m - \mathbf{A} \succeq \mathbf{0}. \end{aligned} \tag{1.26}$$

Observe that (1.25) and (1.26) are always feasible because $\mathbf{V} = \mathbf{0}$ satisfies all primal constraints and $\sum_{m=1}^{M} \mathbf{B}_m \succ \mathbf{0}$ implies dual feasibility. Therefore, strong duality holds (see Lemma 2.4) and the KKT conditions are necessary and sufficient for any optimal solution to (1.25):

$$\operatorname{tr}(\mathbf{B}_m\mathbf{V}) \le b_m \quad \forall m, \tag{1.27}$$

$$\sum_{m=1}^{M} \lambda_m \mathbf{B}_m - \mathbf{A} \succeq \mathbf{0}, \tag{1.28}$$

$$\lambda_m (\operatorname{tr}(\mathbf{B}_m\mathbf{V}) - b_m) = 0 \quad \forall m, \tag{1.29}$$

$$\operatorname{tr}\left(\mathbf{V}\left(\sum_{m=1}^{M} \lambda_m \mathbf{B}_m - \mathbf{A}\right)\right) = 0, \tag{1.30}$$

$$\mathbf{V} \succeq 0, \quad \lambda_m \ge 0 \quad \forall m. \tag{1.31}$$

To prove the sufficiency of rank-one solutions $\mathbf{V} = \mathbf{v}\mathbf{v}^H$, we consider the following alternative optimization problem

$$\begin{aligned} \underset{\mathbf{v}}{\text{maximize}} \;& \mathbf{v}^H \mathbf{A} \mathbf{v} \\ \text{subject to} \;& \mathbf{v}^H \mathbf{B}_m \mathbf{v} \leq b_m \quad \forall m. \end{aligned} \tag{1.32}$$

We want to show that every optimal solution $\mathbf{v}^*$ to (1.32) also satisfies (1.27)–(1.31) for $\mathbf{V} = \mathbf{v}^*(\mathbf{v}^*)^H$ and thus is optimal for (1.25). Although the cost function in (1.32) is generally nonconvex, the constraint functions are convex and thus the KKT conditions are necessary for $\mathbf{v}^*$ (see Lemma 2.2). Now, observe that (1.26) is also the dual problem of (1.32), therefore the feasibility is ensured by the same argument as above. Furthermore, (1.27) and (1.28) are satisfied by $\mathbf{v}^*$ and its corresponding Lagrange multipliers $\mu_m^*$. Next, (1.29) follows from the corresponding complementarity condition $\mu_m^*(\mathbf{v}^H \mathbf{B}_m \mathbf{v} - b_m) = 0$. Finally, (1.30) follows from multiplying the stationarity condition of (1.32), $\big(\sum_{m=1}^M \lambda_m \mathbf{B}_m - \mathbf{A}\big)\mathbf{v} = \mathbf{0}$, with $\mathbf{v}^H$ from the right-hand side. □

Before we show the sufficiency of rank-one signal correlation matrices for the performance region $\mathcal{R}$, we show the corresponding sufficiency for the channel gain regions $\Omega_1, \ldots, \Omega_{K_r}$.

---

**Lemma 1.7.** All upper boundary points of the channel gain region $\Omega_k$ in some arbitrary direction $\mathbf{e} \in \{-1, +1\}^{K_r}$ can be achieved by signal correlation matrices with $\operatorname{rank}(\mathbf{S}_k) \leq 1$.

---

*Proof.* Since $\Omega_k$ is convex and compact, the boundary can be achieved using the Supporting hyperplane theorem [273, Theorem 1.5] by the following optimization problem

$$\begin{aligned} \underset{\mathbf{S}_k \succeq 0}{\text{maximize}} \;& \sum_{i=1}^{K_r} \lambda_i x_{ki}(\mathbf{S}_k) \\ \text{subject to} \;& \operatorname{tr}(\mathbf{Q}_{lk} \mathbf{S}_k) \leq q_{lk} \quad \forall l. \end{aligned} \tag{1.33}$$

The objective function in (1.33) can be rewritten as

$$\begin{aligned}\sum_{i=1}^{K_r}\lambda_i x_{ki}(\mathbf{S}_k) &= \sum_{i=1}^{K_r}\lambda_i \mathbf{h}_i^H \mathbf{C}_i \mathbf{D}_k \mathbf{S}_k \mathbf{D}_k^H \mathbf{C}_i^H \mathbf{h}_i \\ &= \sum_{i=1}^{K_r}\lambda_i \mathrm{tr}(\mathbf{D}_k^H \mathbf{C}_i^H \mathbf{h}_i \mathbf{h}_i^H \mathbf{C}_i \mathbf{D}_k \mathbf{S}_k) \\ &= \mathrm{tr}\Bigg(\underbrace{\sum_{i=1}^{K_r}\lambda_i \mathbf{D}_k^H \mathbf{C}_i^H \mathbf{h}_i \mathbf{h}_i^H \mathbf{C}_i \mathbf{D}_k}_{\mathbf{A}_k}\mathbf{S}_k\Bigg).\end{aligned} \tag{1.34}$$

This is an optimization problem of the form (1.25) and thus the existence of solutions with $\mathrm{rank}(\mathbf{S}_k) \leq 1$ follows from Lemma 1.6. □

Note that $\mathrm{rank}(\mathbf{S}_k) \leq 1$ implies that the signal correlation matrix $\mathbf{S}_k$ is either rank one or identically zero; $\mathbf{S}_k = \mathbf{0}_N$ means no transmission.

By Lemma 1.7, the sufficiency of single-stream beamforming follows immediately for the performance region.

---

**Theorem 1.8.** Every point in the performance region $\mathcal{R}$ (including the weak Pareto boundary) can be achieved using single-stream beamforming (i.e., $\mathrm{rank}(\mathbf{S}_k) \leq 1\ \forall k$).

---

*Proof.* Lemma 1.7 shows that the boundary of each channel gain region $\Omega_k$ is obtained by $\mathbf{S}_k$ with $\mathrm{rank}(\mathbf{S}_k) \leq 1$. Since the strong Pareto boundary of the performance region is achieved by transmit strategies which achieve also the boundary of the channel gain regions (see Lemma 1.5), sufficiency of $\mathrm{rank}(\mathbf{S}_k) \leq 1$ follows. To show that also points on the *weak* Pareto boundary (and all other points in $\mathcal{R}$) are achievable by rank-one solutions, we can simply repeat the approach in the proof of Lemma 1.2 (which showed that $\mathcal{R}$ is normal by fixing the beamforming directions and changing the power allocation). □

The implication of Theorem 1.8 is that any operating point in $\mathcal{R}$ (and particularly Pareto optimal points) can be achieved using single-stream beamforming, thus all tentative solutions to the MOP in (1.19)

are achievable by $\mathbf{S}_k = \mathbf{v}_k \mathbf{v}_k^H$ for some *beamforming vectors* $\mathbf{v}_k \in \mathbb{C}^{N \times 1}\ \forall k$. Without loss of generality, we can reformulate (1.19) as

$$\begin{aligned} &\underset{\mathbf{v}_1,\ldots,\mathbf{v}_{K_r}}{\text{maximize}} \ \{g_1(\text{SINR}_1),\ldots,g_{K_r}(\text{SINR}_{K_r})\} \\ &\text{subject to } \text{SINR}_k = \frac{|\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_k \mathbf{v}_k|^2}{\sigma_k^2 + \sum\limits_{i \neq k} |\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_i \mathbf{v}_i|^2} \quad \forall k, \\ &\qquad \sum_{k=1}^{K_r} \mathbf{v}_k^H \mathbf{Q}_{lk} \mathbf{v}_k \leq q_l \quad \forall l. \end{aligned} \tag{1.35}$$

Considering (1.35) instead of (1.19) greatly reduces the search space for optimal solutions and makes the solution easier to implement in practice, because vector coding or successive interference cancelation are required if $\text{rank}(\mathbf{S}_k) > 1$ [89]. The problem formulation in (1.35) will be used as the starting point in the remainder of this tutorial.

### 1.5.2 Conditions for Full Power Usage

If only the total transmit power over all base stations is constrained, it is trivial to prove that any Pareto optimal solution to (1.19) and (1.35) will use all available power. Under general power constraints, it may be better not to use full power at each transmitter or antenna; there is a balance between increasing channel gains of useful signals and limiting the interference. This is illustrated by the following toy example, which is based on [18].

---

**Example 1.10 (Limited Power Usage).** Consider a two-user interference channel with single-antenna base stations ($K_t = K_r = 2$, $N_1 = N_2 = 1$) and the channel vectors $\mathbf{h}_1 = [1\ \sqrt{1/10}]^T$ and $\mathbf{h}_2 = [\sqrt{1/2}\ 1]^T$. $\text{BS}_j$ transmits to $\text{MS}_j$ and coordinates interference to both users, meaning that $\mathbf{D}_1 = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right]$, $\mathbf{D}_2 = \left[\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right]$, and $\mathbf{C}_1 = \mathbf{C}_2 = \mathbf{I}_2$. The per-transmitter power is constrained as $\text{tr}\,(\mathbf{D}_j \mathbf{S}_j) \leq 20\ \forall j$.

The single-user point of $\text{MS}_1$ is achieved by $\mathbf{S}_1 = 20\mathbf{D}_1$ and $\mathbf{S}_2 = \mathbf{0}_2$, while the corresponding point for $\text{MS}_2$ is achieved by $\mathbf{S}_1 = \mathbf{0}_2$ and $\mathbf{S}_2 = 20\mathbf{D}_2$. Observe that only the base station associated with the active user is satisfying its power constraint with equality.

Furthermore, the operating point where both users have exactly the same SINR is achieved by $\mathbf{S}_1 = 10\mathbf{D}_1$ and $\mathbf{S}_2 = 20\mathbf{D}_2$. This transmit strategy gives $\mathrm{SINR}_1 = \mathrm{SINR}_2 = \frac{10}{3}$. Observe that only $\mathrm{BS}_2$ uses full power and if $\mathrm{BS}_1$ would increase its power then $\mathrm{SINR}_2$ decreases. This shows that this is a strong Pareto optimal point.

---

In principle, knowing that a certain constraint is active (i.e., satisfied with equality at the optimal solution) removes one dimension from the resource allocation problem. The following theorem provides conditions for when full power should be used in general multi-cell systems.

---

**Theorem 1.9.** The following holds for the multi-objective resource allocation problems (1.19) and (1.35):

- Every weak Pareto optimal point can be achieved by a transmit strategy that satisfies at least one power constraint with equality.
- If only the total power per transmitter is constrained, then every strong Pareto optimal point requires that $\mathrm{BS}_j$ uses full power if $\mathcal{D}_j \neq \emptyset$ and the channels $\mathbf{h}_{jk}$ for all users $k \in \mathcal{C}_j$ are linearly independent.

---

*Proof.* If $q_l = 0$ for some $l$, the first part of the theorem is always satisfied. Now assume that $q_l > 0 \, \forall l$. Let $\mathbf{S}_1^*, \ldots, \mathbf{S}_{K_r}^*$ be a transmit strategy that achieves the weak Pareto boundary and assume that all power constraints in (1.4) are inactive. We define

$$\varsigma = \max_{1 \leq l \leq L} \sum_{k=1}^{K_r} \frac{\mathrm{tr}(\mathbf{Q}_{lk}\mathbf{S}_k^*)}{q_l} \tag{1.36}$$

and note that $\varsigma > 1$ since all constraints are inactive. The alternative strategy $\varsigma\mathbf{S}_1^*, \ldots, \varsigma\mathbf{S}_{K_r}^*$ will satisfy all constraints and at least one of them will be active. The performance is not decreased since $\varsigma$ can be seen as decreasing the relative noise power in each SINR in (1.11). Thus, there always exists a solution with at least one active constraint.

The second part is proved by contradiction. Suppose $\widetilde{\mathbf{S}}_1, \ldots, \widetilde{\mathbf{S}}_{K_r}$ achieves a strong Pareto optimal point and that $\mathrm{BS}_j$ is not using full power (but satisfies the conditions in the theorem); that is,

$$\sum_{k=1}^{K_r} \mathrm{tr}\left(\mathbf{Q}_{jk}^{\text{per-BS}} \widetilde{\mathbf{S}}_k\right) < q_j \tag{1.37}$$

where $\mathbf{Q}_{jk}^{\text{per-BS}}$ was defined in (1.10). The assumption of linear independence means that it exists $k \in \mathcal{D}_j$ with

$$\mathbf{h}_{jk} \notin \mathrm{span}\left(\bigcup_{i \in \mathcal{C}_j \setminus \{k\}} \{\mathbf{h}_{ji}\}\right). \tag{1.38}$$

Therefore, it exists a unit-norm vector $\mathbf{v} \neq \mathbf{0}_{N_j \times 1}$ such that $\mathbf{h}_{jk}^H \mathbf{v} \neq 0$ and $\mathbf{h}_{ji}^H \mathbf{v} = 0$ for all $i \in \mathcal{C}_j \setminus \{k\}$ (i.e., a zero-forcing vector). Then, the alternative signal correlation matrix $\mathbf{S}_k = \widetilde{\mathbf{S}}_k + \widetilde{\mathbf{v}}\widetilde{\mathbf{v}}^H$ with

$$\widetilde{\mathbf{v}} = \left[\mathbf{0}_{1 \times N_1 + \cdots + N_{j-1}} \sqrt{q_j - \sum_k \mathrm{tr}(\mathbf{Q}_{jk}^{\text{per-BS}} \widetilde{\mathbf{S}}_k)} \mathbf{v}^T \; \mathbf{0}_{1 \times N_{j+1} + \cdots + \times N_{K_t}}\right]^T \tag{1.39}$$

will strictly increase the signal power and cause exactly the same inter-user interference as $\widetilde{\mathbf{S}}_k$. As $g_k(\cdot)$ is strictly increasing we have unilaterally improved the performance of $\mathrm{MS}_k$ which is a contradiction to the strong Pareto optimality. □

The first implication from Theorem 1.9 is that at least one power constraint should be active at any Pareto optimal point. Second, observe that the linear independence of user channels is a very mild condition when $|\mathcal{C}_j| \leq N_j$ (e.g., satisfied with probability one when the channel realizations are drawn from a stochastic distribution with non-singular covariance matrices). Roughly speaking, the fewer users that a base station coordinates interference to, the more power is used at this base station at strong Pareto optimal points. The condition on linear independence can be relaxed to the existence of (at least) one

user in $\mathcal{D}_j$ with a channel linearly independent to all other users in $\mathcal{C}_j$ that are actually scheduled (i.e., receive nonzero signal power).

## 1.6 Subjective Solutions to Resource Allocation

Recall that the Pareto boundary of the performance region contains all tentative solutions to the MOP in (1.35), each representing a certain tradeoff between the users' performance. Whenever the utopia point is outside of the performance region, there is no objectively optimal resource allocation — there are multiple strong Pareto optimal points and none of these are distinctly better than the others. To actually compare the merits of different Pareto optimal points, the *system designer* (or decision maker) needs to bring in its own subjective perspective on system utility. Different methods to obtain subjectively optimal solutions are outlined in this section and will be the subject of the subsequent sections of this tutorial.

A common approach is to let the system designer describe its preferences as an aggregate system utility function $f : \mathcal{R} \to \mathbb{R}$ that takes any point in $\mathcal{R}$ as input and produces a scalar value describing how preferable this point is (large output means high preference).

---

**Definition 1.13 (System Utility Function).** A *system utility function* is denoted $f(g_1(\mathrm{SINR}_1), \ldots, g_{K_r}(\mathrm{SINR}_{K_r}))$ and is Lipschitz continuous[16] and *monotonically increasing*[17] on $[\mathbf{0}, \mathbf{u}]$.

---

This definition incorporates most system utility functions that appear in literature. In fact, many frequently used functions are *strictly* increasing functions, as seen in the following example [130, 168].

---

**Example 1.11 (System Utility Functions).** For a given operating point $\mathbf{g} = (g_1, \ldots, g_{K_r}) \in \mathcal{R}$, the following system utility functions

[16] A function $f : [\mathbf{a}, \mathbf{b}] \to \mathbb{R}$ is *Lipschitz continuous with Lipschitz constant* $L_f$ if $|f(\mathbf{g}) - f(\mathbf{g}')| \leq L_f \|\mathbf{g} - \mathbf{g}'\|_1$ for all $\mathbf{g}, \mathbf{g}' \in [\mathbf{a}, \mathbf{b}]$.

[17] A function $f : \mathbb{R}^n \to \mathbb{R}$ is *monotonically increasing* if for any $\mathbf{g}, \mathbf{g}' \in \mathbb{R}^n$ such that $\mathbf{g} \geq \mathbf{g}'$ it follows that $f(\mathbf{g}) \geq f(\mathbf{g}')$. The function is *strictly* monotonically increasing if for any $\mathbf{g}, \mathbf{g}'' \in \mathbb{R}^n$ such that $\mathbf{g} > \mathbf{g}''$, it also follows that $f(\mathbf{g}) > f(\mathbf{g}'')$.

satisfy[18] Definition 1.13:

- Weighted arithmetic mean: $f(\mathbf{g}) = \sum_k w_k g_k$
  (also known as weighted sum utility);
- Weighted geometric mean: $f(\mathbf{g}) = \prod_k g_k^{w_k}$
  (also known as weighted proportional fairness [130]);
- Weighted harmonic mean: $f(\mathbf{g}) = \left(\sum_k \frac{w_k}{g_k}\right)^{-1}$;
- Weighted max-min fairness: $f(\mathbf{g}) = \min_k \frac{g_k}{w_k}$
  (also known as weighted worst-user performance);
- Weighted compromise: $f(\mathbf{g}) = -(\sum_k (w_k(r_k^* - g_k))^p)^{1/p}$
  (for some reference point $\mathbf{r}^* \in \mathbb{R}_+^n \setminus \mathcal{R}$ and $1 \leq p \leq \infty$).

The weighting factors $w_k \geq 0$ can be taken to have unit sum, $\sum_{k=1}^{K_r} w_k = 1$, without loss of generality. In case of equal weighting factors, the arithmetic mean maximizes the aggregate system utility $\sum_k g_k$, while the geometric mean, harmonic mean, and max-min fairness gradually sacrifice aggregate utility to achieve more fairness among the users. For a given type of system utility function, the weighting factors can compensate for heterogeneous user channel conditions, handle delay constraints, enforce subscription profiles, etc.

---

There are other system utility functions, for example, the $\alpha$-proportional fairness in [179] that bridges the gap between proportional fairness and max-min fairness by varying a parameter (the arithmetic and harmonic means are also represented by certain parameter values). Weighted utilities for best-effort users are given in [112].

Based on a system utility function, the multi-objective optimization problem in (1.35) can be converted (called *scalarization*) to the

[18] Every continuously differentiable function is locally Lipschitz continuous, but some functions are not globally Lipschitz since the first derivative becomes infinite when approaching the origin. The weighted geometric mean $\prod_k g_k^{w_k}$ has such problems, but this can be resolved by optimizing $\prod_k g_k^{cw_k}$ instead where $c$ is selected to make $cw_k > 1 \forall k$. The weighted harmonic mean also needs additional treatment.

following single-objective optimization problem

$$\begin{aligned} &\underset{\mathbf{v}_1,\ldots,\mathbf{v}_{K_r}}{\text{maximize}} \; f\left(g_1(\text{SINR}_1),\ldots,g_{K_r}(\text{SINR}_{K_r})\right) \\ &\text{subject to } \text{SINR}_k = \frac{|\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_k \mathbf{v}_k|^2}{\sigma_k^2 + \sum_{i \neq k} |\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_i \mathbf{v}_i|^2} \quad \forall k, \\ &\sum_{k=1}^{K_r} \mathbf{v}_k^H \mathbf{Q}_{lk} \mathbf{v}_k \leq q_l \quad \forall l. \end{aligned} \tag{1.40}$$

This problem has a single (nonunique) solution, because the system utility function resolves the conflicting interests in the MOP. The selection of $f(\cdot)$ is therefore very important and should be based on a profound knowledge of $\mathcal{R}$ — the alternative of just selecting $f(\cdot)$ out of the blue corresponds to making decisions without knowing the alternatives. Two of the main objectives of this tutorial is to characterize the performance region and develop a framework for solving any single-objective resource allocation problem of the form (1.40). The latter can be viewed as a *network utility maximization* [40, 53, 131, 194], thus we can utilize many of the results on distributed optimization that has been developed under this umbrella; see Section 4.2.

---

**Remark 1.2 (All Utility Functions are Subjective).** Observe that all utility functions are subjective by nature, because each function imposes a certain order of vectors in the performance region and in $\mathbb{R}_+^{K_r}$. Although this transforms the resource allocation into the tractable form (1.40) where there is a single solution, this is only because all other Pareto optimal points are discarded by the choice of $f(\cdot)$. Therefore, we stress that the particular choice of $f(\cdot)$ should always be clearly motivated in research papers and not considered as given beforehand.

---

The basic connection between $\mathcal{R}$ and $f(\cdot)$ is given by the following important result.

---

**Lemma 1.10.** If $f(\cdot)$ is an increasing function, then the global optimum to (1.40) is attained on $\partial^+\mathcal{R}$. In addition, for any $\tilde{\mathbf{g}} \in \partial^+\mathcal{R}$ there exists a (strictly) increasing $f(\cdot)$ for which (1.40) has $\tilde{\mathbf{g}}$ as global optimum.

---

*Proof.* For the first statement, assume that $\bar{\mathbf{g}} \notin \partial^+\mathcal{R}$ is a global optimum to (1.40). By the definition of the weak Pareto boundary and using that $f(\cdot)$ is increasing, there exist a point $\mathbf{g}' \in \partial^+\mathcal{R}$ with $\mathbf{g}' \geq \bar{\mathbf{g}}$. This point satisfies $f(\mathbf{g}') \geq f(\bar{\mathbf{g}})$ and therefore also solves (1.40).

The second statement is proved using the weighted max-min fairness function $f(\mathbf{g}) = \min_{\{k:\tilde{g}_k>0\}} g_k/\tilde{g}_k$ for given $\tilde{\mathbf{g}} = (\tilde{g}_1, \ldots, \tilde{g}_{K_r}) \in \partial^+\mathcal{R}$. Obviously, $\max_{\mathbf{g}\in\mathcal{R}} f(\mathbf{g}) \geq f(\tilde{\mathbf{g}}) = 1$ and assume for the purpose of contradiction that there exists $\mathbf{g}^* \in \mathcal{R}$ that achieves strict inequality. This means that $\mathbf{g}^* > \tilde{\mathbf{g}}$ and thus $\tilde{\mathbf{g}}$ cannot be a weak Pareto optimal point since it requires $\{\mathbf{y}' \in \mathbb{R}_+^n : \mathbf{y}' > \tilde{\mathbf{g}}\} \cap \mathcal{R} \neq \emptyset$ (see Definition 1.10). This contradiction yields $\max_{\mathbf{g}\in\mathcal{R}} f(\mathbf{g}) = f(\tilde{\mathbf{g}})$ and thus $\tilde{\mathbf{g}}$ is the (nonunique) global optimum. □

Based on this lemma, we only need to search the weak Pareto boundary of $\mathcal{R}$ to solve any resource allocation problem of the form (1.40). Unfortunately, this is not as simple as it seems; we will show in Section 2 that (1.40) can only be solved in an efficient manner in certain special cases (e.g., depending on $f(\cdot)$, the number of transmit antennas, and the structure of the power constraints).

Similar to Lemma 1.10, there is an important connection between (1.40) and the channel gain regions.

---

**Corollary 1.11.** Suppose the solution to the optimization problem in (1.40) is achieved by signal correlation matrices $\mathbf{S}_1, \ldots, \mathbf{S}_{K_r}$ (with $\text{rank}(\mathbf{S}_k) \leq 1 \forall k$). Each $\mathbf{S}_k$ achieves a point on the upper boundary of the corresponding channel gain region $\Omega_k$ in direction $\mathbf{e}_k$ for all $k$.

---

*Proof.* The corollary follows from the monotonicity of $f(\cdot)$, Lemma 1.5, and Lemma 1.10. □

It is important to note that the set of transmit strategies that achieve points on the upper boundaries of the channel gain regions is much larger than the set of transmit strategies that achieves operating points on the Pareto boundary of $\mathcal{R}$, which again is much larger than the set of transmit strategies that maximizes $f(\cdot)$ in (1.40). The reason is that the upper boundary of *each* of the $K_r$ channel gain

regions has dimension $K_r - 1$ whereas the Pareto boundary of $\mathcal{R}$ has only dimension $K_r - 1$.

### 1.6.1 Four Methods to Solve Resource Allocation Problems

We have shown how scalarization converts the MOP in (1.35) into a single-objective problem (1.40) with a single solution. There are different ways of utilizing scalarization for finding a Pareto optimal point that makes the system designer satisfied. The preferable approach depends on how well the system designer can specify its subjective views in mathematical terms, and whether the system designer is taking an active or passive part in the optimization. The different methods can be categorized as follows [38, 324]:

(1) *No-preference methods* are applied when the system designer has no subjective preference on the final solution. To emphasize neutrality, (1.40) can be solved using a weighted system utility function (see Example 1.11) where the weighting factors are used for normalization (i.e., using the utopia point for weighting as $w_k = \frac{u_k}{\sum_{i=1}^{K_r} u_i}$).

(2) *A priori methods* are used when the system designer has a clear invariable goal, corresponding to a certain $f(\cdot)$. For instance, an optimistic reference point $\mathbf{r}^*$ might be given in advance and the optimal solution minimizes the distance to this point as $f(\mathbf{g}) = -\|\mathbf{r}^* - \mathbf{g}\|_p$ in the $L_p$-norm (i.e., a compromise problem). Maximizing the sum utility is another example. Any prior knowledge of the performance region and system-wide preference on the final solution should be taken into account when selecting $f(\cdot)$.

(3) *A posteriori methods* generate a set of sample points on the Pareto boundary (the whole set is infinite and nontrivial to characterize) and let the system designer select among these points. Based on Lemma 1.10, sample points are achieved by solving (1.40) for a set of different system utility functions. For example, a certain type of function can be selected from Example 1.11 and the weighting factors are then varied over

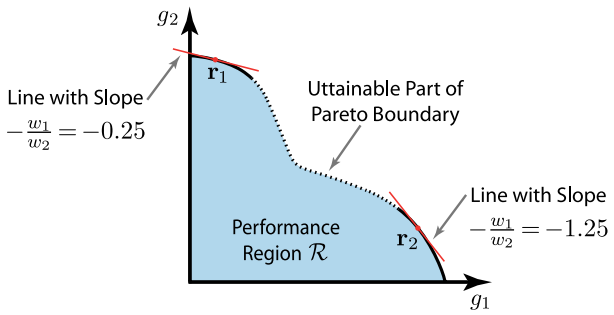
a grid. Keep in mind that the whole Pareto boundary cannot be reached by all types of functions (see Remark 1.3).

(4) *Interactive methods* can be viewed as an iterative combination of *a priori* and *a posteriori* methods, where each iteration generates new sample points on the Pareto boundary based on previous suggestions from the system designer. The advantage of this approach is that the preference of the system designer can be modified as the shape of Pareto boundary (i.e., the different alternatives) is learned, thus giving a kind of psychological convergence to the final solution.

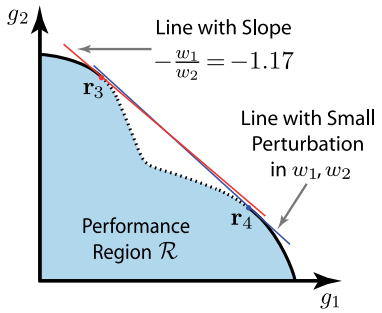
All of these methods involve one or multiple scalarizations of the MOP into SOPs of the form (1.40). Section 2 will therefore be devoted to solving SOP for any choice of $f(\cdot)$. Section 3 derives structure on the optimal transmit strategies and parameterizes the Pareto boundary. Based on the knowledge and experience from these sections, we will return to the aforementioned four methods in Section 3.5. We will then shed light on how these methods can be formulated and implemented efficiently for practical resource allocation.

---

**Remark 1.3 (Shortcomings of Weighted Arithmetic Mean).** It has become a common practice to optimize the weighted arithmetic mean (e.g., the weighted sum information rate) in the area of communications. This could make sense when $\mathcal{R}$ is convex, which holds for the ideal capacity region but not necessarily in other scenarios. Even if all possible weights are considered, the weighted arithmetic mean only finds Pareto optimal points that coincide with the convex hull of $\mathcal{R}$; this is illustrated in Figure 1.16(a). The weights are often viewed as the relative priority of different users, but the coupling is complicated and can in general be misleading. First, the notion of priority makes sense in a local area of the performance region, but the global interpretation of the weighting is not easily characterized [216]. This is particularly evident for nonconvex performance regions, because a small perturbation in the weights can greatly affect the optimal operating point; see Figure 1.16(b). Second, the physical setup makes it easier to simultaneously serve spatially separated users (rather than co-located users) and




(a) Not all Pareto points are achievable when maximizing weighted arithmetic means.




(b) Small perturbations in the weights can have large consequences.

Fig. 1.16 Example of maximization of the weighted arithmetic mean $w_1 g_1 + w_2 g_2$ for a nonconvex performance region. The weights $w_1, w_2$ define a line (or hyperplane of dimension $K_r - 1$) that is moved away from the origin until it leaves the performance region; the final intersection with the Pareto boundary gives the optimal operating point. (a) shows that certain points of the Pareto boundary can never be attained by maximizing a weighted arithmetic mean; (b) shows that a small perturbation in the weights can move the optimal solution from one side of the gap to the other side (i.e., from $r_3$ to $r_4$).

thus promotes unbalanced allocation of resources; see further examples on inter-criteria correlation in [258]. Third, the linearity of $f(\cdot)$ implicitly assumes that degrading the performance of one user can be fully compensated by improving for other users, which might not be reasonable in practice [38]. In fact, the *law of diminishing marginal utility* suggests that $f(\cdot)$ should be nonlinear since users become increasingly satisfied with their current performance and less interested in further improvements [223]. Nevertheless, maximizing the weighted arithmetic mean guarantees Pareto optimality and has a simple geometric

interpretation (see Figure 1.16), but the system designer should be aware of the limitations and select the weights carefully.

---

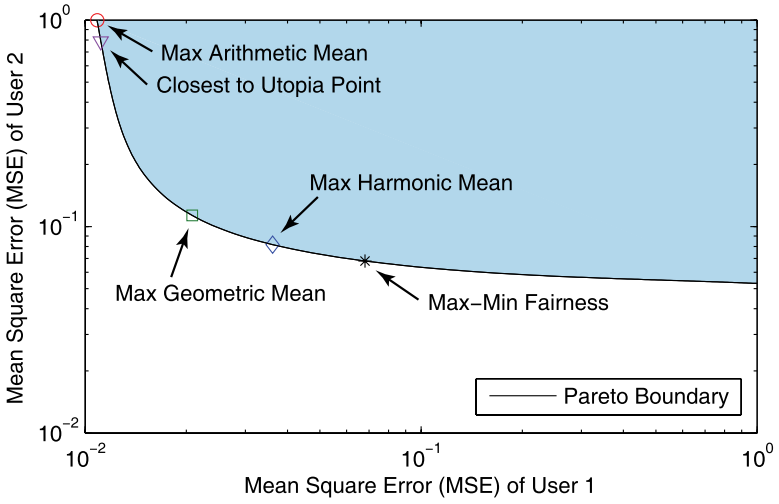
**Remark 1.4 (Game Theoretic Approaches).** Game theory provides an alternative approach to MOPs where the users are seen as players that compete for resources. The game can be formulated in a variety of ways, but the Pareto boundary describes the efficient outcomes for any cooperative game. This approach makes particular sense for ad hoc networks in unlicensed bands and cognitive radio, where there is no joint decision-making and users are indeed competing for spectrum. We refer to [68, 140, 171, 230] and references therein for further details.

---

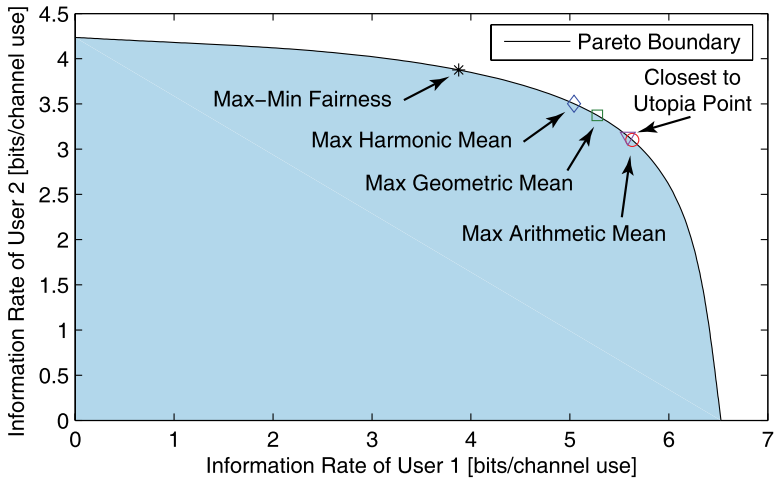
## 1.7 Numerical Examples

In this section, we provide a numerical example that illustrates various concepts defined in this section. We consider a simple scenario with $K_r = 2$ users, $N = 3$ transmit antennas, and global joint transmission (as in Example 1.3). The channel vectors are generated as $\mathbf{h}_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_N)$ (i.e., uncorrelated Rayleigh fading) and we assume per-antenna power constraints with $q_l = 10$ (i.e., 10 dBm). The average single-user SNR $\frac{\mathbb{E}\{q_l \|\mathbf{h}_k\|_2^2\}}{\sigma_k^2}$ is $q_l N$ for User 1 and $q_l \frac{N}{4}$ for User 2, creating an asymmetry that will highlight properties of different system utility functions.

Figure 1.17 shows the performance regions for a single random channel realization for different user performance functions. In Figure 1.17(a), the additive inverse of the MSE is considered (i.e., $g_k(\mathrm{SINR}_k) = \frac{\mathrm{SINR}_k}{1+\mathrm{SINR}_k}$ to make $g_k(0) = 0$), but the figure axes show MSEs to enhance viewing. The information rate $g_k(\mathrm{SINR}_k) = \log_2(1 + \mathrm{SINR}_k)$ is the user performance function in Figure 1.17(b). In both cases, the optimal operating points are shown for the five functions in Example 1.11: arithmetic mean (sum utility), geometric mean (proportional fairness), harmonic mean, max-min fairness, and distance to the utopia point. The weighting factors are $w_1 = w_2 = \frac{1}{2}$.




(a)




(b)

Fig. 1.17 Performance regions for a single channel realizations for different user performance functions: (a) the inverse MSE; and (b) information rate. The Pareto boundary is indicated along with the optimal operating points for different system utility functions.

It is clear that the optimal operating points for these system utility functions are on the Pareto boundary (confirming Lemma 1.10), but at quite different places. As noted in Example 1.11, the arithmetic mean only cares about the aggregate system utility and ignores which user

who gets the performance, while max-min fairness makes sure that all users get exactly the same performance. The geometric mean and harmonic mean are in between these extremes, taking both aggregate system utility and user fairness into account. Searching for the point with the smallest Euclidean distance to the utopia point is similar to maximizing the arithmetic mean. By changing the weighting factors in Example 1.11, the optimal point for a certain type of system utility function can be moved around on the Pareto boundary; in fact, the Pareto boundaries in Figure 1.17 were generated by solving weighted max-min fairness problems over a fine grid of weighting factors.

## 1.8 Summary and Outline

Coordinated multi-cell multi-antenna communication provides an opportunity to increase the system-wide spectral efficiency, as compared to traditional multi-cell setups built on strict interference avoidance. There are many similarities between the single-cell and multi-cell downlink, which can be utilized to bring insights from one case to the other. However, there are also important differences that need to be modeled and managed properly. In this tutorial, we defined a general system model based on *dynamic cooperation clusters* and arbitrary linear power constraints. The main idea behind such clusters is that each base station coordinates interference to exactly those users whom it causes non-negligible interference, while only sending data to a subset of them. As exemplified in this section, this framework can jointly describe many important multi-cell scenarios, including the Wyner model, interference channel, coordinated beamforming, global joint transmission, cognitive radio, and spectrum sharing.

The user performance depends on functions of the SINRs (e.g., information rate, MSE, or error probability), which in turn depends on the selection of signal correlation matrices. Each signal correlation matrix will generally affect all users, which can be illustrated by channel gain regions. These regions were proved to be convex and compact, and the upper boundaries in different directions represent maximization of the received signal power at different users. The joint selection of signal correlation matrices is called resource allocation and can be formulated as

a multi-objective optimization problem. There is not a single solution to such a problem, but many possible tradeoffs between maximizing performance for individual users and maximizing the aggregate utility of the whole system. This tradeoff is illustrated by the performance region $\mathcal{R}$, which was proved to be compact and normal. The Pareto boundary of $\mathcal{R}$ contains all resource allocations that can be regarded optimal. Furthermore, it was shown that all Pareto optimal points can be achieved using single-stream beamforming and optimality conditions for using full transmit power was derived.

To solve the multi-objective resource allocation problem it is necessary to conclude which Pareto optimal points that are preferable for the system. There are different categories of methods and most of them include the selection of a system utility function that assigns a value to each point in the performance region indicating the subjective preference of the system designer. This function can, for example, be the sum utility or max-min fairness. This scalarizes the multi-objective problem to a single-objective problem with a single solution.

### 1.8.1 Outline

Section 2 shows how to solve any single-objective optimization problem. It becomes clear that some problem formulations enable practically efficient algorithms while others can only be optimally solved for offline benchmarking. Section 3 reduces the search-space by parameterizing the optimal transmit strategies and thereby characterizing the Pareto boundary. Section 3 also provides guidelines for formulating and solving multi-objective resource allocation problem in computationally efficient manners.

Finally, Section 4 generalizes the system model to include practical nonidealities, such as CSI uncertainty, hardware impairments, and limited backhaul signaling. It will be shown which results on optimal resource allocation in Sections 2 and 3 that can be easily generalized, and which become intractable. The design of dynamic cooperation clusters and multi-cell scheduling is also discussed. Furthermore. we describe extensions to multi-cast transmission, multi-carrier systems, multi-antenna users, cognitive radio, and physical layer security.

# 2

# Optimal Single-Objective Resource Allocation

The purpose of this section is to provide a systematic framework for solving single-objective resource allocation problems, under the general multi-cell system model defined in Section 1. Recall that this optimization problem was formulated in (1.35) as

$$\begin{aligned} &\underset{\mathbf{v}_1,\ldots,\mathbf{v}_{K_r}}{\text{maximize}} \; f\left(g_1(\text{SINR}_1),\ldots,g_{K_r}(\text{SINR}_{K_r})\right) \\ &\text{subject to } \text{SINR}_k = \frac{|\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_k \mathbf{v}_k|^2}{\sigma_k^2 + \sum\limits_{i \neq k} |\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_i \mathbf{v}_i|^2} \quad \forall k, \\ &\qquad \sum_{k=1}^{K_r} \mathbf{v}_k^H \mathbf{Q}_{lk} \mathbf{v}_k \leq q_l \quad \forall l. \end{aligned} \tag{2.1}$$

The user performance functions $g_k(\cdot)$ are continuous and strictly monotonically increasing, while the system utility function $f(\cdot)$ is Lipschitz continuous and monotonically increasing.

In the process of finding the globally optimal solution to (2.1), Section 2.1 provides some basic results from optimization theory, including classification of optimization problems and Lagrange multiplier theory. Next, Section 2.2 presents some important special cases

when (2.1) is convex and can be solved efficiently. Section 2.3 describes two systematic algorithms for solving any problem of the form in (2.1) with guaranteed convergence to the global optimum. These iterative algorithms originate from the monotonic optimization literature in [218, 274, 275] and utilize the special cases in Section 2.2 to achieve efficient subproblems. Finally, Section 2.4 illustrates the large differences in computational complexity for solving different instances of (2.1). Matlab code for some of the algorithms developed in this section is available for download in [19].

## 2.1 Introduction to Single-Objective Optimization Theory

This section reviews some basic terminology and results in optimization theory, and exemplify their impact on the resource allocation problem in (2.1). These results are utilized throughout of this tutorial.

Consider a *single-objective optimization problem (SOP)* that can be expressed as

$$\begin{aligned} &\underset{\mathbf{x}}{\text{minimize}} \ f_0(\mathbf{x}) \\ &\text{subject to } \mathbf{x} \in \mathcal{X}, \end{aligned} \tag{2.2}$$

where $\mathbf{x} \in \mathbb{R}^n$ is called the *optimization variable* and belongs to the closed *feasible set* $\mathcal{X}$. The feasible set is a subset of some box $[\mathbf{a}, \mathbf{b}] \subseteq \mathbb{R}^n$ that we assume to be compact. The function $f_0 : \mathbb{R}^n \to \mathbb{R}$ is the *cost function* and is assumed to be continuously differentiable over $[\mathbf{a}, \mathbf{b}]$. A feasible vector $\mathbf{x}^* \in \mathcal{X}$ is called an *optimal solution* to (2.2) if it provides the smallest value (called the *optimal value*), $f_0(\mathbf{x}^*)$, on the cost function among all $\mathbf{x} \in \mathcal{X}$. If the feasible set is empty (i.e., $\mathcal{X} = \emptyset$), the optimal value is conventionally set to $+\infty$.

To enable analysis and numerical computations, it is often more convenient to write the SOP on *standard form* as

$$\begin{aligned} &\underset{\mathbf{x}}{\text{minimize}} \ f_0(\mathbf{x}) \\ &\text{subject to } f_m(\mathbf{x}) \leq 0 \quad m = 1, \ldots, M, \end{aligned} \tag{2.3}$$

where the $M$ functions $f_m : \mathbb{R}^n \to \mathbb{R}$ are the *constraint functions*. Any constrained SOP can be rewritten on standard form [37] (but the

dimension of $\mathbf{x}$ might change) and (2.3) is equivalent to (2.2) if we set

$$\mathcal{X} = \left\{\mathbf{x} \in \mathbb{R}^n \colon f_m(\mathbf{x}) \leq 0 \quad m = 1, \ldots, M\right\}. \tag{2.4}$$

---

**Remark 2.1 (Maximization).** The SOP on standard form considers minimization of a cost function $f_0$, but this is equivalent to maximization of the additive inverse $-f_0$ under identical constraints.

---

**Example 2.1 (Resource Allocation on Standard Form).** The resource allocation problem in (2.1) can be expressed as

$$\begin{aligned} &\underset{\mathbf{g}}{\text{minimize}} \;\; -f(\mathbf{g}) \\ &\text{subject to} \;\; \mathbf{g} \in \mathcal{R} \end{aligned} \tag{2.5}$$

where the optimization variable $\mathbf{g} = [g_1(\text{SINR}_1) \ldots g_{K_r}(\text{SINR}_{K_r})]^T$ represents the user performance, $-f(\mathbf{g})$ is the cost function, and the performance region $\mathcal{R}$ equals the feasible set. This formulation shows that resource allocation means searching $\mathcal{R}$ for the vector that optimizes system utility.

To achieve a formulation on standard form, denote the concatenation of all beamforming vectors as $\mathbf{v} = [\mathbf{v}_1^T \ldots \mathbf{v}_{K_r}^T]^T \in \mathbb{C}^{NK_r}$ and let $\mathbf{x} = \begin{bmatrix} \Re(\mathbf{v}) \\ \Im(\mathbf{v}) \end{bmatrix}$ be the optimization variable. The cost function is $f_0(\mathbf{x}) = -f\left(g_1(\text{SINR}_1), \ldots, g_{K_r}(\text{SINR}_{K_r})\right)$ and observe that $\text{SINR}_k$ is a function of $\mathbf{x}$. The constraints are given by $f_l(\mathbf{x}) = \mathbf{x}^H \begin{bmatrix} \Re(\mathbf{Q}_l) & -\Im(\mathbf{Q}_l) \\ \Im(\mathbf{Q}_l) & \Re(\mathbf{Q}_l) \end{bmatrix} \mathbf{x} - q_l$ for $l = 1, \ldots, L$, where $\mathbf{Q}_l = \text{diag}(\mathbf{Q}_{l1}, \ldots, \mathbf{Q}_{lK_r})$.

---

### 2.1.1 Classification and Computational Complexity

The standard form in (2.3) provides a compact way of representing any SOP, but additional information is required to analyze the problem and devise suitable numerical algorithms. Fortunately, it is not necessary to build the analysis from scratch for any set of cost function and constraint functions, but there are some important classes of problems where certain numerical algorithms can be applied to solve any instance of the class [10, 12, 37, 274]. Some important classes are now defined.

**Definition 2.1.** A SOP on standard form is called a

- *linear problem* if $f_0, \ldots, f_M$ are linear/affine functions.[1] The feasible set $\mathcal{X}$ becomes a convex polytope in $\mathbb{R}^n$.
- *convex problem* if $f_0, \ldots, f_M$ are convex functions.[2] The feasible set $\mathcal{X}$ becomes a convex set in $\mathbb{R}^n$.
- *quasi-convex problem* if $f_0, \ldots, f_M$ are quasi-convex functions.[3] The feasible set $\mathcal{X}$ becomes a convex set in $\mathbb{R}^n$.
- *monotonic problem* if $f_0, \ldots, f_M$ are monotonic functions (any combination of increasing and decreasing functions). The feasible set $\mathcal{X}$ becomes a mutually normal set.[4]

These four classes represent successively more general conditions: every linear problem is also convex, every convex problem is also quasi-convex, and every quasi-convex problem is also monotonic.[5] This relationship is illustrated in Figure 2.1. Practical optimization problems could be difficult to classify in this way and reformulations are sometimes necessary to reveal a hidden underlying structure. The authors of [196] note that there is no systematic way of identifying and extracting an underlying structure, but it is rather an art that includes making good changes of variables and relaxations. Examples of such reformulations are found in [11, 29, 167, 168, 296].

Most optimization problems have no closed-form solutions, but can still be solved numerically (to any accuracy $\varepsilon > 0$ on the optimal value).

[1] A function $f_m : \mathbb{R}^n \to \mathbb{R}$ is called *affine* on $[\mathbf{a}, \mathbf{b}]$ if for any $\mathbf{x}_1, \mathbf{x}_2 \in [\mathbf{a}, \mathbf{b}]$ and $t \in [0, 1]$, $f_m(t\mathbf{x}_1 + (1-t)\mathbf{x}_2) = tf_m(\mathbf{x}_1) + (1-t)f_m(\mathbf{x}_2)$.

[2] A function $f_m : \mathbb{R}^n \to \mathbb{R}$ is called *convex* on $[\mathbf{a}, \mathbf{b}]$ if for any $\mathbf{x}_1, \mathbf{x}_2 \in [\mathbf{a}, \mathbf{b}]$ and $t \in [0, 1]$, $f_m(t\mathbf{x}_1 + (1-t)\mathbf{x}_2) \leq tf_m(\mathbf{x}_1) + (1-t)f_m(\mathbf{x}_2)$.

[3] A function $f_m : \mathbb{R}^n \to \mathbb{R}$ is called *quasi-convex* on $[\mathbf{a}, \mathbf{b}]$ if for any $\mathbf{x}_1, \mathbf{x}_2 \in [\mathbf{a}, \mathbf{b}]$ and $t \in [0, 1]$, $f_m(t\mathbf{x}_1 + (1-t)\mathbf{x}_2) \leq \max(f_m(\mathbf{x}_1), f_m(\mathbf{x}_2))$.

[4] A set $\mathcal{S}$ is *mutually normal* on $[\mathbf{a}, \mathbf{b}]$ if it can be written as $\mathcal{S} = \mathcal{T}_1 \cap ([\mathbf{a}, \mathbf{b}] \backslash \mathcal{T}_2)$ for two normal sets $\mathcal{T}_1, \mathcal{T}_2$ on $[\mathbf{a}, \mathbf{b}]$. The relative complement $[\mathbf{a}, \mathbf{b}] \backslash \mathcal{T}_2$ is called a *conormal set*.

[5] Quasi-convex functions are not necessarily monotonic, thus it is not trivial to see that any quasi-convex problem is also a monotonic problem. However, a quasi-convex function can be written as the difference of two monotonically increasing functions [275], which is rather straightforward to rewrite as a monotonic problem on standard form; see [274].

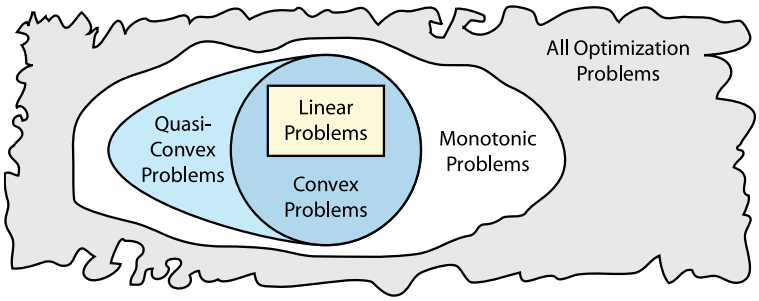



Fig. 2.1 Classification of single-objective optimization problems in Definition 2.1. Linear, convex, quasi-convex, and monotonic problems have successively more general conditions on the functions $f_0, \ldots, f_M$.

Classification of a problem enables the use of numerical algorithms designed for this class. For example, linear problems can be solved very efficiently by the *simplex method* [136]. This method has an average-case computational complexity that only grows polynomially with the problem size (e.g., the number of variables $n$ and number of constraints $M$), but the worst-case complexity is exponential. *Interior-point methods* can be applied to both linear and convex problems with a polynomial worst-case complexity (at least under mild conditions such as self-concordance [37]). General-purpose implementations of interior-point methods are available in `SeDuMi` [256] and `SDPT3` [271]. The use of these implementations can be simplified by the high-level modeling languages `CVX` [95] and `Yalmip` [161]. These implementations are particularly good at solving convex problems with second-order cone constraints [160] and semi-definite constraints, whereof the former is particularly important in this section.

---

**Example 2.2 (Second-Order Cone Constraint).** A *second-order cone constraint* is given by

$$f_m(\mathbf{x}) = \|\mathbf{A}_m\mathbf{x} + \mathbf{b}_m\|_2 + \mathbf{c}_m^T\mathbf{x} + d_m \tag{2.6}$$

and is convex for any positive integer $n_m$ and parameters $\mathbf{A}_m \in \mathbb{R}^{n_m \times n}$, $\mathbf{b}_m \in \mathbb{R}^{n_m}$, $\mathbf{c}_m \in \mathbb{R}^n$, and $d_m \in \mathbb{R}$.

The power constraints for the resource allocation problem in (2.1) can be written as second-order cones

$$f_l(\mathbf{x}) = \left\| \begin{bmatrix} \Re(\mathbf{Q}_l) & -\Im(\mathbf{Q}_l) \\ \Im(\mathbf{Q}_l) & \Re(\mathbf{Q}_l) \end{bmatrix}^{1/2} \mathbf{x} \right\|_2 - \sqrt{q_l} \tag{2.7}$$

for the optimization variable $\mathbf{x} = \begin{bmatrix} \Re(\mathbf{v}) \\ \Im(\mathbf{v}) \end{bmatrix}$ (see Example 2.1).

---

It is important to differentiate between *globally optimal points* $\mathbf{x}^*$ (minimizing the cost in $\mathcal{X}$) and *locally optimal points* that provide the lowest cost among the feasible points in their immediate surroundings.[6] As noted by Rockafellar in [213], there is a great watershed between convex problems and nonconvex problems; every locally optimal solution to a convex problem is also globally optimal, while this is not the case for general nonconvex problems [37].[7] Therefore, the entire feasible set $\mathcal{X}$ basically needs to be searched when solving nonconvex problems, which corresponds to a complexity that grows exponentially with the problem size. Practical algorithms for nonconvex problems are typically designed to only search for locally optimal points, which might be achieved with manageable complexity.

In terms of complexity, quasi-convex problems actually belong to category of convex problems, because these can be solved by a limited sequence of convex subproblems [37, Subsection 4.2.5]. General monotonic problems have however exponential worst-case complexity, but we can avoid searching the entire feasible set by utilizing the monotonicity; if $f_0$ is monotonically decreasing and $\bar{\mathbf{x}}$ is found to be a feasible point, then any $\mathbf{x} \leq \bar{\mathbf{x}}$ provides higher cost and can be immediately discarded. The area of monotonic optimization is relatively new, although monotonicity constraints (e.g., free disposability) have appeared in economical applications for a long time [192]. In the early 2000s, Tuy et al. proposed two iterative algorithms that utilize monotonicity when solving

[6] Formally, a point $\bar{\mathbf{x}}$ is called locally optimal if there exist $\epsilon > 0$ such that $f_0(\bar{\mathbf{x}}) \leq f_0(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$ satisfying $\|\bar{\mathbf{x}} - \mathbf{x}\|_2 < \epsilon$.

[7] In addition, infeasibility of convex problems is easily detected (e.g., using the dual function defined in Subsection 2.1.2), while infeasibility might be difficult to detect for general nonconvex problems [167]. The resource allocation problem in (2.1) fortunately has second-order cone constraints and will (almost) always be feasible.

monotonic problems: the *polyblock outer approximation (PA) algorithm* in [218, 274] and the *branch-reduce-and-bound (BRB) algorithm* in [275]. These algorithms have exponential worst-case complexity, but provide a structured approach that (at least) can solve small problems.

This section will show that the multi-cell resource allocation problem in (2.1) is linear, convex, quasi-convex, or monotonic depending on the scenario. As convex problems are easily implemented and solved using general-purpose implementations of interior-point methods (as mentioned earlier), for each scenario we either show how to reformulate (2.1) into a convex problem or give algorithms that solve it as a sequence of convex problems. To this end, we first review some basic results on duality, bounding of the optimal value, and necessary (and sometimes sufficient) conditions on the optimal solution.

---

**Remark 2.2 (Complex-Valued Optimization Variables).** The literature on optimization theory usually considers real-valued optimization variables $\mathbf{x}$, but most results can be readily extended to complex-valued variables $\mathbf{x} \in \mathbb{C}^n$ if the cost and constraint functions are defined as $f_m : \mathbb{C}^n \to \mathbb{R}$ for $m = 0, \ldots, M$. Observe that any complex-valued scalar $c$ can be described by the two real-valued scalars $\Re(c), \Im(c)$, thus problems with complex-valued variables can be rewritten on standard form, for example using the rule $\mathbf{x}^H \mathbf{A} \mathbf{x} = \begin{bmatrix} \Re(\mathbf{x}) \\ \Im(\mathbf{x}) \end{bmatrix}^H \begin{bmatrix} \Re(\mathbf{A}) & -\Im(\mathbf{A}) \\ \Im(\mathbf{A}) & \Re(\mathbf{A}) \end{bmatrix} \begin{bmatrix} \Re(\mathbf{x}) \\ \Im(\mathbf{x}) \end{bmatrix}$ for Hermitian matrices $\mathbf{A} \in \mathbb{C}^{n \times n}$. However, such reformulations are often unnecessary because the definitions of linear, convex, and quasi-convex problems (see Definition 2.1) are applicable also for complex-valued variables. The modeling languages `CVX` and `Yalmip` also handle such variables.

---

### 2.1.2 Lagrange Multiplier Theory

Lagrange multiplier theory provides useful tools to analyze, bound, and solve optimization problems on standard form. In particular, it gives optimality conditions for identifying potential solutions to constrained optimization problems. These conditions generalize a well-known result in unconstrained optimization, namely that the global minimum $\bar{\mathbf{x}}$ of

$f(\mathbf{x})$ satisfies $\nabla f(\bar{\mathbf{x}}) = \mathbf{0}$. This subsection reviews concepts and results that are utilized in this tutorial, while further details and proofs are available in [37, Chapter 5].

---

**Definition 2.2 (Lagrangian).** The *Lagrangian function* $\mathcal{L} : [\mathbf{a}, \mathbf{b}] \times \mathbb{R}^M \to \mathbb{R}$ associated with (2.3) is

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = f_0(\mathbf{x}) + \sum_{m=1}^{M} \lambda_m f_m(\mathbf{x}). \tag{2.8}$$

The *Lagrange multiplier* $\lambda_m$ is associated with the $m$th constraint and the vector $\boldsymbol{\lambda} = [\lambda_1 \dots \lambda_M]^T$ is the *Lagrange multiplier vector* for (2.3).

The *Lagrange dual function* $h : \mathbb{R}^M \to \mathbb{R}$ is the minimum value of the Lagrangian function over $\mathbf{x}$,

$$h(\boldsymbol{\lambda}) = \inf_{\mathbf{x} \in [\mathbf{a}, \mathbf{b}]} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}). \tag{2.9}$$

---

The idea behind the Lagrangian function is to augment the cost function $f_0(\mathbf{x})$ with the constraints, such that constraint violations are penalized with an increased cost. Since the constraints are to be fulfilled, the simplest approach would be to let the cost become infinite when outside the feasible set. Such hard penalization stands in contrast to the soft penalization in $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})$, where a constraint violation is weighted linearly by its corresponding Lagrange multiplier.

Observe that the dual function is the pointwise infimum of a family of affine functions of $\boldsymbol{\lambda}$, thus it is concave even if (2.3) is a nonconvex problem. On the other hand, it might be difficult to compute the infimum, which is necessary to explicitly derive the dual function.

The dual function provides a bound on the optimal value.

---

**Lemma 2.1.** The dual function yields lower bounds on the optimal value of (2.3). For any $\boldsymbol{\lambda} \geq \mathbf{0}$ we have

$$h(\boldsymbol{\lambda}) \leq f_0(\mathbf{x}^*). \tag{2.10}$$

---

*Proof.* Based on [37, Subsection 5.1.3], suppose $\bar{\mathbf{x}}$ is a feasible vector for (2.3) (i.e., $f_m(\bar{x}) \leq 0 \, \forall m$) and observe that $\sum_{m=1}^{M} \lambda_m f_m(\mathbf{x}) \leq 0$, for any $\boldsymbol{\lambda} \geq \mathbf{0}$, since all terms are nonpositive. As a result,

$$h(\boldsymbol{\lambda}) = \inf_{\mathbf{x} \in [\mathbf{a}, \mathbf{b}]} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) \leq \mathcal{L}(\bar{\mathbf{x}}, \boldsymbol{\lambda}) \leq f_0(\bar{\mathbf{x}}) \tag{2.11}$$

for all feasible points $\bar{\mathbf{x}} \in [\mathbf{a}, \mathbf{b}]$, including the optimal solutions. □

Lemma 2.1 provides a lower bound on the optimal solution of (2.3) that holds for any feasible choice of Lagrange multipliers, thus the closest lower bound is obtained by maximizing the lower bound.

---

**Definition 2.3 (Lagrange Dual Problem).** The *Lagrange dual problem* associated with (2.3) is

$$\begin{aligned} &\underset{\boldsymbol{\lambda}}{\text{maximize}} \; h(\boldsymbol{\lambda}) \\ &\text{subject to } \boldsymbol{\lambda} \geq \mathbf{0}. \end{aligned} \tag{2.12}$$

In this context, the original problem in (2.3) is called the *primal problem*. The optimal vector of the dual problem is denoted $\boldsymbol{\lambda}^*$.

---

Interestingly, the Lagrange dual problem in (2.12) is always a convex optimization problem, since the objective to be maximized is concave and the constraint is convex. This is independent of whether the primal problem in (2.3) is convex or not. On the other hand, the dual function is not necessarily differentiable.

### 2.1.3 Optimality Conditions and Strong Duality

There are many important connections between the optimal solution $\mathbf{x}^*$ of the primal problem and the Lagrange multiplier vector $\boldsymbol{\lambda}$. Particularly the *Karush–Kuhn–Tucker conditions (KKT conditions)* can be used to identify solution candidates.

---

**Definition 2.4 (KKT Conditions).** Let $\mathbf{x}^*$ be an optimal solution to the primal problem (2.3). The *KKT conditions* say that there exist

a unique Lagrange multiplier vector $\boldsymbol{\lambda}^*$ such that

$$\nabla f_0(\mathbf{x}^*) + \sum_{m=1}^{M} \lambda_m^* \nabla f_m(\mathbf{x}^*) = \mathbf{0}, \tag{2.13}$$

$$f_m(\mathbf{x}^*) \leq 0 \quad m = 1, \ldots, M, \tag{2.14}$$

$$\lambda_m^* \geq 0 \quad m = 1, \ldots, M, \tag{2.15}$$

$$\lambda_m^* f_m(\mathbf{x}^*) = 0 \quad m = 1, \ldots, M. \tag{2.16}$$

These conditions are known as *stationarity*, *primal feasibility*, *dual feasibility*, and *complementary slackness*, respectively.

These conditions are generally neither sufficient nor necessary for the optimal solution. The extra conditions for becoming necessary are known as *constraint qualifications* and typically require some kind of linear independence among gradients of the active constraints; see [12]. The following simple condition is sufficient under convex constraints.

**Lemma 2.2 (Slater's Constraint Qualification).** If all constraint functions $f_m(\mathbf{x})$ are convex and it exists $\mathbf{x} \in [\mathbf{a}, \mathbf{b}]$ such that $f_m(\mathbf{x}) < 0$ for all nonaffine constraints, then the KKT conditions are necessary for the corresponding optimization problem.

This lemma originates from [249] and we use the formulation in [12, Chapter 3]. Only the constraints need to be convex to satisfy Slater's constraint qualification, thus we have the following result.

**Example 2.3 (KKT Conditions in Resource Allocation).** The resource allocation problem in (2.1) has convex constraints (see Example 2.2). Suppose all beamforming vectors are zero, $\mathbf{v}_k = \mathbf{0}\,\forall k$, then $f_l(\mathbf{0}) < 0$ for all power constraints with $q_l > 0$. In addition, all constraints with $q_l = 0$ can be reformulated as affine equality constraints $\mathbf{Q}_{lk}^{1/2}\mathbf{v}_k = \mathbf{0}$. Therefore, the power constraints satisfy Slater's constraint qualification and the KKT conditions are necessary for all tentative solutions to (2.1).

The KKT conditions are broadly related to the property of strong duality, as will be shown below. Observe that the optimal value of the dual problem in (2.12) is always smaller than or equal to the value of the primal problem in (2.3), thus

$$h(\boldsymbol{\lambda}^*) \leq f_0(\mathbf{x}^*). \tag{2.17}$$

Equality would mean that the best bound obtained from the Lagrange dual function is tight, but equality is generally not achieved.

---

**Definition 2.5 (Strong Duality).** The difference $f_0(\mathbf{x}^*) - h(\boldsymbol{\lambda}^*)$ is the *optimal duality gap* and is always nonnegative. The case when the optimal duality gap is zero is called *strong duality*.

---

The dual problem provides the optimal value of the primal problem under strong duality, giving an alternative way of solving the primal problem. Strong duality also makes the KKT conditions necessary.

---

**Lemma 2.3 (KKT Conditions under Strong Duality).** If strong duality holds, then the KKT conditions are necessary for the optimal solution of the corresponding optimization problem.

---

*Proof.* Suppose that strong duality holds, let $\mathbf{x}^*$ be a primal optimal solution, and let $\boldsymbol{\lambda}^*$ be a dual optimal solution. This means that

$$\begin{aligned} f_0(\mathbf{x}^*) = h(\boldsymbol{\lambda}^*) &= \inf_{\mathbf{x}\in[\mathbf{a},\mathbf{b}]} \left( f_0(\mathbf{x}) + \sum_{m=1}^{M} \lambda_m^* f_m(\mathbf{x}) \right) \\ &\leq f_0(\mathbf{x}^*) + \sum_{m=1}^{M} \lambda_m^* f_m(\mathbf{x}^*) \leq f_0(\mathbf{x}^*). \end{aligned} \tag{2.18}$$

The two inequalities must hold with equality, thus it follows that $\mathbf{x}^*$ minimizes $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}^*)$ and the gradient is zero at $\mathbf{x}^*$:

$$\nabla f_0(\mathbf{x}^*) + \sum_{m=1}^{M} \lambda_m^* \nabla f_m(\mathbf{x}^*) = \mathbf{0}. \tag{2.19}$$

In addition, we have $\sum_{m=1}^{M} \lambda_m^* f_m(\mathbf{x}^*) = 0$ and since $\lambda_m \geq 0$ it follows that

$$\lambda_m^* f_m(\mathbf{x}^*) = 0 \quad m = 1, \ldots, M. \tag{2.20}$$

The combination of primal feasibility of $\mathbf{x}^*$, dual feasibility of $\boldsymbol{\lambda}^*$, (2.19), and (2.20) is exactly the KKT conditions. □

For convex problems, KKT conditions and strong duality are particularly important as these often are both sufficient and necessary.

---

**Lemma 2.4 (KKT Conditions for Convex Problems).** If the cost function is convex and Slater's constraint qualification is satisfied, then strong duality holds and the KKT conditions are both necessary and sufficient for the optimal solution.

---

This lemma provides a simple way to prove strong duality for convex problems before actually solving the problem — this is why problems in this category can be solved relatively efficiently. Strong duality can also be shown to hold for certain nonconvex problems, but it generally requires numerical computation of the optimal duality gap.

---

**Remark 2.3 (Saddle Point Interpretation).** Strong duality can be interpreted as the existence of a saddle point in the Lagrangian function, meaning that

$$\sup_{\boldsymbol{\lambda} \succeq \mathbf{0}} \inf_{\mathbf{x} \in [\mathbf{a}, \mathbf{b}]} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = \inf_{\mathbf{x} \in [\mathbf{a}, \mathbf{b}]} \sup_{\boldsymbol{\lambda} \succeq \mathbf{0}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}). \tag{2.21}$$

This equivalence holds under certain properties on $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})$, for example, if $\mathcal{L}$ is convex in $\mathbf{x}$ and lower semi-continuous for every $\boldsymbol{\lambda} \succeq \mathbf{0}$ and $\mathcal{L}$ is also concave in $\boldsymbol{\lambda}$ and upper semi-continuous for every $\mathbf{x} \in [\mathbf{a}, \mathbf{b}]$; see [70, Theorem 1] for some general conditions.

---

## 2.2 Convex Optimization for Resource Allocation

In this section, we investigate under which conditions the single-objective resource allocation problem in (2.1) is linear, convex, or

quasi-convex. Recall that these classes of problems can be solved efficiently (e.g., using interior-point methods [256, 271]).

The problem (2.1) has convex constraints (see Example 2.2). Therefore, the classification strongly depends on the cost function $-f\left(g_1(\mathrm{SINR}_1),\ldots,g_{K_r}(\mathrm{SINR}_{K_r})\right)$, which unfortunately is a complicated function that seems nonconvex; $f(\cdot)$ depends on the SINRs which in turn are nonconvex functions of the beamforming vectors $\mathbf{v}_1,\ldots,\mathbf{v}_{K_r}$. To pinpoint the main cause of nonconvexity, we represent the SINRs by auxiliary optimization variables $\gamma_k$ such that $\gamma_k = \mathrm{SINR}_k$. We then rewrite (2.1) as

$$\begin{aligned} \underset{\mathbf{v}_k,\gamma_k\,\forall k}{\text{minimize}} \quad & -f\left(g_1(\gamma_1),\ldots,g_{K_r}(\gamma_{K_r})\right) \\ \text{subject to} \quad & |\mathbf{h}_k^H\mathbf{C}_k\mathbf{D}_k\mathbf{v}_k|^2 \geq \gamma_k\Big(\sigma_k^2 + \sum_{i\neq k}|\mathbf{h}_k^H\mathbf{C}_k\mathbf{D}_i\mathbf{v}_i|^2\Big) \quad \forall k, \\ & \sum_{k=1}^{K_r}\mathbf{v}_k^H\mathbf{Q}_{lk}\mathbf{v}_k \leq q_l \quad \forall l. \end{aligned} \tag{2.22}$$

The second row of (2.22) represents the auxiliary SINR constraints $\gamma_k \leq \mathrm{SINR}_k$ and the optimal solution always gives equality in these constraints. The main complication lies in the SINR constraints, because $-f\left(g_1(\gamma_1),\ldots,g_{K_r}(\gamma_{K_r})\right)$ is a convex function with respect to $\gamma_1,\ldots,\gamma_{K_r}$ for many $f(\cdot)$ and $g_k(\cdot)$ of practical interest.

---

**Example 2.4 (Some Convex and Concave Functions).** A continuous twice differentiable function is convex (concave) if the second-order derivative is nonnegative (nonpositive). For functions of several variables, this extends to a positive (negative) semi-definite Hessian.

Some typical convex functions are $x^2$, $e^x$, and $-\log_2(x)$.

Some typical concave functions are $\log_2(x)$, $-e^x$, and $\sqrt{x}$.

Linear functions, such as $x$ and $-x$, are both convex and concave.

---

**Example 2.5 (Concavity of Performance Functions).** The information rate and the MSE are concave user performance functions (see Examples 1.6 and 1.7), which is easily seen from the nonpositive second-order derivatives. The BER for $M$-QAM with $M \in \{4, 16, 64, 256\}$

also gives concave performance functions [189] (see Example 1.8 for $M = 16$). The same holds for the symbol error rate (SER) under arbitrary constellations, while the BER and pairwise error probability (PEP) are only guaranteed to be concave at high SINR; see [163]. Sigmoid functions can describe certain application-oriented utilities [145] and are only concave if the SINR exceeds a certain threshold.

All system utility functions in Example 1.11 are concave functions (e.g., arithmetic/geometric/harmonic mean).[8] In fact, the so-called *law of diminishing marginal utility* suggests that all system utility functions are concave [223], because users generally become less interested in further improvements as their performance increases. The composite function $f(g_1(\gamma_1), \ldots, g_{K_r}(\gamma_{K_r}))$ is concave with respect to $\gamma_1, \ldots, \gamma_{K_r}$ whenever both $f(\cdot)$ and $g_k(\cdot)$ are concave for all $k$.

---

In other words, it is generally the SINR constraints that prevent (2.22) from being a convex problem. These constraints are nonconvex because of the multiplication between $\gamma_k$ (the SINR value at $\text{MS}_k$) and $\sum_{i \neq k} |\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_i \mathbf{v}_i|^2$ (the inter-user interference caused to $\text{MS}_k$). Three approaches to resolve the non-convexity can be envisioned:

(1) Fix the inter-user interference caused to each user;
(2) Fix the SINR value at each user;
(3) Turn the multiplication into addition by change of variables.

None of these approaches can be applied successfully to any resource allocation problem, but they will help identifying special cases when (2.1) has a hidden convex structure and thus can be solved efficiently. The division between convex and nonconvex resource allocation problems is illustrated in Figure 2.2. The special cases with convexity are interesting and useful on their own, but will also be used as subproblems when solving general nonconvex resource allocation problems in Section 2.3.

---

[8] To exploit the inherent concavity, it might be necessary to reformulate $f(\mathbf{g})$ into an equivalent form; the weighted geometric mean should have exponents greater than one, the maximization of the weighted harmonic mean is equivalent to maximizing $f(\mathbf{g}) = -\left(\sum_k \frac{w_k}{g_k}\right)$, and the exponent $1/p$ can be dropped for the weighted compromise.

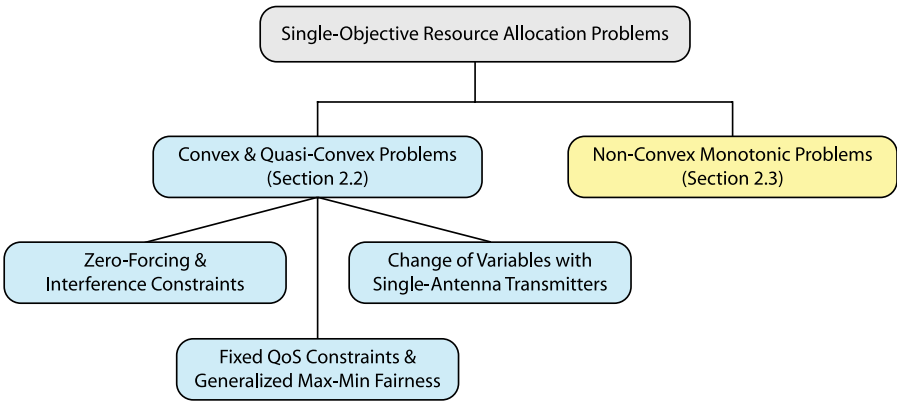



Fig. 2.2 The division of single-objective resource allocation between convex and nonconvex problems. Three types of convex problem formulations are described in this section, based on fixing the inter-user interference, fixing the SINR at each user, or changing variables.

### 2.2.1 Zero-Forcing and Interference Constraints

This subsection will show that the resource allocation problem becomes convex if the power of the inter-user interference is known *a priori*. An important special case is so-called *zero-forcing beamforming*[9] [23, 46, 85, 115, 252, 297, 305], where the beamforming vectors are selected to cause zero interference to nonintended users. This condition greatly simplifies the beamforming design by reducing the search-space (i.e., beamforming vectors should lie in the nullspace of the co-user channels), but has also practical importance in cognitive radio (see Section 4.8) and in high-SNR scenarios where inter-user interference greatly dominates the noise term in the SINR expression.

Zero-forcing can be relaxed into *interference-constrained beamforming* [26, 143, 215, 325] where the inter-user interference at $\text{MS}_k$ is not nulled but should be below some threshold $\Gamma_k \geq 0$. This relaxation is

[9] Zero-forcing is also known as *channel inversion* because the goal is to make $\mathbf{H}_{\text{tot}}\mathbf{V}_{\text{tot}}$ a diagonal matrix, where $\mathbf{H}_{\text{tot}} = [\mathbf{C}_1^H\mathbf{h}_1 \ldots \mathbf{C}_{K_r}^H\mathbf{h}_{K_r}]^H$ is the joint channel matrix and $\mathbf{V}_{\text{tot}} = [\mathbf{D}_1\mathbf{v}_1 \ldots \mathbf{D}_{K_r}\mathbf{v}_{K_r}]$ is the joint beamforming matrix. Under a total power constraint, the diagonalization is achieved by setting $\mathbf{V}_{\text{tot}} = \mathbf{H}_{\text{tot}}^{-1}$. Under general power constraints and flexible power allocation, the channel inverse becomes a generalized inverse [297] and lacks a simple closed-form expression.

reasonable because nulling the interference is usually an overreaction; CSI uncertainty makes it impossible in practice and it is unnecessary to suppress the interference far below the background noise. The corresponding interference constraints are

$$\sum_{i\neq k} |\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_i \mathbf{v}_i|^2 \leq \Gamma_k \quad \forall k \tag{2.23}$$

$$\Leftrightarrow$$

$$\sum_{i\neq k} \mathbf{v}_i^H (\mathbf{D}_i^H \mathbf{C}_k^H \mathbf{h}_k \mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_i) \mathbf{v}_i \leq \Gamma_k \quad \forall k.$$

Observe that this constraint has the same form as the power constraints in (1.4) with $\mathbf{Q}_{li} = \mathbf{D}_i^H \mathbf{C}_k^H \mathbf{h}_k \mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_i$ for $i \neq k$, $\mathbf{Q}_{lk} = \mathbf{0}_N$, and $q_l = \Gamma_k$. This subsection therefore considers the special case when there are $K_r$ interference constraints of the form in (2.23), in addition to the $L$ regular power constraints:

$$\begin{aligned}
&\underset{\mathbf{v}_k, \gamma_k \, \forall k}{\text{minimize}} \quad -f\left(g_1(\gamma_1), \ldots, g_{K_r}(\gamma_{K_r})\right) \\
&\text{subject to } |\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_k \mathbf{v}_k|^2 \geq \gamma_k \Big( \sigma_k^2 + \sum_{i\neq k} |\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_i \mathbf{v}_i|^2 \Big) \quad \forall k, \\
&\qquad \sum_{i\neq k} \mathbf{v}_i^H (\mathbf{D}_i^H \mathbf{C}_k^H \mathbf{h}_k \mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_i) \mathbf{v}_i \leq \Gamma_k \quad \forall k, \\
&\qquad \sum_{k=1}^{K_r} \mathbf{v}_k^H \mathbf{Q}_{lk} \mathbf{v}_k \leq q_l \quad \forall l.
\end{aligned} \tag{2.24}$$

For this problem, the SINR of $\text{MS}_k$ can be lower-bounded as

$$\text{SINR}_k = \frac{|\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_k \mathbf{v}_k|^2}{\sigma_k^2 + \sum\limits_{i\neq k} |\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_i \mathbf{v}_i|^2} \geq \frac{|\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_k \mathbf{v}_k|^2}{\sigma_k^2 + \Gamma_k} \tag{2.25}$$

by replacing the actual interference at $\text{MS}_k$ with the corresponding interference constraint. Observe that all feasible solutions must satisfy (2.25) with equality if $\Gamma_k = 0$, while this is not necessarily the case when $\Gamma_k > 0$ (i.e., it might be optimal to cause less interference than allowed). Using the lower bound in (2.25), the resource allocation problem in (2.24) can be solved as follows.

**Theorem 2.5.** For fixed $\Gamma_1, \ldots, \Gamma_{K_r} \in \mathbb{R}_+$, the optimization problem

$$\begin{aligned} \underset{\mathbf{v}_k, \gamma_k \, \forall k}{\text{minimize}} \quad & -f\left(g_1(\gamma_1), \ldots, g_{K_r}(\gamma_{K_r})\right) \\ \text{subject to} \quad & |\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_k \mathbf{v}_k|^2 \geq \gamma_k(\sigma_k^2 + \Gamma_k) \quad \forall k, \\ & \sum_{i \neq k} \mathbf{v}_i^H (\mathbf{D}_i^H \mathbf{C}_k^H \mathbf{h}_k \mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_i) \mathbf{v}_i \leq \Gamma_k \quad \forall k, \\ & \sum_{k=1}^{K_r} \mathbf{v}_k^H \mathbf{Q}_{lk} \mathbf{v}_k \leq q_l \quad \forall l \end{aligned} \tag{2.26}$$

is solved by considering the semi-definite relaxation with $\mathbf{S}_k = \mathbf{v}_k \mathbf{v}_k^H$:

$$\begin{aligned} \underset{\mathbf{S}_k \succeq \mathbf{0}_N, \gamma_k \, \forall k}{\text{minimize}} \quad & -f\left(g_1(\gamma_1), \ldots, g_{K_r}(\gamma_{K_r})\right) \\ \text{subject to} \quad & \mathrm{tr}(\mathbf{D}_k^H \mathbf{C}_k^H \mathbf{h}_k \mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_k \mathbf{S}_k) \geq \gamma_k(\sigma_k^2 + \Gamma_k) \quad \forall k, \\ & \sum_{i \neq k} \mathrm{tr}(\mathbf{D}_i^H \mathbf{C}_k^H \mathbf{h}_k \mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_i \mathbf{S}_i) \leq \Gamma_k \quad \forall k, \\ & \sum_{k=1}^{K_r} \mathrm{tr}(\mathbf{Q}_{lk} \mathbf{S}_k) \leq q_l \quad \forall l. \end{aligned} \tag{2.27}$$

The relaxed problem (2.27) is convex if $f\left(g_1(\gamma_1), \ldots, g_{K_r}(\gamma_{K_r})\right)$ is concave and it always has rank-one solutions that also solve (2.26).

*Proof.* Lemma 1.6 and Theorem 1.8 can be applied to see that the relaxed problem always has rank-one solutions, as originally shown in [26, 297]. If an optimization procedure still delivers a high-rank solution $\mathbf{S}_k^*$, one can find $\mathbf{v}_k^*$ by maximizing $\Re(\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_k \mathbf{v}_k)$ under the interference constraints $|\mathbf{h}_i^H \mathbf{C}_i \mathbf{D}_k \mathbf{v}_k|^2 \leq \mathrm{tr}(\mathbf{D}_k^H \mathbf{C}_i^H \mathbf{h}_i \mathbf{h}_i^H \mathbf{C}_i \mathbf{D}_k \mathbf{S}_k^*)$ $\forall i \neq k$ and power constraints $\mathbf{v}_k^H \mathbf{Q}_{lk} \mathbf{v}_k \leq \mathrm{tr}(\mathbf{Q}_{lk} \mathbf{S}_k^*)$ $\forall l$. □

This theorem solves (2.24) in polynomial time if the system utility function is concave (which is often the case, see Example 2.5) and if all interference constraints are active at the optimal solution. The latter is always the case when $\Gamma_k = 0 \, \forall k$, but some interference constraints can in general be inactive and thereby enable improvements. In such

a case, $\Gamma_k$ can be reduced for the inactive constraints and then (2.26) is solved again. This iterative approach is not guaranteed to solve the original problem in (2.24), but successively finds better approximations. Another approach is to use the achieved solution as a starting-point for a fairness-profile optimization described later in this section (see Example 2.8). This will provide a weak Pareto optimal point, but not necessarily the one solving the original problem.

Instead of having one interference constraint $\Gamma_k$ per user that represents the aggregate inter-user interference that can be caused to $\mathrm{MS}_k$, it is possible to have $K_r - 1$ interference constraints $\Gamma_{ik}$, where each represents the interference that transmission to a particular co-user $\mathrm{MS}_i$ may cause to $\mathrm{MS}_k$ for $i \neq k$. This leads to interference constraints of the form $|\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_i \mathbf{v}_i|^2 \leq \Gamma_{ik}$ for all $k, i$ with $i \neq k$. This formulation generally provides lower performance, but might be useful as it decouples the beamforming selection and thus enables simple parametrizations (see Subsection 3.2.1) and distributed optimization (see Subsection 4.2.1).

---

**Remark 2.4 (Nonzero Solutions).** Zero-forcing constraints with $\Gamma_k = 0$ require $\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_i \mathbf{v}_i = 0$ for all $i \neq k$, which either requires that $\mathbf{D}_i \mathbf{v}_i$ is orthogonal to $\mathbf{C}_k^H \mathbf{h}_k$ or that $\mathbf{D}_i \mathbf{v}_i = \mathbf{0}$. Since the latter case would give $\mathrm{SINR}_i = 0$, it is desirable to operate in the former case where each beamforming vector is orthogonal to all co-user channels. However, this is only possible if there are sufficient degrees-of-freedom in the system; that is, if the set of co-user channels are not spanning the whole space. It is difficult to give a general condition on the existence of non-zero solutions, but $N_j \geq |\mathcal{C}_j| \; \forall j$ is necessary under coordinated beamforming (see Example 1.2) while $N \geq K_r$ is necessary under global joint transmission (see Example 1.3). Interference-constrained beamforming with $\Gamma_k > 0$ does not exhibit such restrictions.

---

**Remark 2.5 (Simplifying the General Problem).** This subsection assumed that the interference constraints (2.23) were part of the problem to be solved, meaning that our goal is to solve (2.24). It is also possible to add interference constraints to the general problem (2.1) for the purpose of simplifying the problem, while striving

for an optimal solution to the original non-interference-constrained problem. This heuristic approach is further discussed in Section 3.4 and makes sense from a theoretical standpoint, because interference-constrained beamforming provides the optimal solution to the general problem (2.1) if the interference constraints happen to equal the interference caused by the optimal solution to (2.1) [26, 215, 325]. This feature is utilized in [215] to solve general resource allocation problems.

### 2.2.2 Fixed Quality-of-Service Requirements

While the previous subsection considered fixed inter-user interference, we now consider the second approach for achieving convex problem formulations: fix the SINR value of each user. This special case is particularly important since it highlights a fundamental connection between beamforming optimization in the downlink and receive combining in a related uplink scenario. Furthermore, Subsection 2.2.3 will show that the fixed SINR values can be relaxed into searching for the optimal solution along a one-dimensional curve in the performance region.

Consider the case when the system designer knows exactly which performance each user should be allocated; the goal is to achieve $g_k(\mathrm{SINR}_k) = r_k^*$ for some given parameters $r_1^* \geq 0, \ldots, r_{K_r}^* \geq 0$. The resource allocation then consists of finding beamforming vectors that achieve this operating point, which is known as having fixed *quality-of-service (QoS) requirements* [11, 18, 59, 208, 209, 226, 296, 308]. This can be represented by the system utility function

$$f(g_1, \ldots, g_{K_r}) = \begin{cases} 0, & \min_{\{k: r_k^* > 0\}} \frac{g_k}{r_k^*} \geq 1, \\ -\infty, & \text{otherwise}, \end{cases} \tag{2.28}$$

which is zero if the QoS requirements are fulfilled. If the QoS requirements are unattainable (due to power constraints and/or inter-user interference), then the system utility is set to $-\infty$ which is the conventional way of saying that the feasible set is empty. Plugging (2.28)

into (2.22) yields the following resource allocation problem

$$\begin{aligned} &\text{find } \mathbf{v}_1 \ldots, \mathbf{v}_{K_r} && (2.29)\\ &\text{subject to } |\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_k \mathbf{v}_k|^2 \geq g_k^{-1}(r_k^*)\Big(\sigma_k^2 + \sum_{i \neq k} |\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_i \mathbf{v}_i|^2\Big) \quad \forall k, \\ &\qquad \sum_{k=1}^{K_r} \mathbf{v}_k^H \mathbf{Q}_{lk} \mathbf{v}_k \leq q_l \quad \forall l, \end{aligned}$$

where we utilized that the QoS requirements are infeasible exactly when $f(\cdot) \neq 0$. Observe that there is no cost function in (2.29), meaning that we are satisfied with finding any feasible solution to (2.29). This type of problem is known as a *feasibility problem* and can also be written as a minimization of a cost function that always equals zero. A preference of solutions that use little power can be induced by replacing the upper bound $q_l$ of each power constraint with $\beta q_l$ and then minimize over $\beta$:

$$\begin{aligned} &\underset{\mathbf{v}_k \, \forall k, \beta}{\text{minimize}} \ \beta && (2.30)\\ &\text{subject to } |\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_k \mathbf{v}_k|^2 \geq g_k^{-1}(r_k^*)\Big(\sigma_k^2 + \sum_{i \neq k} |\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_i \mathbf{v}_i|^2\Big) \quad \forall k, \\ &\qquad \sum_{k=1}^{K_r} \mathbf{v}_k^H \mathbf{Q}_{lk} \mathbf{v}_k \leq \beta q_l \quad \forall l. \end{aligned}$$

This reformulation of (2.29) into a power minimization under QoS requirements resembles how the problem was originally posed in [71, 208, 282]. The power minimization formulation might be more computationally tractable than (2.29) since the feasible set is larger; we accept $\beta > 1$ which means using more power than is actually available. In other words, the optimal solution $\{\mathbf{v}_k^*\}, \beta^*$ to (2.30) only satisfies the original power constraints in (1.4) if $\beta^* \leq 1$. The QoS requirements are infeasible if $\beta^* > 1$. Infeasibility can be handled by either reducing QoS constraints (e.g., by scaling down the power as $\mathbf{v}_k^*/\sqrt{\beta^*}$) or by removing the users that are hardest to serve [253].

The following theorem shows that both (2.29) and (2.30) can be cast as convex optimization problems.

**Theorem 2.6.** The optimization problems (2.29) and (2.30) become convex problems if the QoS constraints $|\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_k \mathbf{v}_k|^2 \geq g_k^{-1}(r_k^*)(\sigma_k^2 + \sum_{i \neq k} |\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_i \mathbf{v}_i|^2)$ are rewritten as

$$\begin{aligned} &\left\| \begin{array}{c} \mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_1 \mathbf{v}_1 \\ \vdots \\ \mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_{K_r} \mathbf{v}_{K_r} \\ \sigma_k \end{array} \right\| \leq \sqrt{\frac{1 + g_k^{-1}(r_k^*)}{g_k^{-1}(r_k^*)}} \Re(\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_k \mathbf{v}_k) \quad \forall k, \\ &\Im(\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_k \mathbf{v}_k) = 0 \quad \forall k, \end{aligned} \tag{2.31}$$

where the first row contains second-order cone constraints and the second row contains linear constraints.

*Proof.* Since the power constraints are convex (see Example 2.2) and the cost functions are convex, only the QoS constraints need reformulation. As in [11], we observe that the phase of $\mathbf{v}_k$ can be selected in an arbitrary way. This enables us to assume that $\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_k \mathbf{v}_k > 0$, which makes the square root of $|\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_k \mathbf{v}_k|^2$ well-defined. By reshuffling the constraints and taking the square root, we achieve (2.31). □

In other words, the resource allocation problem with QoS requirements can be solved with a computational complexity that only scales polynomially with the number of antennas $N$, users $K_r$, and power constraints $L$ [10, Chapter 6]. The exact complexity depends on current systems conditions and the choice of numerical algorithm (e.g., interior-point methods [256, 271]). In the special case of coordinated beamforming with single-antenna transmitters (see Example 1.2), the problem can even be reduced to a linear power allocation problem by setting $p_k = \|\mathbf{D}_k \mathbf{v}_k\|_2^2$:

$$\begin{aligned} &\underset{p_k \geq 0 \forall k, \beta}{\text{minimize}} \ \beta \\ &\text{subject to } \|\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_k\|_2^2 p_k \geq g_k^{-1}(r_k^*) \Big( \sigma_k^2 + \sum_{i \neq k} \|\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_i\|_2^2 p_i \Big) \quad \forall k, \\ &\qquad \sum_{k=1}^{K_r} \operatorname{tr}(\mathbf{D}_k^H \mathbf{Q}_{lk} \mathbf{D}_k) p_k \leq \beta q_l \quad \forall l. \end{aligned} \tag{2.32}$$

This fundamental type of power allocation problem was formulated already in the 1960s by Bock and Ebstein [32]. Applications in the area of cellular communications have also existed for many years; see for example [52, 137, 190, 304, 314, 316].

Next, we derive the Lagrange dual problem to (2.29) which has a conceptually important form.

---

**Theorem 2.7.** A Lagrange dual problem to (2.29) is[10]

$$\begin{aligned} &\underset{\lambda_k\,\forall k,\,\mu_l\,\forall l}{\text{maximize}} \ \sum_{k=1}^{K_r}\lambda_k - \sum_{l=1}^{L}\mu_l \qquad (2.33)\\ &\text{subject to } \ \mu_l \geq 0, \lambda_k \geq 0 \quad \forall k,l,\\ &\qquad \max_{\bar{\mathbf{v}}_k} \frac{\frac{\lambda_k}{\sigma_k^2}\bar{\mathbf{v}}_k^H \mathbf{D}_k^H \mathbf{C}_k^H \mathbf{h}_k \mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_k \bar{\mathbf{v}}_k}{\bar{\mathbf{v}}_k^H \big(\sum_l \frac{\mu_l}{q_l}\mathbf{Q}_{lk} + \sum_{i\neq k}\frac{\lambda_i}{\sigma_i^2}\mathbf{D}_k^H \mathbf{C}_i^H \mathbf{h}_i \mathbf{h}_i^H \mathbf{C}_i \mathbf{D}_k\big)\bar{\mathbf{v}}_k} = g_k^{-1}(r_k^*) \ \forall k. \end{aligned}$$

If the primal problem is feasible, then strong duality holds and thus the optimal values coincide as $\sum_{k=1}^{K_r}\lambda_k - \sum_{l=1}^{L}\mu_l = 0$.

---

*Proof.* The cost function (2.28) is not continuous, but if the primal problem is feasible then we operate in a range where strong duality follows from Slater's constraint qualification (see Lemma 2.4). The Lagrangian function associated with (2.29) is

$$\begin{aligned} &\mathcal{L}(\{\mathbf{v}_k\}, \boldsymbol{\lambda}, \boldsymbol{\mu})\\ &\quad = 0 + \sum_{l=1}^{L}\mu_l\left(\frac{1}{q_l}\sum_{k=1}^{K_r}\mathbf{v}_k^H\mathbf{Q}_{lk}\mathbf{v}_k - 1\right)\\ &\qquad + \sum_{k=1}^{K_r}\lambda_k\left(1 + \frac{1}{\sigma_k^2}\sum_{i\neq k}|\mathbf{h}_k^H\mathbf{C}_k\mathbf{D}_i\mathbf{v}_i|^2 - \frac{1}{\sigma_k^2\gamma_k}|\mathbf{h}_k^H\mathbf{C}_k\mathbf{D}_k\mathbf{v}_k|^2\right) \end{aligned}$$

[10] This problem formulation includes terms of the form $\mu_l/q_l$ which requires that $q_l > 0$. However, for every $q_l = 0$ we can simply replace the corresponding Lagrange multiplier $\mu_l$ with $\tilde{\mu}_l = \mu_l q_l$ in (2.33) to make the dual problem well-defined.

$$= \sum_{k=1}^{K_r} \lambda_k - \sum_{l=1}^{L} \mu_l + \sum_{k=1}^{K_r} \mathbf{v}_k^H \left( \sum_{l=1}^{L} \frac{\mu_l}{q_l} \mathbf{Q}_{lk} + \sum_{i \neq k} \frac{\lambda_i}{\sigma_i^2} \mathbf{D}_k^H \mathbf{C}_i^H \mathbf{h}_i \mathbf{h}_i^H \mathbf{C}_i \mathbf{D}_k - \frac{\lambda_k}{\sigma_k^2 \gamma_k} \mathbf{D}_k^H \mathbf{C}_k^H \mathbf{h}_k \mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_k \right) \mathbf{v}_k. \tag{2.34}$$

This expression is achieved by first dividing the power constraints by $q_l$ and the QoS constraints by $\sigma_k^2 \gamma_k$ (where $\gamma_k = g_k^{-1}(r_k^*)$), and then apply Definition 2.2. The second equality follows from rewriting the Lagrangian function in the same way as in [308, Proposition 1]. Minimizing (2.34) with respect to $\{\mathbf{v}_k\}$ gives a finite solution only if

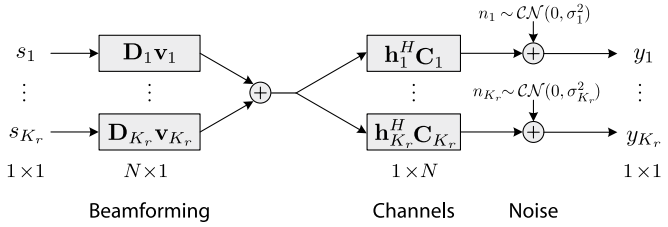
$$\left( \sum_{l=1}^{L} \frac{\mu_l}{q_l} \mathbf{Q}_{lk} + \sum_{i \neq k} \frac{\lambda_i}{\sigma_i^2} \mathbf{D}_k^H \mathbf{C}_i^H \mathbf{h}_i \mathbf{h}_i^H \mathbf{C}_i \mathbf{D}_k - \frac{\lambda_k}{\sigma_k^2 \gamma_k} \mathbf{D}_k^H \mathbf{C}_k^H \mathbf{h}_k \mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_k \right) \succeq \mathbf{0} \quad \forall k \tag{2.35}$$

and the corresponding minimum is $\sum_{k=1}^{K_r} \lambda_k - \sum_{l=1}^{L} \mu_l$ (achieved for $\mathbf{v}_k = \mathbf{0}_{N \times 1}$). Using [308, Lemma 1], the dual feasibility constraint (2.35) is equivalent to

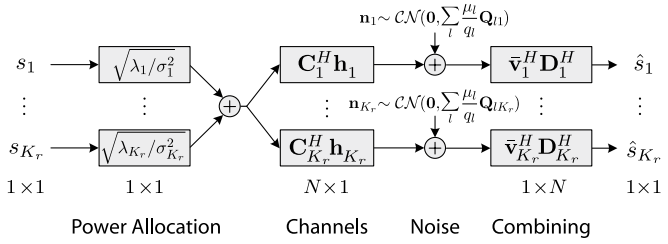
$$\begin{aligned} \gamma_k &\geq \frac{\lambda_k}{\sigma_k^2} \mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_k \left( \sum_{l=1}^{L} \frac{\mu_l}{q_l} \mathbf{Q}_{lk} + \sum_{i \neq k} \frac{\lambda_i}{\sigma_i^2} \mathbf{D}_k^H \mathbf{C}_i^H \mathbf{h}_i \mathbf{h}_i^H \mathbf{C}_i \mathbf{D}_k \right)^{\dagger} \mathbf{D}_k^H \mathbf{C}_k^H \mathbf{h}_k \\ &= \max_{\bar{\mathbf{v}}_k} \frac{\frac{\lambda_k}{\sigma_k^2} \bar{\mathbf{v}}_k^H \mathbf{D}_k^H \mathbf{C}_k^H \mathbf{h}_k \mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_k \bar{\mathbf{v}}_k}{\bar{\mathbf{v}}_k^H \left( \sum_l \frac{\mu_l}{q_l} \mathbf{Q}_{lk} + \sum_{i \neq k} \frac{\lambda_i}{\sigma_i^2} \mathbf{D}_k^H \mathbf{C}_i^H \mathbf{h}_i \mathbf{h}_i^H \mathbf{C}_i \mathbf{D}_k \right) \bar{\mathbf{v}}_k}, \end{aligned} \tag{2.36}$$

where the equality follows from introducing maximization over an auxiliary variable $\bar{\mathbf{v}}_k \in \mathbb{C}^{N \times 1}$. Its optimal value is given by (2.37) in Corollary 2.8, because (2.36) is a generalized Rayleigh quotient. The constraint (2.36) is active at the optimum of the dual problem for all $k$ (otherwise we can increase some $\lambda_k$ and thereby increase the dual function), thus we have the Lagrange dual problem in (2.33). □

This theorem establishes what is known as *uplink–downlink duality* [30, 226, 282, 283, 315]; the last line of (2.33) has the form of an uplink




(a) Downlink Communication using Linear Beamforming




(b) Virtual Uplink Communication using Linear Receive Combining

Fig. 2.3 Block diagram of multi-cell communications for: (a) the downlink; and (b) the virtual uplink achieved by uplink–downlink duality.

SINR for (reciprocal) transmission from $K_r$ single-antenna users to $K_t$ multi-antenna base stations. The uplink scenario that would give these SINRs is illustrated in Figure 2.3. With the uplink interpretation, the dual variable $\lambda_k$ is the uplink power of the signal from $\text{MS}_k$ (scaled by the downlink noise variance), $\bar{\mathbf{v}}_k$ is the receive combining vector used for reception of this signal, and $\mu_l$ is an uplink noise variance (scaled by the downlink power constraints).

Uplink–downlink duality implies that if a set of QoS requirements is feasible in the downlink, then this set is also feasible in the uplink and vice versa. Furthermore, there is an important relationship between the primal and dual variables.

---

**Corollary 2.8.** The optimal beamforming vector $\mathbf{v}_k^*$ to (2.29) is equal to the optimal dual variable

$$\bar{\mathbf{v}}_k^* = \left( \sum_{l=1}^{L} \frac{\mu_l}{q_l} \mathbf{Q}_{lk} + \sum_{i=1}^{K_r} \frac{\lambda_i}{\sigma_i^2} \mathbf{D}_k^H \mathbf{C}_i^H \mathbf{h}_i \mathbf{h}_i^H \mathbf{C}_i \mathbf{D}_k \right)^{\dagger} \mathbf{D}_k^H \mathbf{h}_k \tag{2.37}$$

up to a scaling factor.

---

*Proof.* The stationarity KKT condition (2.13) becomes

$$\mathbf{0} = \frac{\partial \mathcal{L}}{\partial \mathbf{v}_k} = 2\Bigg( \sum_{l=1}^{L} \frac{\mu_l}{q_l} \mathbf{Q}_{lk} + \sum_{i \neq k} \frac{\lambda_i}{\sigma_i^2} \mathbf{D}_k^H \mathbf{C}_i^H \mathbf{h}_i \mathbf{h}_i^H \mathbf{C}_i \mathbf{D}_k - \frac{\lambda_k}{\sigma_k^2 \gamma_k} \mathbf{D}_k^H \mathbf{C}_k^H \mathbf{h}_k \mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_k \Bigg) \mathbf{v}_k. \tag{2.38}$$

By defining the scalar $d_k = \frac{\lambda_k}{\sigma_k^2} \frac{(1+\gamma_k)}{\gamma_k} \mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_k \mathbf{v}_k$, using that $\mathbf{C}_k \mathbf{D}_k = \mathbf{D}_k$, and multiplying by a Moore–Penrose pseudo-inverse, (2.38) becomes

$$\mathbf{v}_k = d_k \left( \sum_{l=1}^{L} \frac{\mu_l}{q_l} \mathbf{Q}_{lk} + \sum_{i=1}^{K_r} \frac{\lambda_i}{\sigma_i^2} \mathbf{D}_k^H \mathbf{C}_i^H \mathbf{h}_i \mathbf{h}_i^H \mathbf{C}_i \mathbf{D}_k \right)^{\dagger} \mathbf{D}_k^H \mathbf{h}_k \tag{2.39}$$

and we identify $\bar{\mathbf{v}}_k$ from (2.37), which solves (2.36). □

This corollary shows that the optimal beamforming direction in the downlink is equivalent to the optimal receive combining in the uplink — this is quite intuitive if interpreted as turning the head toward the audience when speaking and pointing the ears in the same direction when listening. The proof of this relationship was however an important breakthrough as it is analytically simpler to select receive combining vectors than transmit beamforming; the former only affects the intended user while the latter affects all the users. Although the directions are equivalent, the corresponding power allocations are generally different between the downlink and the dual uplink (but there is a simple matrix transformation, see Subsection 3.2.3).

The duality is particularly strong in the case of a total power constraint (i.e., $L = 1$, $\mathbf{Q}_{lk} = \mathbf{I}_N \, \forall k$); the dual uplink then represents a problem formulation that is practically important for the uplink; see [30, 226, 283]. The duality can in this case be utilized to design iterative fixed-point algorithms that quickly find the optimal dual variables and thereby solve both the downlink and uplink problems [42, 59, 208, 226, 227, 296]. We refer to [227, 228] for further details on such algorithms and the related topic of general interference functions. Fixed-point algorithms are less useful in the general multi-cell

case (although an outer optimization procedure can be applied to take care of general power constraints [59, 308]). In fact, the dual problem in Theorem 2.7 is more of a *virtual* multi-cell uplink scenario than a practically reasonable problem formulation; the uplink noise in (2.33) is determined by the dual variables $\mu_l$ and the cost function, $\sum_{k=1}^{K_r} \lambda_k - \sum_{l=1}^{L} \mu_l$, represents some kind of balance between the uplink transmit power and the uplink noise power. Nevertheless, the multi-cell uplink-downlink duality will be exploited in Section 3 to achieve strong parametrizations of the optimal beamforming. It will also be an enabler for truly distributed resource allocation in Section 4.

### 2.2.3 Quasi-Fixed Quality-of-Service Requirements

The previous subsection showed that resource allocation with fixed QoS requirements leads to convex optimization problems. This important result is utilized in this subsection to achieve efficient solutions to a wider class of resource allocation problems where the QoS requirements are flexible but governed by a single parameter.[11] To describe this structure in general terms, we consider a continuous vector-valued function $\mathbf{r}(\tau) = [r_1(\tau) \ldots r_{K_r}(\tau)]^T$ of the scalar parameter $\tau \in \mathbb{R}_+$. This function is assumed to be strictly monotonically increasing, thus whenever $\tau_1 > \tau_2 \geq 0$ we have $r_k(\tau_1) \geq r_k(\tau_2)$ $\forall k$ and there is at least one strict inequality. Observe that $\mathbf{r}(\tau)$ for $\tau \in [0, \tau^{\text{upper}}]$ describes a one-dimensional curve that connects the points $\mathbf{r}(0)$ and $\mathbf{r}(\tau^{\text{upper}})$ and constantly moves away from the origin; see Figure 2.4. If the curve is plotted against the performance region $\mathcal{R}$, we have the following result.

---

**Lemma 2.9.** Consider the curve generated by a continuous strictly monotonically increasing function $\mathbf{r} : \mathbb{R}_+ \to \mathbb{R}_+^{K_r}$. If $\mathbf{r}(0) \in \mathcal{R}$ and $\mathbf{r}(\tau^{\text{upper}}) \notin \mathcal{R}$ for some $\tau^{\text{upper}} > 0$, then the curve intersects the Pareto

[11] This subsection considers optimization of the QoS under fixed power constraints, while (2.30) in the previous subsection minimizes the transmit power under fixed QoS requirements. Note that these problems are each other's inverses; if the optimal QoS achieved in this subsection is used as QoS requirements in (2.30), then the two problems will have the same optimal beamforming. However, the problem formulation in this subsection is often preferable as it always gives a Pareto optimal point, while (2.30) requires that a good operating point is known beforehand — which is generally not easy to achieve.

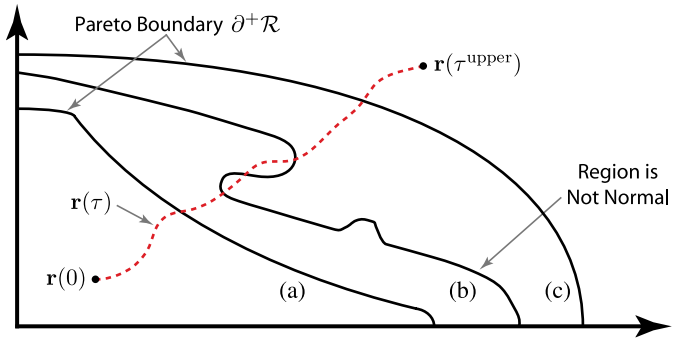



Fig. 2.4 Illustration of a one-dimensional curve generated by the strictly increasing vector-valued function $\mathbf{r}(\tau)$ for $\tau \in [0, \tau^{\text{upper}}]$. If $\mathbf{r}(0)$ is inside a normal region and $\mathbf{r}(\tau^{\text{upper}})$ is outside, then the curve intersects the Pareto boundary only once. For the non-normal region (b) the curve leaves the region and then comes back again.

boundary of $\mathcal{R}$ exactly once. This happens at $\tau \in [\tau_1^*, \tau_2^*]$ where $\tau_1^* \leq \tau_2^*$. There is always a unique intersection point $\tau_1^* = \tau_2^*$ when the weak and strong Pareto boundary coincides.

---

*Proof.* There will be at least one intersection with the weak Pareto boundary $\partial^+\mathcal{R}$, due to the continuity of $\mathbf{r}(\tau)$ and that $\mathcal{R}$ is compact and normal. Suppose it exists $\tau_1^* < \tau_2^*$ such that $\mathbf{r}(\tau_1^*), \mathbf{r}(\tau_2^*) \in \partial^+\mathcal{R}$ while $\mathbf{r}(\tau_3) \notin \partial^+\mathcal{R}$ for some $\tau_3 \in [\tau_1^*, \tau_2^*]$. The definition of weak Pareto optimal points then implies that $\mathbf{r}(\tau_1^*)$ cannot be Pareto optimal either, which is a contradiction. Consequently, the intersection occurs for all points in the interval $[\tau_1^*, \tau_2^*]$. If the weak and strong Pareto boundary coincides, then intersection point must be unique due to the definition of strong Pareto optimal points. □

This lemma proves that a strictly increasing curve that leaves the performance region intersects the Pareto boundary exactly once. This might seem trivial, but it requires that the region is normal (as proved in Lemma 1.10). This property is illustrated in Figure 2.4, where (a) and (c) are normal regions while (b) is nonnormal and thus some increasing curves can cross the boundary multiple times. There is only one intersection point in most cases, but if the curve enters the boundary at a

weak Pareto optimal point then it might follow the boundary until a strong Pareto optimal point is found and then leave it.

Suppose we optimize over $\tau$ to find the outermost intersection point, this can be formulated as an optimization problem.

---

**Theorem 2.10.** Consider the optimization problem

$$\begin{aligned} &\underset{\mathbf{v}_1,\ldots,\mathbf{v}_{K_r},\tau}{\text{maximize}} \ \tau \\ &\text{subject to } r_k(\tau) = g_k(\text{SINR}_k) \quad \forall k, \\ &\qquad \sum_{k=1}^{K_r} \mathbf{v}_k^H \mathbf{Q}_{lk} \mathbf{v}_k \leq q_l \quad \forall l, \\ &\qquad \tau \in [0, \tau^{\text{upper}}] \end{aligned} \tag{2.40}$$

for a strictly increasing function $\mathbf{r}(\tau)$. This problem can be solved by line-search over the range $\mathcal{T} = [0, \tau^{\text{upper}}]$. For a given $\tau^{\text{candidate}} \in \mathcal{T}$, the convex feasibility problem (2.29) is solved for $r_k^* = r_k(\tau^{\text{candidate}}) \forall k$. If the problem is feasible, all $\tilde{\tau} \in \mathcal{T}$ with $\tilde{\tau} < \tau^{\text{candidate}}$ are removed from $\mathcal{T}$. Otherwise, all $\tilde{\tau} \in \mathcal{T}$ with $\tilde{\tau} \geq \tau^{\text{candidate}}$ are removed.

Initial feasibility of (2.40) is checked by (2.29) for $r_k^* = r_k(0)$. The optimum is achieved at $\tau^{\text{upper}}$ if (2.29) is feasible for $r_k^* = r_k(\tau^{\text{upper}})$.

---

*Proof.* The convex feasibility problem (2.29) checks whether a point $\mathbf{r}^*$ is inside $\mathcal{R}$ or not. As $\mathbf{r}(\tau)$ is strictly increasing, (2.40) is infeasible if $\mathbf{r}(0) \notin \mathcal{R}$ and is solved at $\tau^{\text{upper}}$ if $\mathbf{r}(\tau^{\text{upper}}) \in \mathcal{R}$. In any other case, Lemma 2.9 shows that $\mathbf{r}(\tau)$ intersects $\partial^+ \mathcal{R}$ once and there is a unique last intersection point $\mathbf{r}(\tau^{\text{optimal}})$ for some $\tau^{\text{optimal}} \in [0, \tau^{\text{upper}}]$. Therefore, the range $\mathcal{T}$ can be divided into two parts: one part is inside of $\mathcal{R}$ and one part is outside. The intersection can be found (to any accuracy $\delta$) by a line-search that iteratively checks if a point $\mathbf{r}(\tau^{\text{candidate}})$ is inside $\mathcal{R}$ by solving (2.29). □

Theorem 2.10 shows that optimization along a strictly increasing curve $\mathbf{r}(\tau)$ can be solved by line-search over the range of $\tau$, where the subproblems are convex feasibility problems. This means that (2.40) is a quasi-convex problem [37]. The *bisection method* is an efficient line-search procedure where each iteration consists of checking the feasibility

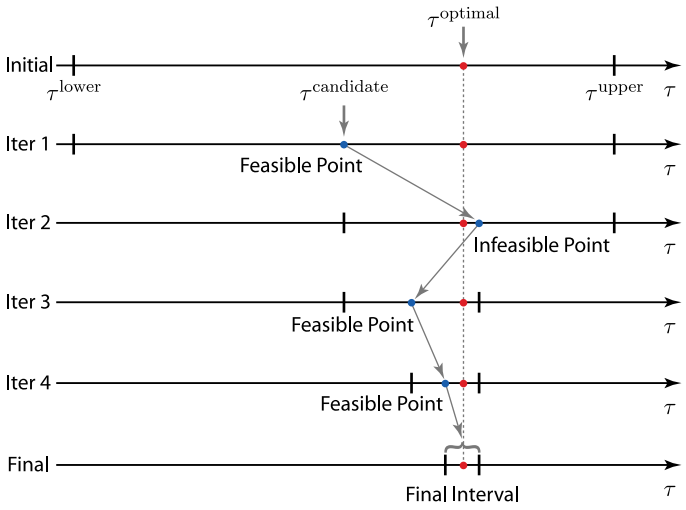



Fig. 2.5 Illustration of the bisection method that searches the range $\mathcal{T} = [\tau^{\text{lower}}, \tau^{\text{upper}}]$ to find $\tau^{\text{optimal}}$. The feasibility at the midpoint $\tau^{\text{candidate}}$ is checked in each iteration (i.e., is $\tau^{\text{candidate}} \leq \tau^{\text{optimal}}$?) and half the interval is removed based on the answer.

at the midpoint of the current range [37], thus the range is halved at each iteration. The bisection method is illustrated in Figure 2.5 and the approach is described in Algorithm 1. The number of iterations in the bisection method scales only logarithmical with the desired width $\delta$ of the final interval — precisely $\lceil \log_2(\tau^{\text{upper}}/\delta) \rceil$ feasibility problems will be solved. As this variable is bounded by a constant, the computational complexity is just a constant times the complexity of the convex feasibility problem (2.29) solved in each iteration. In other words, the worst-case computationally complexity is polynomial in the number of antennas $N$, users $K_r$, and power constraints $L$ [10, Chapter 6].

Theorem 2.10 shows how to solve a class of quasi-convex problems. These are connected to a certain type of resource allocation problems.

---

**Corollary 2.11.** Consider a resource allocation problem of the form (2.1) with $f(\mathbf{g}) = \min_k \upsilon_k(g_k)$, for some continuous and strictly increasing functions $\upsilon_k : \mathbb{R}_+ \to \mathbb{R}_+$ that satisfy $\upsilon_k(0) = 0$. This problem is solved by Theorem 2.10 for $r_k(\tau) = \upsilon_k^{-1}(\tau)$ and some $\tau^{\text{upper}}$ that satisfies $\mathbf{r}(\tau^{\text{upper}}) \notin \mathcal{R}$.

---

**Algorithm 1**: Optimization Along a Strictly Increasing Curve

**Result**: Solves optimization problem in (2.40).
**Input**: Lower bound $\tau^{\text{lower}}$ and upper bound $\tau^{\text{upper}}$ on $\tau$;
**Input**: Line-search accuracy $\delta$;
**while** $\tau^{\text{upper}} - \tau^{\text{lower}} > \delta$ **do**
  Set $\tau^{\text{candidate}} = \frac{\tau^{\text{lower}} + \tau^{\text{upper}}}{2}$;
  Set $r_k^* = r_k(\tau^{\text{candidate}}) \quad \forall k$;
  **if** *Problem* (2.29) *is feasible for these* $\{r_k^*\}$ **then**
    Set $\{\mathbf{v}_k^{\text{lower}}\}$ as the solution to (2.29);
    Set $\tau^{\text{lower}} = \tau^{\text{candidate}}$;
  **else**
    Set $\tau^{\text{upper}} = \tau^{\text{candidate}}$;
Set $\tau_{\text{final}}^{\text{lower}} = \tau^{\text{lower}}$ and $\tau_{\text{final}}^{\text{upper}} = \tau^{\text{upper}}$;
**Output**: Final interval $[\tau_{\text{final}}^{\text{lower}}, \tau_{\text{final}}^{\text{upper}}]$ for $\tau$;
**Output**: Best feasible solution $\{\mathbf{v}_k^{\text{lower}}\}$;

*Proof.* Suppose the optimal value is $f(\mathbf{g}^*) = \tau^{\text{optimal}}$, then there exists an optimal solution with $\upsilon_k(g_k) = \tau^{\text{optimal}}$ for all $k$. This is equivalent to $g_k = \upsilon_k^{-1}(\tau^{\text{optimal}})$, which is the last intersection point between $\mathbf{r}(\tau)$ and the weak Pareto boundary of $\mathcal{R}$. □

Resource allocation problems covered by Corollary 2.11 concentrate on the worst-user performance, but can still take many different forms. The following examples are illustrated in Figure 2.6.

**Example 2.6 ($\epsilon$-Constraint Optimization).** The $\epsilon$-constraint optimization represents maximizing the performance of $\text{MS}_k$, while guaranteeing that $g_i \geq \epsilon_i$ for all $i$ [38, 98, 123, 149, 278, 292, 293]. This problem is solved by Theorem 2.10 using $r_k(\tau) = \tau + \epsilon_k$ and $r_i(\tau) = \epsilon_i$.

**Example 2.7 (Max-Min Fairness).** Max-min fairness optimization is given by $f(\mathbf{g}) = \min_k g_k$ [42, 226, 227, 270, 296]. This problem is

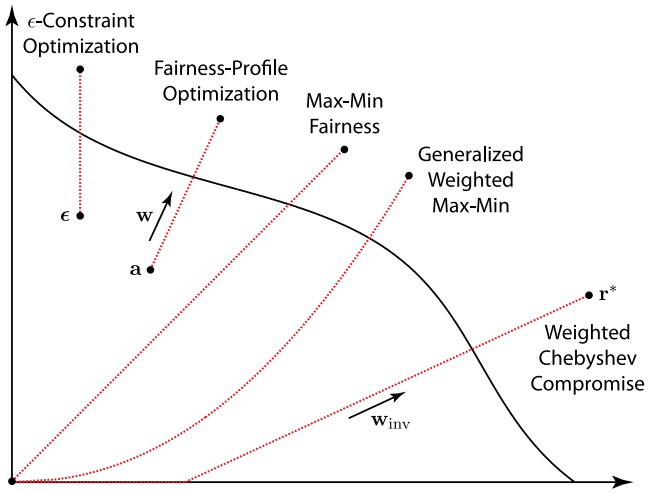



Fig. 2.6 Illustration of the strictly increasing curves $\mathbf{r}(\tau)$ that represents $\epsilon$-constraint optimization (Example 2.6), max-min fairness (Example 2.7), fairness-profile optimization (Example 2.8), weighted Chebyshev compromise (Example 2.9), and generalized weighted max-min optimization (Example 2.10). These problems are solved in polynomial time using Theorem 2.10 and Corollary 2.11.

solved by Theorem 2.10 using $\mathbf{r}(\tau) = [\tau \dots \tau]^T$, which corresponds to searching on a line in the direction $[1 \dots 1]^T$ from the origin.

---

**Example 2.8 (Fairness-Profile Optimization (FPO)).** Fairness-profile optimization (FPO) is given by

$$f(\mathbf{g}) = \begin{cases} \min_{\{k:\, w_k > 0\}} \frac{g_k - a_k}{w_k}, & \min_k g_k - a_k \geq 0, \\ -\infty, & \text{otherwise.} \end{cases} \tag{2.41}$$

This is a generalization of max-min optimization in Example 2.7 where two fairness constraints[12] have been added [17, 26, 126, 144, 185, 193, 325]:

[12] The fairness constraints have important bargaining interpretations in cooperative game-theoretic setups where users compete for resources [193]: The so-called Kalai–Smorodinsky objective function can be formulated as (2.41) using $\mathbf{w} = \mathbf{u} - \mathbf{a}$ as the weighing factors. The vector $\mathbf{a}$ is the disagreement point used if bargaining fails, while $\mathbf{w}$ is the direction from $\mathbf{a}$ toward the utopia point $\mathbf{u}$. Bargaining thus improves user performance proportionally to the performance each user would achieve with TDMA.

(1) Each user has a lowest acceptable performance level $g_k(\text{SINR}_k) \geq a_k$ for some $a_k \geq 0$;
(2) The aggregate performance above this level (i.e., $\sum_k (g_k - a_k)$) is divided such that each user gets a predefined fraction $w_k \geq 0$.[13]

This problem is solved by Theorem 2.10 using $r_k(\tau) = w_k\tau + a_k$, which corresponds to searching on a line segment from $\mathbf{a} = [a_1 \dots a_{K_r}]^T$ to some infeasible point $\mathbf{r}(\tau^{\text{upper}})$ in the direction of $\mathbf{w} = [w_1 \dots w_{K_r}]^T$.

---

**Example 2.9 (Weighted Chebyshev Compromise).** For a given reference point $\mathbf{r}^* \in \mathbb{R}_+^n \setminus \mathcal{R}$, the weighted compromise problem $f(\mathbf{g}) = -(\sum_k (w_k(r_k^* - g_k))^p)^{1/p}$ finds the closest feasible point in the weighted $L_p$-norm. This problem can be solved by Theorem 2.10 if we consider the $L_\infty$-norm (also known as Chebyshev metric), which corresponds to $f(\mathbf{g}) = -\max_k w_k(r_k^* - g_k)$ [278].

To find the appropriate curve $\mathbf{r}(\tau)$, note that one solution is given by $w_k(r_k^* - g_k) = a \; \forall k$ for some appropriate value of $a$. This can be rewritten as $g_k = r_k^* - \frac{a}{w_k}$, which reveals that we should search on a line in the direction of $\mathbf{w}_{\text{inv}} = [\frac{1}{w_1} \dots \frac{1}{w_{K_r}}]^T$ that intersects with $\mathbf{r}^*$. This line will generally not pass through the origin. The performance $g_k$ of $\text{MS}_k$ is only positive when $a \leq w_k r_k^*$, thus the operator $[\cdot]_+$ will be used to ensure that negative performance entries are replaced by zero. The strictly increasing curve can be expressed as $\mathbf{r}(\tau) = [\mathbf{r}^* - (1 - \tau)c\mathbf{w}_{\text{inv}}]_+$ for $\tau \in [0, 1]$, where $c = \max_k w_k r_k^*$ is the value of $a$ where all users achieve zero performance.

---

**Example 2.10 (Generalized Weighted Max-Min Optimization).** Max-min optimization can be generalized as $f(\mathbf{g}) = \min_k \tilde{w}_k(g_k)$, where $\tilde{w}_k(\cdot)$ is a strictly increasing weighting

[13] To see that the weighting factors equal the fraction of aggregate performance allocated to each user, note that one of the optimal solutions to (2.41) is when $(g_k - a_k)/w_k$ is the same for all active users.

function. This function can describe, for example, a multiplicative weighting $\tilde{w}_k(g_k) = w_k g_k$ or a weighting exponent $\tilde{w}_k(g_k) = g_k^{w_k}$, for some fixed $w_k > 0$. This is equivalent to Corollary 2.11 with $\upsilon_k(\cdot) = \tilde{w}_k(\cdot)$ and is solved by searching along $r_k(\tau) = \tilde{w}_k^{-1}(\tau)$, which in general is not a line.

---

Corollary 2.11 requires an initial upper bound $\tau^{\text{upper}}$ satisfying $\mathbf{r}(\tau^{\text{upper}}) \notin \mathcal{R}$. If not given in advance, $\tau^{\text{upper}}$ can be selected as follows:

- $\tau^{\text{upper}} = \min_k \upsilon_k(u_k)$ for utopia point $\mathbf{u} = [u_1 \dots u_{K_r}]^T$.
- $\tau^{\text{upper}} = \min_k \upsilon_k\left(g_k\left(\frac{\kappa_k \|\mathbf{D}_k^H \mathbf{h}_k\|_2^2}{\sigma_k^2}\right)\right)$, where $\kappa_k$ is a bound on the maximum transmit power and can be calculated as the smallest positive eigenvalue of $\frac{\mathbf{D}_k^H \mathbf{Q}_{lk} \mathbf{D}_k}{q_l \text{tr}(\mathbf{D}_k)}$ among all $l$.
- $\tau^{\text{upper}} = \min_k \lim_{\rho \to \infty} \upsilon_k(g_k(\rho))$, which is only useful if $g_k(\rho) \to c < \infty$ as $\rho \to \infty$.

The first alternative is based on the utopia point and thus provides the tightest search range, but at the expense of solving $K_r$ single-user problems (see Lemma 1.3). The second alternative ignores inter-user interference and assumes that the highest power available in some spatial direction can be used in any direction. The third alternative is the simplest because it ignores both inter-user interference and power constraints.

---

**Remark 2.6 (Non-Uniqueness).** Although the quasi-convexity of max-min optimization has been known in the communication community for at least a decade [270, 303], it is not always embraced in the literature; for example, the property is not exploited when solving such problems in [133, 250], leading to unnecessary high computational complexity. The reason might be that the nondifferentiable min-operator makes the problem look nonsmooth. In fact, if Algorithm 1 finds a weak Pareto optimal point, then it also exists strong Pareto optimal points that give the same optimal system utility but where a (strict) subset of users achieve higher performance [172]. It is easy to

refine the solution to one of these strong Pareto optimal points (e.g., by $\epsilon$-constraint optimization done sequentially for all users), but at the expense of increasing the complexity with approximately a factor $K_r$. However, it can be very difficult to find the lexicographic[14] optimal solution [222]. Algorithm 1 can be applied to other system models, but the subproblems might not be convex in these cases (for example, see multi-cast transmission in Section 4.4).

### 2.2.4 Change of Variables

The third approach to achieve convex problem formulations (as outlined in Section 2.2) is to make a change of variables. The idea is to turn the multiplication between $\gamma_k$ and $\sum_{i\neq k}|\mathbf{h}_k^H\mathbf{C}_k\mathbf{D}_i\mathbf{v}_i|^2$ in the SINR constraints into an addition by using logarithms. This must be done in a clever way to make each term convex, although the logarithm is a concave function. This is typically only possible when a single antenna is transmitting to each user (see also Remark 2.8).

In this subsection, we consider the special case of coordinated beamforming with single-antenna transmitters (see Example 1.2), thus $\mathbf{h}_k^H\mathbf{C}_k\mathbf{D}_i$ and $\mathbf{D}_k\mathbf{v}_k$ have at most one nonzero element. We can therefore define $p_k = \|\mathbf{D}_k\mathbf{v}_k\|_2^2$ and turn (2.22) into

$$\begin{aligned}
&\underset{p_k\geq 0,\gamma_k\,\forall k}{\text{minimize}} \quad -f\left(g_1(\gamma_1),\ldots,g_{K_r}(\gamma_{K_r})\right) && (2.42)\\
&\text{subject to } \|\mathbf{h}_k^H\mathbf{C}_k\mathbf{D}_k\|_2^2 p_k \geq \gamma_k\Big(\sigma_k^2 + \sum_{i\neq k}\|\mathbf{h}_k^H\mathbf{C}_k\mathbf{D}_i\|_2^2 p_i\Big) \quad \forall k,\\
&\qquad\qquad \sum_{k=1}^{K_r}\mathrm{tr}(\mathbf{D}_k^H\mathbf{Q}_{lk}\mathbf{D}_k)p_k \leq q_l \quad \forall l.
\end{aligned}$$

Similar to [29, 168], we now make a change of variables: $p_k \to \tilde{p}_k, \gamma_k \to \tilde{\gamma}_k$ with $p_k = e^{\tilde{p}_k}$ and $\gamma_k = g_k^{-1}(e^{\tilde{\gamma}_k})$. This corresponds to measuring

[14] The lexicographic solution jointly maximizes the worst-user performance, second-worst-user performance, and so on.

transmit power in log-scale and (2.42) becomes

$$\begin{aligned}
\underset{\tilde{p}_k,\tilde{\gamma}_k\,\forall k}{\text{minimize}} \;& -f\left(e^{\tilde{\gamma}_1},\ldots,e^{\tilde{\gamma}_{K_r}}\right) \\
\text{subject to } & \log\left(\frac{\sigma_k^2}{\|\mathbf{h}_k^H\mathbf{C}_k\mathbf{D}_k\|_2^2}e^{-\tilde{p}_k}+\sum_{i\neq k}\frac{\|\mathbf{h}_k^H\mathbf{C}_k\mathbf{D}_i\|_2^2}{\|\mathbf{h}_k^H\mathbf{C}_k\mathbf{D}_k\|_2^2}e^{\tilde{p}_i-\tilde{p}_k}\right) \\
& \qquad + \log\left(g_k^{-1}(e^{\tilde{\gamma}_k})\right)\leq 0 \quad \forall k, \\
& \sum_{k=1}^{K_r}\mathrm{tr}(\mathbf{D}_k^H\mathbf{Q}_{lk}\mathbf{D}_k)e^{\tilde{p}_k}\leq q_l \quad \forall l.
\end{aligned} \tag{2.43}$$

Observe that we also have taken the logarithm of both sides in the SINR constraints and gathered all the terms.

---

**Theorem 2.12.** The transformed optimization problem in (2.43) is convex if both $-f(e^{\tilde{\gamma}_1},\ldots,e^{\tilde{\gamma}_{K_r}})$ and $\log(g_k^{-1}(e^{\tilde{\gamma}_k}))\,\forall k$ are convex with respect to $\tilde{\gamma}_1,\ldots,\tilde{\gamma}_{K_r}$.

---

*Proof.* The problem is convex if every term is convex. Under the stipulated conditions, it remains to check that $\log(c_k e^{-\tilde{p}_k}+\sum_{i\neq k}d_k e^{\tilde{p}_i-\tilde{p}_k})$ is convex for any $c_k,d_k\geq 0$ and that the power constraints are convex. The former can be checked by straightforward differentiation [168], while the latter follows since the exponential function is convex. □

The conditions in Theorem 2.12 are not satisfied by all system utility and user performance functions, but for several of practical interest.

---

**Corollary 2.13.** The cost function $-f(e^{\tilde{\gamma}_1},\ldots,e^{\tilde{\gamma}_{K_r}})$ is convex for the weighted geometric mean and the weighted harmonic mean. Convexity is also satisfied for the weighted compromise if only $\mathbf{g}\geq\mathbf{r}/p$ are of interest (where $\mathbf{r}$ is the reference point). In addition, $\log(g_k^{-1}(e^{\tilde{\gamma}_k}))$ is convex for the information rate and for the MSE.

---

*Proof.* A continuous and differentiable function is convex if the Hessian is positive semi-definite. The weighted geometric mean can be written as $\sum_k w_k\log(g_k)$, thus $-f(e^{\tilde{\gamma}_1},\ldots,e^{\tilde{\gamma}_{K_r}})=-\sum_k\tilde{\gamma}_k$ which is both

a convex and concave function. The weighted harmonic mean can be written as $-\sum_k \frac{w_k}{g_k}$, thus $-f(e^{\tilde{\gamma}_1},\ldots,e^{\tilde{\gamma}_{K_r}}) = \sum_k w_k e^{-\tilde{\gamma}_k}$, which is a convex function.

The second-order derivative of $\sum_k w_k^p (r_k - e^{\tilde{\gamma}_k})^p$ with respect to $\tilde{\gamma}_k$ is $pw_k^p e^{\tilde{\gamma}_k}(r_k - e^{\tilde{\gamma}_k})^{p-2}(pe^{\tilde{\gamma}_k} - r_k)$, which is nonnegative if we restrict the search-space to $e^{\tilde{\gamma}_k} \geq r_k/p$. The convexity of $\log(g_k^{-1}(e^{\tilde{\gamma}_k}))$ for the information rate and the MSE follows from checking the second-order derivatives. □

Although (2.43) is convex in many scenarios with single-antenna transmitters, the optimization problem might be difficult to implement in a way that the high-level modeling languages `CVX` [95] and `Yalmip` [161] will accept. On the other hand, single-antenna coordinated beamforming is a special case of limited practical interest — the convexity results in this subsection mainly show that some optimization problems are significantly easier to solve in the single-antenna case, since the transmitted signals have fixed spatial directivity and the beamforming design reduces to power allocation. This can be utilized in the following way.

---

**Remark 2.7 (Power Allocation for Heuristic Beamforming).** Suppose the beamforming vectors are decomposed as $\mathbf{v}_k = \sqrt{p_k}\bar{\mathbf{v}}_k$ for all $k$, where $\bar{\mathbf{v}}_k$ are the normalized beamforming directions and $p_k \geq 0$ are the corresponding power allocation coefficients. If the beamforming directions are fixed, the remaining resource allocation problem can be expressed as (2.42) (by replacing $\|\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_i\|_2^2$ with $|\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_i \bar{\mathbf{v}}_i|^2$ everywhere). This subproblem is convex in many cases, which indicates that finding the optimal beamforming directions is the difficult part in multi-antenna resource allocation. In other words, the computational complexity can be greatly reduced by selecting the beamforming directions heuristically. Different beamforming parametrizations and common heuristic approaches are described in Section 3.

---

**Remark 2.8 (Interference Functions).** Theorem 2.12 shows that the variable substitution $p_k = e^{\tilde{p}_k}$ can extract hidden convexity in

scenarios with single-antenna transmitters. This is no coincident, but provably the only substitution that can be applied for *all* problems where $f(g_1(\mathrm{SINR}_1),\ldots,g_{K_r}(\mathrm{SINR}_{K_r}))$ can be written as $\sum_k \omega_k \tilde{g}_k(\mathrm{SINR}_k)$ for some weighting factors $\omega_k \geq 0$ and some functions $\tilde{g}_k$ for which $\tilde{g}_k(e^{\tilde{p}_k})$ is concave [29]. The weighted geometric and harmonic means of information rates can be expressed in this way, while the arithmetic mean cannot. This result can be generalized beyond the SINR expression $\frac{\|\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_k\|_2^2 p_k}{\sigma_k^2 + \sum_{i \neq k} \|\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_i\|_2^2 p_i}$ used in this subsection; [29] extends the convexification to SINRs based on so-called log-convex interference functions [31]. These alternative SINR expressions can, for example, describe worst-case interference or uplink transmission to multi-antenna receivers. We refer to [227, 228] for further details on general interference functions.

### 2.2.5 Summary of Convexity Classification

We conclude the section on convexity by summarizing which resource allocation problems are linear, convex, or quasi-convex (and which are not). Under the assumption that the user performance functions are concave (which is often satisfied, see Example 2.5), Table 2.1 shows the classification for maximizing the weighted arithmetic mean, weighted geometric mean, weighted harmonic mean, weighted max-min fairness, weighted compromise, or having fixed QoS requirements. Three different system scenarios are considered: the general case, zero-forcing constraints, and single-antenna coordinated beamforming.

Table 2.1. Summary of classification for resource allocation problems.

| | **System scenarios** | | |
|---|---|---|---|
| | **General** | **Zero-forcing** | **Single-antenna** |
| **Arithmetic mean** | NP-hard | Convex | NP-hard |
| **Geometric mean** | NP-hard | Convex | Convex |
| **Harmonic mean** | NP-hard | Convex | Convex |
| **Max-Min fairness** | Quasi-convex | Quasi-convex | Quasi-convex |
| **Compromise** | NP-hard | Convex | Convex/NP-hard |
| **Fixed QoS** | Convex | Convex | Linear |

There are several optimization scenarios in Table 2.1 that have *not* been proved to be linear, convex, or quasi-convex in this tutorial. These scenarios are however analyzed in [157, 168] and the authors show that these problems are NP-hard. A main characteristic of NP-hard problems is that there are no known algorithms that solve them in polynomial time, and it is widely believed that there exist no such algorithms. The weighted arithmetic mean is NP-hard for any number of transmit antennas, while the weighted geometric and harmonic means are NP-hard for single-cell and interference channels with $N_j > 1$. We will not dig deeper into the notion and proofs of NP-hardness herein, but simply label these problems as NP-hard in Table 2.1. A recent survey on the NP-hardness of these resource allocation problems and related problems is available in [104].

From Table 2.1 it is clear that only resource allocation problems that maximize the weighted max-min fairness or have fixed QoS requirements are *always* solvable in polynomial time. Furthermore, zero-forcing constraints lead to convex problems for all of the considered system utility functions. Looking at single-antenna coordinated beamforming, it is clear that optimization of the arithmetic mean is the most difficult problem as it is the only one that cannot be solved in polynomial time.

---

**Remark 2.9 (Freedom is Problematic).** Roughly speaking, resource allocation problems are only convex when the cost function and/or power constraints greatly limit the degrees-of-freedom for selecting beamforming vectors. The zero-forcing and single-antenna cases remove much of the freedom of choice in the spatial dimension. Similarly, fixed QoS requirements and max-min fairness strictly specify the amount and/or fraction of resources that each user should be allocated. The arithmetic mean with per-transmitter power constraints represents the other extreme: the utility function leaves all fairness decisions to the optimization process and the transmit power can be allocated freely over each antenna array. Consequently, this is the most difficult problem to solve.

---

Nonconvex resource allocation problems are not without structure; all resource allocation problems are monotonic and this property can be utilized to solve the problems in a structured way, as shown in the next section.

## 2.3 Monotonic Optimization for Resource Allocation

In this section, we will solve the multi-cell resource allocation problem in (2.1) for any system utility function $f(\cdot)$ and user performance functions $g_k(\cdot)$. As these are increasing functions (and the power constraints are convex), (2.1) is always a monotonic optimization problem. It will be useful to express (2.1) as a search in the performance region,

$$\begin{aligned} &\underset{\mathbf{g}}{\text{maximize}} \ f(\mathbf{g}) \\ &\text{subject to } \mathbf{g} \in \mathcal{R}, \end{aligned} \tag{2.44}$$

instead of using standard form. We emphasize that even if we select $f(\cdot)$ as a concave function and $\mathcal{R}$ happens to be a convex set, (2.44) is generally not considered a convex problem. The reason is that $\mathcal{R}$ is not defined by a finite set of convex inequality constraints, as required for convex problems on standard form. Instead, checking if $\mathbf{r} \in \mathbb{R}_+^{K_r}$ belongs to $\mathcal{R}$ is a convex feasibility problem with QoS requirements, which can be solved as in Subsection 2.2.2.

As compared to arbitrary nonconvex problems, monotonic problems have the important property that the optimum lies on the Pareto boundary of $\mathcal{R}$ (see Lemma 1.10). This property should certainly be utilized when devising a numerical algorithm for solving the problem. The naive approach would be to generate a large set of Pareto optimal points, preferably by some approach that finds Pareto optimal points with polynomial computational complexity (e.g., the fairness-profile optimization problem in Example 2.8 can be used, if the weights $\{w_k\}$ are varied over a fine grid). However, there are more intelligent and systematic algorithms than this naive approach. These algorithms concentrate on searching parts of the Pareto boundary that give large values on $f(\cdot)$.

This section describes two general algorithms for solving monotonic problems[15]: the *polyblock outer approximation (PA) algorithm* from [218, 274] and the *branch-reduce-and-bound (BRB) algorithm* from [275]. Both algorithms are designed to iteratively improve a lower bound $f_{\min}$ and an upper bound $f_{\max}$ on the optimal value of (2.44). Convergence to the global optimum will be guaranteed in the sense that

$$f_{\max} - f_{\min} < \varepsilon \tag{2.45}$$

is achieved in finitely many iterations, for any accuracy $\varepsilon > 0$. The algorithms also find an *$\varepsilon$-optimal solution* $\mathbf{g}_\varepsilon^*$, which is a feasible point with $f_{\min} = f(\mathbf{g}_\varepsilon^*)$. In general, the number of iterations scales exponentially with the number of users $K_r$, which is an inescapable consequence of solving a problem that generally is NP-hard (see Subsection 2.2.5).

---

**Remark 2.10 (Importance of Lipschitz Continuity).** The system utility function is assumed to be Lipschitz continuous (see Definition 1.13), which provides a limit on how fast the function varies. If the function is also differentiable, [277, Theorem 4] shows that the brute force approach[16] has a worst-case complexity of $c_{K_r}(\frac{L_f}{\varepsilon})^{K_r}$, where $L_f$ is the Lipschitz constant and $c_{K_r}$ is a constant that depends on the number of users. This provides an upper bound on the run time for any sensible algorithm — the PA and BRB algorithms have much faster convergence [26].

Lipschitz continuity is a sufficient condition for guaranteeing an $\varepsilon$-optimal solution in a finite number of iterations, but other assumptions that involve bounded derivatives are also possible; see [277]. However, if we do not impose any restrictions on $f(\cdot)$ then we generally cannot even find an $\varepsilon$-optimal solution in finite time [39].

---

[15] This tutorial describes adaptations of the PA and BRB algorithms that utilize specific properties of the resource allocation problem in (2.1) and (2.44). We refer to [218, 274, 275] for the generic algorithms that solve any monotonic problem.

[16] This corresponds to placing a fine grid over $[\mathbf{0}, \mathbf{u}]$ where the distance between adjacent points are $L_f/\varepsilon$. The performance and feasibility of each grid point need to be checked.

### 2.3.1 Lower and Upper Bounds in a Box

An essential step in the PA and BRB algorithms is that of bounding the highest feasible performance in a box $\mathcal{M} = [\mathbf{a}, \mathbf{b}] \subseteq \mathbb{R}_+^{K_r}$. This means finding a lower bound $f_{\min,\mathcal{M}}$ and an upper bound $f_{\max,\mathcal{M}}$ on the optimal solution to

$$\begin{aligned} &\underset{\mathbf{g}}{\text{maximize}} \ f(\mathbf{g}) \\ &\text{subject to } \mathbf{g} \in \mathcal{R} \cap \mathcal{M}. \end{aligned} \tag{2.46}$$

By utilizing that $f(\cdot)$ is increasing, the trivial bounds are

$$f_{\min,\mathcal{M}}^{\text{trivial}} = \begin{cases} f(\mathbf{a}), & \mathcal{R} \cap \mathcal{M} \neq \emptyset, \\ -\infty, & \text{otherwise}, \end{cases} \quad f_{\max,\mathcal{M}}^{\text{trivial}} = \begin{cases} f(\mathbf{b}), & \mathcal{R} \cap \mathcal{M} \neq \emptyset, \\ -\infty, & \text{otherwise}. \end{cases} \tag{2.47}$$

These bounds represent the performance in the lower and upper corners of the box, but only if the box has a nonempty overlap with the performance region — this is equivalent to $\mathbf{a} \in \mathcal{R}$, which is easily checked by solving the feasibility problem (2.29) with $\mathbf{a}$ as the QoS requirements.

As will become clear later, tighter bounds than (2.47) are necessary in the PA algorithm and will improve the convergence of the BRB algorithm. Any Pareto optimal point $\mathbf{g}' \in \partial^+\mathcal{R} \cap \mathcal{M}$ might give a reasonable lower bound, while an upper bound can be achieved by projecting $\mathbf{g}'$ onto the different outer sides of the box. This bounding procedure is formalized as follows and illustrated in Figure 2.7.

---

**Lemma 2.14.** Consider a box $\mathcal{M} = [\mathbf{a}, \mathbf{b}] \subseteq \mathbb{R}_+^{K_r}$ with $\mathcal{M} \cap \mathcal{R} \neq \emptyset$ and a strictly increasing curve $\mathbf{r}(\tau)$ satisfying $\mathbf{r}(0) = \mathbf{a}$ and $\mathbf{r}(\tau^{\text{upper}}) = \mathbf{b}$ for some given $\tau^{\text{upper}} > 0$. The feasible performance in $\mathcal{M}$ can be lower and upper bounded as

$$\begin{aligned} f_{\min,\mathcal{M}} &= f(\mathbf{n}) \\ f_{\max,\mathcal{M}} &= \max_k f(\underbrace{\mathbf{b} - [\mathbf{b} - \mathbf{m}]_k \mathbf{e}_k}_{=\mathbf{z}_k}), \end{aligned} \tag{2.48}$$

where $\mathbf{e}_k$ denotes the $k$th column of $\mathbf{I}_{K_r}$, $\mathbf{n} = \mathbf{r}(\tau_{\text{final}}^{\text{lower}})$ and $\mathbf{m} = \mathbf{r}(\tau_{\text{final}}^{\text{upper}})$ with $[\tau_{\text{final}}^{\text{lower}}, \tau_{\text{final}}^{\text{upper}}]$ being the final interval when solving (2.40) using Algorithm 1 (for some given line-search accuracy $\delta > 0$).

---

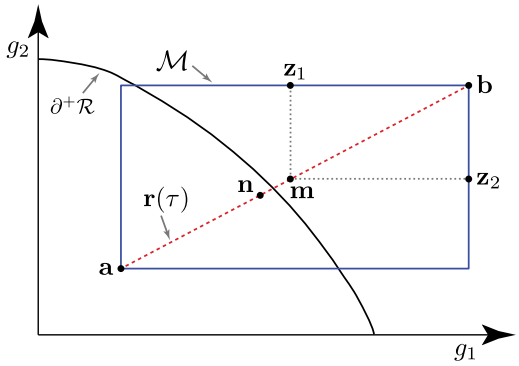



Fig. 2.7 Illustration of the bounding procedure in Lemma 2.14. The line-search between $\mathbf{a}$ and $\mathbf{b}$ results in a feasible point $\mathbf{n}$ and an infeasible point $\mathbf{m}$. The points $\mathbf{z}_k$ give upper bounds and are computed from $\mathbf{m}$.

*Proof.* The curve-search procedure in Algorithm 1 provides a final interval $[\tau_{\text{final}}^{\text{lower}}, \tau_{\text{final}}^{\text{upper}}]$, where the lower bound gives a feasible point $\mathbf{n} \in \mathcal{R}$. Every feasible point, including $\mathbf{n}$, gives a lower bound on the optimal solution. There are no feasible points $\mathbf{g} \in \mathcal{M}$ with $\mathbf{g} > \mathbf{m}$ as $\mathcal{R}$ is normal. The extreme points in $\mathcal{M}$ where all elements but one are larger than in $\mathbf{m}$ are $\mathbf{z}_k = \mathbf{b} - [\mathbf{b} - \mathbf{m}]_k \mathbf{e}_k$, for $k = 1, \ldots, K_r$, and can potentially be feasible. Thus, $\max_k f(\mathbf{z}_k)$ provides an upper bound on the feasible performance in $\mathcal{M}$. □

This lemma bounds the performance by searching on an increasing curve that connects the lower and upper corners of the box. Traditionally, this curve is a simple line $\mathbf{r}(\tau) = \mathbf{a} + \tau \frac{\mathbf{b}-\mathbf{a}}{\|\mathbf{b}-\mathbf{a}\|_1}$ with $\tau \in [0, \|\mathbf{b} - \mathbf{a}\|_1]$, which corresponds to the FPO problem in Example 2.8.[17] This line-search approach was suggested in [218, 274, 275] and utilized for multi-cell resource allocation with single-antenna interference channels in [206], coordinated MISO beamforming in [153, 276], and general multi-cell MISO systems in [26]. Other types of curves $\mathbf{r}(\tau)$ can also be used to capture certain properties of $f(\cdot)$ — one should always try to utilize any structure that exists in the problem.

[17] The line can be defined using other normalizations than $\|\mathbf{b} - \mathbf{a}\|_1$, but the $L_1$-norm is suitable in our applications because the aggregate performance of an operating point $\mathbf{n}$ is given by $\|\mathbf{n}\|_1$.

Lemma 2.14 can be applied whenever the feasibility problem (2.29) can be solved efficiently; Section 4 shows that this is possible under more general system conditions than assumed in Section 1.

---

**Remark 2.11 (Other Bounding Procedures).** Bounding the performance in a box is the most important and difficult step in monotonic optimization. Therefore, it is of profound importance to exploit any additional structure that might exist in the problem formulation. Instead of making a single curve-search in $\mathcal{M} \cap \mathcal{R}$ (as in Lemma 2.14), [123, 292, 293] suggest solving the $\epsilon$-constraint optimization problem (see Example 2.6) for each user with $\epsilon_k = a_k$ for the others. This approach might enable tighter bounds than Lemma 2.14, but generally has higher computational complexity (due to the $K_r$ optimization procedures at each iteration).

It is also possible to make a change of variables and thereby consider the intersection of a box $\mathcal{M}$ with some other $K_r$-dimensional region that is normal. In the two-user scenario, a region based on a beamforming parametrization is used in [118], which enables very efficient line-search (see also Example 3.1). If the system utility and user performance functions are concave, then a region based on interference constraints is taken in [215], which enables bounding operations based on interference-constrained beamforming (see Subsection 2.2.1).

In the case of weighted sum information rate optimization, the cost function represents the difference of two convex functions [272]. This property is utilized in [3, 133, 301] for single-antenna transmitters and simple power constraints. The region can then be based on power allocation coefficients and the bounding procedure consists of a sequence of approximate convex problems, which has much faster convergence than the general multi-antenna case. This approach can also be applied in the MISO case [67], but only under simple power constraints where so-called SINR balancing can be efficiently solved using interference functions [227, 228].

---

### 2.3.2 Polyblock Outer Approximation (PA) Algorithm

The globally optimal solution to (2.44) lies on the Pareto boundary $\partial^+\mathcal{R}$ of the performance region $\mathcal{R}$ (see Lemma 1.10). The PA

algorithm searches for the solution by approximating the region and iteratively refines the approximation. The algorithm is not applied directly onto (2.44) but on the perturbed problem

$$\begin{aligned} &\underset{\mathbf{g}}{\text{maximize}}\ \tilde{f}(\mathbf{g}) = f([\mathbf{g}-\mathbf{s}]_+ + \mathbf{s}), \\ &\text{subject to } \mathbf{g} \in \mathcal{R} \end{aligned} \tag{2.49}$$

for some parameter vector $\mathbf{s} > \mathbf{0}$. The operator $[\cdot]_+$ replaces negative elements with zero, thus $[\mathbf{g}-\mathbf{s}]_+ + \mathbf{s} \geq \mathbf{s}$ is guaranteed. The perturbed problem is approximately equivalent to (2.44) in the sense that $\tilde{f}(\mathbf{g}) - f(\mathbf{g}) \leq L_f \|\mathbf{s}\|_1$ for every $\mathbf{g} \in \mathcal{R}$, where $L_f$ is the Lipschitz constant of the system utility function. Solving (2.49) instead of (2.44) will therefore result in an error not exceeding $L_f\|\mathbf{s}\|_1$, which is manageable if $L_f\|\mathbf{s}\|_1 < \varepsilon$. Furthermore, if the optimal solution $\mathbf{g}^*$ to the original problem satisfies $\mathbf{g}^* \geq \mathbf{s}$, then the perturbation will not impact the solution accuracy. The reason for the perturbation is to prevent numerical convergence issues, for example, when searching for solutions close to an axis [275, 276]. We have the following result.

---

**Lemma 2.15.** Suppose that $\mathbf{g}_{\text{feasible}} \in \mathcal{R}$ and that $\mathbf{g}_{\text{upper}} \notin \mathcal{R}$ upper bounds the performance of the perturbed problem (2.49). If $\tilde{f}(\mathbf{g}_{\text{upper}}) - f(\mathbf{g}_{\text{feasible}}) < \varepsilon$, then $\mathbf{g}_{\text{feasible}}$ is an $\epsilon$-optimal solution to the original problem (2.44).

---

*Proof.* This lemma follows from the fact that $\tilde{f}(\mathbf{g}) \geq f(\mathbf{g})$ for all $\mathbf{g} \in \mathcal{R}$ and that both problems have the same feasible set. □

The PA algorithm solves (2.49) by approximating the feasible set $\mathcal{R}$ from above using polyblocks.

---

**Definition 2.6.** A set $\mathcal{P} \subseteq \mathbb{R}_+^n$ is called a *polyblock* if it is the union of a finite number of boxes $[\mathbf{0}, \mathbf{b}_m]$ with lower corners in the origin.

---

A polyblock $\mathcal{P}$ can be defined by a finite set of vertices $\mathcal{V} = \{\mathbf{b}_1, \ldots \mathbf{b}_{|\mathcal{V}|}\}$, which we write as $\mathcal{P}(\mathcal{V})$. The same polyblock can be expressed using different numbers of vertices $|\mathcal{V}|$, but we are only

interested in the minimal set called the *proper vertices*. In the proper representation, no vertex is dominated by another vertex (i.e., $\mathbf{b} \geq \mathbf{b}'$ does not hold for any $\mathbf{b} \neq \mathbf{b}' \in \mathcal{V}$).

As the system utility function is increasing, the maximum of $\tilde{f}(\mathbf{g})$ for $\mathbf{g} \in \mathcal{P}(\mathcal{V})$ is achieved at a proper vertex. This is the basic idea behind the PA algorithm; if $\mathcal{R}$ is approximated by a polyblock then the strong Pareto boundary is approximated by the proper vertices of this polyblock. This property is exploited as follows [218, 274]:

The PA algorithm constructs a nested sequence of polyblocks which approximates $\mathcal{R}$ from above as

$$\mathcal{P}(\mathcal{V}_1) \supset \mathcal{P}(\mathcal{V}_2) \supset \ldots \supset \mathcal{R} \quad \text{such that} \quad \max_{\mathbf{b}\in\mathcal{V}_n} \tilde{f}(\mathbf{b}) \searrow \max_{\mathbf{g}\in\mathcal{R}} \tilde{f}(\mathbf{g}), \quad (2.50)$$

where $x_n \searrow x$ means that $x_n \to x$ as $n \to \infty$ and that $x_n \geq x_{\tilde{n}} \geq x$ for all $\tilde{n} \geq n$. This approximation procedure is illustrated in Figure 2.8.

To realize the algorithm, we need a way to construct a new polyblock $\mathcal{P}(\mathcal{V}_{n+1})$ from the previous polyblock $\mathcal{P}(\mathcal{V}_n)$ such that the convergence in (2.50) is achieved. It makes sense to modify the vertex in $\mathcal{V}_n$ that provides the current maximum of $\tilde{f}(\cdot)$ over the polyblock:

$$\mathbf{g}^{(n)} = \arg\max_{\mathbf{b}\in\mathcal{V}_n} \tilde{f}(\mathbf{b}). \quad (2.51)$$

The following procedure is suggested in [274, Proposition 17–18].

---

**Lemma 2.16.** Consider the bounding procedure in Lemma 2.14 on the box $\mathcal{M}^{(n)} = [\mathbf{0}, \mathbf{g}^{(n)}]$ using

$$\mathbf{r}(\tau) = \tau \frac{\mathbf{g}^{(n)}}{\|\mathbf{g}^{(n)}\|_1} \quad \tau \in [0, \|\mathbf{g}^{(n)}\|_1]. \quad (2.52)$$

For a given line-search accuracy $\delta > 0$, Lemma 2.14 generates a feasible point $\mathbf{n}^{(n)}$ and a set of points $\{\mathbf{z}_k\}$ that upper bounds the feasible performance in $\mathcal{M}^{(n)}$. Then, the set of vertices

$$\mathcal{V}_{n+1} = (\mathcal{V}_n \backslash \{\mathbf{g}^{(n)}\}) \bigcup_{k:\,[\mathbf{g}^{(n)}]_k > 0} \{\tilde{\mathbf{z}}_k\}, \quad (2.53)$$

where $\tilde{\mathbf{z}}_k = \mathbf{z}_k - [\mathbf{s} - \mathbf{z}_k]_+$, satisfies $\mathcal{P}(\mathcal{V}_n) \supset \mathcal{P}(\mathcal{V}_{n+1}) \supset \mathcal{R}$.

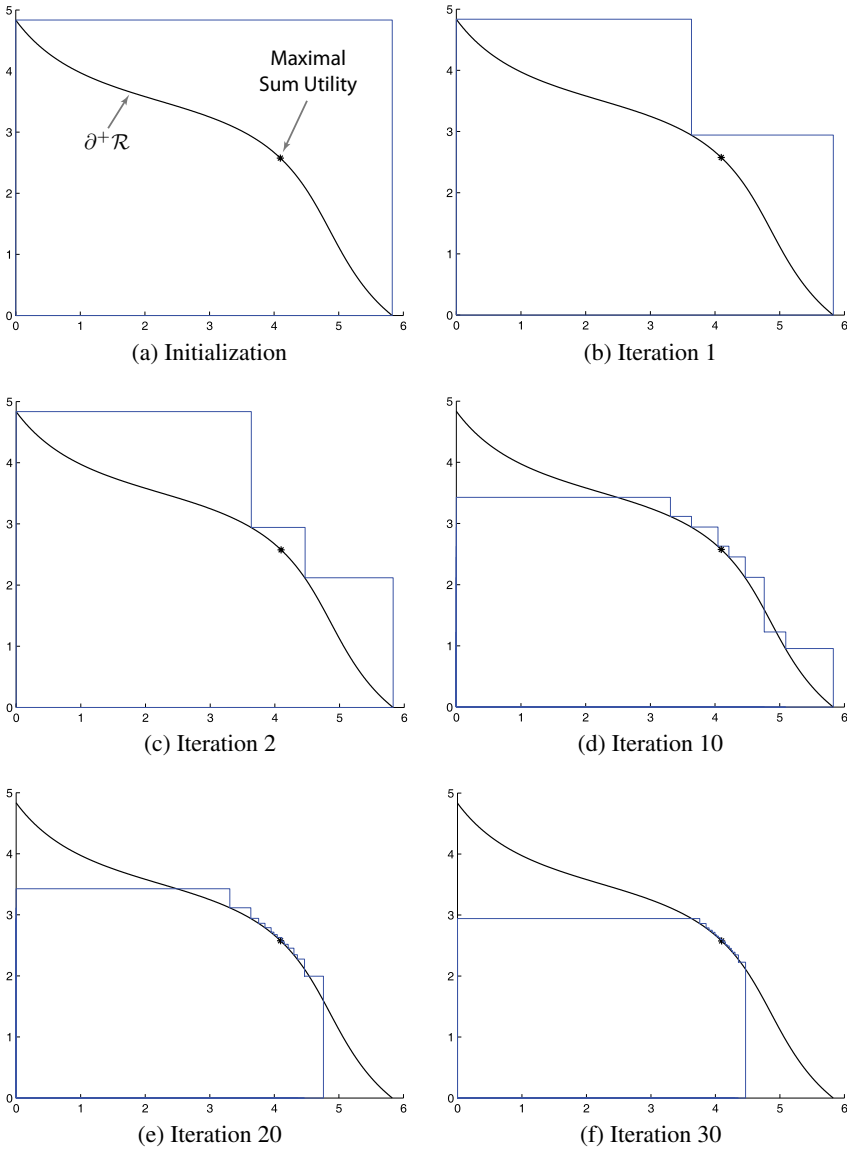



Fig. 2.8 Illustration of the Polyblock outer approximation (PA) algorithm. The sum information rate is maximized by approximating the performance region (around the optimal point) from above using a polyblock. The approximation is iteratively refined and parts that cannot contain the optimal point are removed.

If $\mathcal{V}_n$ only contains proper vertices, then $\mathcal{V}_{n+1}$ has the same property if improper vertices are removed using following the rule:
For every $\mathbf{g} \in \mathcal{V}_n \backslash \{\mathbf{g}^{(n)}\}$ such that $\mathbf{g} \geq \mathbf{m}$ while $[\mathbf{g}]_k < [\mathbf{g}^{(n)}]_k$ for exactly one element $k$, then remove $\tilde{\mathbf{z}}_k$ from $\mathcal{V}_{n+1}$.

---

*Proof.* The update (2.53) constructs a new polyblock by removing some (or all) of the overlap between $\mathcal{P}(\mathcal{V}_n)$ and $\{\mathbf{g} : \mathbf{g} > \mathbf{m}\}$. Since $\mathbf{m}$ is either Pareto optimal or outside $\mathcal{R}$, it follows that $\mathcal{P}(\mathcal{V}_{n+1}) \supset \mathcal{R}$. Observe that $\tilde{f}(\tilde{\mathbf{z}}_k) = \tilde{f}(\mathbf{z}_k)$, thus using $\tilde{\mathbf{z}}_k$ as vertex instead of $\mathbf{z}_k$ will not remove any optimal solution to the perturbed problem.

The deletion rule finds $\mathbf{g} \in \mathcal{V}_n \backslash \{\mathbf{g}^{(n)}\}$ such that $\mathbf{g} \geq \tilde{\mathbf{z}}_k$, thus $\tilde{\mathbf{z}}_k$ is necessarily improper. As $\mathcal{V}_n$ only contains proper vertices and $\tilde{\mathbf{z}}_k$ is only different from $\mathbf{g}^{(n)}$ in the $k$th element, the rule also provides a sufficient condition for improper vertices. □

This way of refining the polyblock is illustrated in Figure 2.9. To remove the shaded area from the polyblock, the number of vertices typically increases by each iteration (with at most $K_r - 1$). Although the increase is linear, it is sometimes necessary to take actions to overcome storage limitations; see [218, 274].

The algorithm can be initialized using the utopia point $\mathbf{u}$ (see Lemma 1.3) as the only vertex, $\mathcal{V}_1 = \{\mathbf{u}\}$, or in some other way that guarantees $\mathcal{P}(\mathcal{V}_1) \supset \mathcal{R}$. The best feasible point $\mathbf{g}_{\text{feasible}}$ known beforehand is also used for initialization. This could, for example, be $\mathbf{g}_{\text{feasible}} = \mathbf{0}_{K_r \times 1}$ or something achieved from some suboptimal resource allocation algorithm (see Section 4.2).

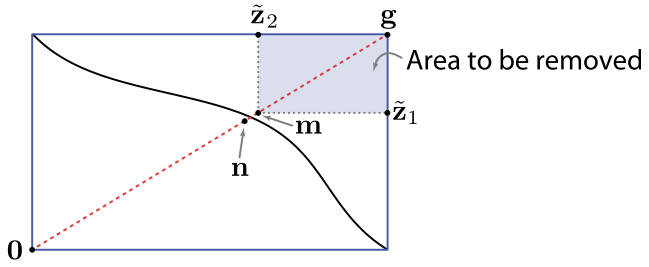



Fig. 2.9 The PA algorithm refines the polyblock in each iteration by removing the shaded area/volume which is outside the performance region $\mathcal{R}$. The refined polyblock is represented by replacing the vertex $\mathbf{g}$ by the new vertices $\tilde{\mathbf{z}}_1, \tilde{\mathbf{z}}_2$.

The bounds on the optimal value are refined in each iteration. At iteration $n$, the current upper bound is $f_{\max} = \max_{\mathbf{b}\in\mathcal{V}_n} \tilde{f}(\mathbf{b})$. Each vertex update using Lemma 2.16 includes a bounding procedure that gives a new feasible point. The current lower bound is the maximal performance among all the feasible points found so far: $f_{\min} = \max_{0\le\ell\le n} f(\mathbf{n}^{(\ell)})$. We can expect faster convergence in the lower bound (which is the *best* among $\mathbf{n}^{(0)},\ldots,\mathbf{n}^{(n)}$) than in the upper bound (which is the *worst* among the vertices in $\mathcal{V}_n$). Consequently, the algorithm usually finds a feasible point close to the optimal solution much earlier than we can formally declare that the point has this property.

The PA algorithm is summarized in Algorithm 2. This formulation is an adaptation of the generic PA algorithm to multi-cell resource allocation problems [218, 274]. It is basically a generalization of algorithms for single-antenna interference channels in [206] and for multi-antenna

**Algorithm 2**: Polyblock Outer Approximation (PA) Algorithm

**Result**: Solves the monotonic optimization problem in (2.44).
**Input**: Feasible solution $\mathbf{g}_{\text{feasible}}$ on (2.44);
**Input**: Solution accuracy $\varepsilon > 0$ and line-search accuracy $\delta > 0$;
**Input**: Initial vertex set $\mathcal{V}_1$ such that $\mathcal{P}(\mathcal{V}_1) \supset \mathcal{R}$;
Set $\mathbf{n}^{(0)} = \mathbf{g}_{\text{feasible}}$, $\mathbf{s} = \frac{\delta}{K_r}\mathbf{1}_{K_r}$, and $n = 1$;
Set $f_{\min} = f(\mathbf{n}^{(0)})$ and $f_{\max} = \max_{\mathbf{b}\in\mathcal{V}_1} \tilde{f}(\mathbf{b})$;
**while** $f_{\max} - f_{\min} > \varepsilon$ **do**
    Set $\mathbf{g}^{(n)} = \arg\max_{\mathbf{b}\in\mathcal{V}_n} \tilde{f}(\mathbf{b})$;
    Compute $\mathcal{V}_{n+1}$ according to Lemma 2.16 using $\mathcal{M}^{(n)} = [\mathbf{0}, \mathbf{g}^{(n)}]$. Obtain resulting feasible point $\mathbf{n}^{(n)}$;
    **if** $f(\mathbf{n}^{(n)}) > f_{\min}$ **then**
        Set $f_{\min} = f(\mathbf{n}^{(n)})$;
        Set $\mathbf{g}_{\text{feasible}} = \mathbf{n}^{(n)}$;
    Set $f_{\max} = \max_{\mathbf{b}\in\mathcal{V}_{n+1}} \tilde{f}(\mathbf{b})$;
    Remove all $\mathbf{b} \in \mathcal{V}_{n+1}$ with $\tilde{f}(\mathbf{b}) \le f_{\min} + \varepsilon$;
    Set $n = n + 1$;
**Output**: Final interval $[f_{\min}, f_{\max}]$ on optimal value;
**Output**: Feasible point $\mathbf{g}^*_\varepsilon = \mathbf{g}_{\text{feasible}}$ with $f_{\min} = f(\mathbf{g}^*_\varepsilon)$;

coordinated beamforming in [153, 276]. The convergence of the PA algorithm to the global optimum is established by the following theorem.

---

**Theorem 2.17.** For any given accuracy $\varepsilon > 0$, the PA algorithm finds an interval $[f_{\min}, f_{\max}]$ for the optimal value of (2.1) that satisfies $f_{\max} - f_{\min} \leq \varepsilon$, in a finite number of iterations. It is sufficient to have a line-search accuracy $0 < \delta < \frac{\varepsilon}{2L_f}$ and to set $\mathbf{s} = \frac{\delta}{K_r}\mathbf{1}_{K_r}$, where $L_f$ is the Lipschitz constant of $f(\cdot)$ in $[\mathbf{0}, \mathbf{u}]$.

---

*Proof.* The original proof in [274] assumed ideal line-search $\delta = 0$, but can be relaxed using the guidelines in [276]. Suppose for the purpose of contradiction that the algorithm requires infinitely many iterations, then it generates at least one infinite sequence of vertices $\mathbf{g}^{(n_1)}, \mathbf{g}^{(n_2)}, \ldots$ such that $\mathbf{g}^{(n_{h+1})} = \mathbf{g}^{(n_h)} - [\mathbf{g}^{(n_h)} - \mathbf{m}^{(n_h)}]_{k_h}\mathbf{e}_{k_h}$, where $\mathbf{m}^{(n_h)}$ is obtained from the bounding procedure and $k_h \in \{1, \ldots, K_r\}$. Clearly, $\mathbf{g}^{(n_1)} \geq \mathbf{g}^{(n_2)} \geq \ldots \geq \mathbf{0}$, thus the sequence converges to a limit point. Consequently, for any $\xi > 0$ it exists $h_\xi < \infty$ such that $\|\mathbf{g}^{(n_{h+1})} - \mathbf{g}^{(n_h)}\|_2 = [\mathbf{g}^{(n_h)} - \mathbf{m}^{(n_h)}]_{k_h} < \xi$ for all $h \geq h_\xi$.

This means that the difference between new and old vertices approaches zero. It remains to show that also the difference between $f(\mathbf{n}^{(n_h)})$ at the current feasible point $\mathbf{n}^{(n_h)}$ and the current maximum $\tilde{f}(\mathbf{g}^{(n_h)})$ goes below $\varepsilon > 0$, if $\delta$ is selected properly. Note that

$$\begin{aligned}
&\tilde{f}(\mathbf{g}^{(n_h)}) - f(\mathbf{n}^{(n_h)}) \\
&\quad \leq f(\mathbf{g}^{(n_h)}) - f(\mathbf{n}^{(n_h)}) + L_f\|\mathbf{s}\|_1 \\
&\quad = \big(f(\mathbf{g}^{(n_h)}) - f(\mathbf{m}^{(n_h)})\big) + \big(f(\mathbf{m}^{(n_h)}) - f(\mathbf{n}^{(n_h)})\big) + L_f\|\mathbf{s}\|_1 \\
&\quad \leq L_f\xi\frac{K_r\|\mathbf{u}\|_1}{\delta} + L_f\delta + L_f\delta,
\end{aligned} \tag{2.54}$$

where the inequalities follow from that $f(\cdot)$ is Lipschitz continuous, that geometrically $\|\mathbf{g}^{(n_h)} - \mathbf{m}^{(n_h)}\|_1 = \|\mathbf{g}^{(n_{h+1})} - \mathbf{g}^{(n_h)}\|_1 \frac{\|\mathbf{g}^{(n_h)}\|_1}{[\mathbf{g}^{(n_h)}]_{k_h}} < \xi\frac{\|\mathbf{u}\|_1}{\min_k[\mathbf{s}]_k}$, that $\mathbf{s} = \frac{\delta}{K_r}\mathbf{1}_{K_r}$, and that the line-search stops when $\|\mathbf{m}^{(n_h)} - \mathbf{n}^{(n_h)}\|_1 \leq \delta$. If we have $\delta < \frac{\varepsilon}{2L_f}$, then $2L_f\delta < \varepsilon$ and (2.54) becomes

$$\tilde{f}(\mathbf{g}^{(n_{h_\xi})}) - f(\mathbf{n}^{(n_{h_\xi})}) < \varepsilon \tag{2.55}$$

for some finite $h_\xi$ and $0 < \xi \leq \frac{(\varepsilon - 2L_f\delta)\delta}{L_f K_r \|\mathbf{u}\|_1}$, which is a contradiction. This implies $f_{\max} - f_{\min} \leq \varepsilon$ in finitely many iterations. □

The line-search accuracy in Theorem 2.17 is sufficient, but not necessary for convergence. Thus, a rougher accuracy can be used in Algorithm 2, at least initially. Although the algorithm converges, the worst-case convergence speed is generally exponential in the number of users $K_r$. The number of antennas $N$ and power constraints $L$ will however have much smaller impact on the convergence scaling of the PA algorithm, as it approximates the $K_r$-dimensional performance region. The main computational complexity lies in the bounding procedure,[18] which includes a quasi-convex line-search. In practice, it might be necessary to stop the algorithm before it converges, but fortunately $f_{\min}$ is usually closer to the true optimal value than $f_{\max}$ (as noted earlier).

---

**Remark 2.12 (Variations).** There are many variations on the PA algorithm that might improve the convergence speed: (a) An improved vertex update rule is suggested in [275, Proposition 4.2] to remove more in each iteration; (b) the line-search accuracy $\delta$ can be a function of the number of iterations and the vertex $\mathbf{g}^{(n)}$ [118, 276]; (c) the original problem can be perturbed in an adaptive manner to further avoid shallow cuts and jamming [39, 275]; (d) scheduling can be included in the algorithm [39, 276]; and (e) the algorithm can be restarted from the current best solution if the number of vertices grows too large [274].

---

### 2.3.3 Branch-Reduce-and-Bound (BRB) Algorithm

The PA algorithm constructs a series of polyblocks that covers and converges to the performance region $\mathcal{R}$. As the optimal solution lies on the Pareto boundary, it is sufficient to construct a set of boxes that closely approximates the Pareto boundary around the optimal point. This is essentially what is done in the BRB algorithm of [275]; see Figure 2.10.

[18] This is generally the case when solving nonconvex optimization problems.

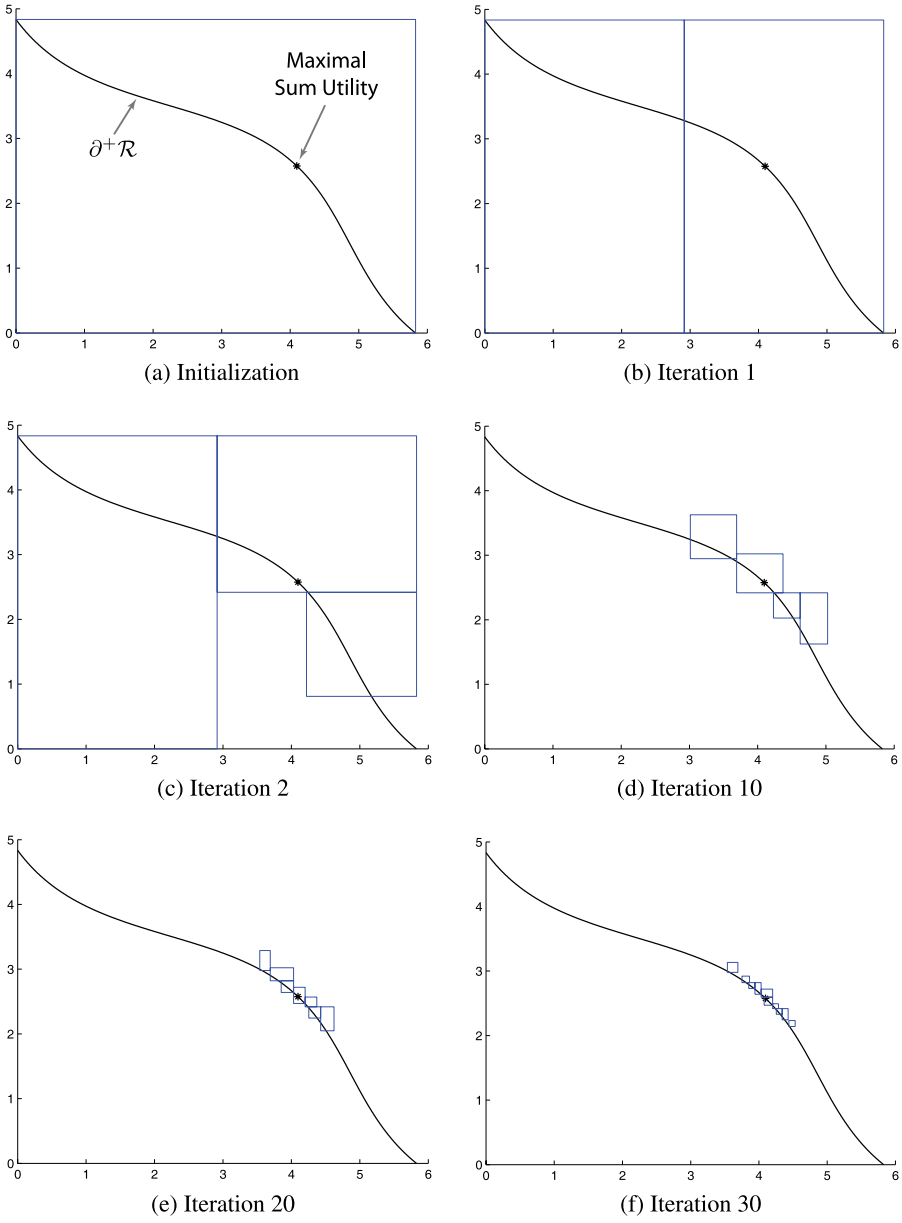



Fig. 2.10 Illustration of the Branch-reduce-and-bound (BRB) algorithm. The sum information rate is maximized by approximating the Pareto boundary of the performance region (around the optimal point) using a set of disjoint boxes. The approximation is iteratively refined and parts that cannot contain the optimal point are removed.

The BRB algorithm maintains a set $\mathcal{N}$ with nonoverlapping boxes that surely covers the parts of the performance region $\mathcal{R}$ where the optimal solutions lie (the solution might be nonunique). Iteratively, we split certain boxes and bounds the performance in these new boxes for the purpose of improving a lower bound $f_{\min}$ and an upper bound $f_{\max}$ on the optimal value of (2.44). To aid this process, a local feasible point $\mathbf{g}_{\mathcal{M}}$ and a local upper bound $\beta(\mathcal{M})$ are stored for each box $\mathcal{M} \in \mathcal{N}$.

Initially, $\mathcal{N} = \{\mathcal{M}_0\}$ for a box $\mathcal{M}_0 = [\mathbf{0}, \mathbf{b}_0] \subseteq \mathbb{R}_+^{K_r}$, where $\mathbf{b}_0$ could be the utopia point $\mathbf{u}$ or some other optimistic point that guarantees $\mathcal{R} \subseteq \mathcal{M}_0$. The initial upper bound is $f_{\max} = f(\mathbf{b}_0)$, while the lower bound is initialized as $f_{\min} = f(\mathbf{g}_{\text{feasible}})$ for some known feasible point $\mathbf{g}_{\mathcal{M}_0} = \mathbf{g}_{\text{feasible}}$ (e.g., $\mathbf{g}_{\text{feasible}} = \mathbf{0}_{K_r \times 1}$ or a point achieved from some suboptimal resource allocation algorithm; see Section 4.2).

Each iteration of the BRB algorithm consists of three steps.

(1) Branch: Divide a box $\mathcal{M}_{\max} \in \mathcal{N}$ into two new boxes.
(2) Reduce: Remove parts of these new boxes that cannot contain optimal solutions.
(3) Bound: Apply the bounding procedure in Lemma 2.14 to one of the new boxes, to improve local and global bounds.

These steps are illustrated in Figure 2.11 and are described next. Each iteration of the BRB algorithm modifies one of the boxes, which

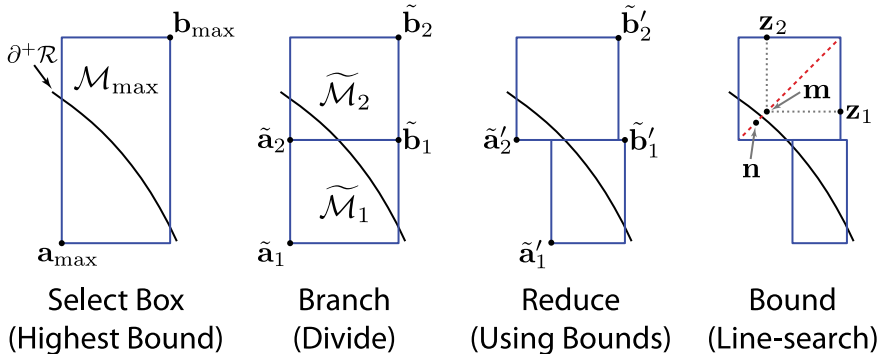



Fig. 2.11 An iteration of the BRB algorithm: A box is selected and branched into two new boxes. These are reduced based on the current bounds on the optimal value. Finally, line search between the lower and upper corners of the outermost box is applied to improve the bounds.

for convergence reasons is

$$\mathcal{M}_{\max} = \arg\max_{\mathcal{M}\in\mathcal{N}} \beta(\mathcal{M}), \tag{2.56}$$

where $\mathcal{M}_{\max} = [\mathbf{a}_{\max}, \mathbf{b}_{\max}]$ contains the current upper bound $f_{\max}$ (recall from the PA algorithm that upper bounds converge more slowly).

**Branch:** First, $\mathcal{M}_{\max}$ is divided into two disjoint boxes $\widetilde{\mathcal{M}}_1, \widetilde{\mathcal{M}}_2$. $\mathcal{M}_{\max}$ is bisected along its longest side (see Figure 2.11) which produces

$$\begin{aligned} \widetilde{\mathcal{M}}_1 &= [\mathbf{a}_{\max}, \mathbf{b}_{\max} - d\mathbf{e}_{\dim}], \\ \widetilde{\mathcal{M}}_2 &= [\mathbf{a}_{\max} + d\mathbf{e}_{\dim}, \mathbf{b}_{\max}], \end{aligned} \tag{2.57}$$

where $\dim = \text{argmax}_k[\mathbf{b}_{\max} - \mathbf{a}_{\max}]_k$, $d = [\mathbf{a}_{\max} + \mathbf{b}_{\max}]_{\dim}/2$, and $\mathbf{e}_k$ is the $k$th column of the identity matrix $\mathbf{I}_{K_r}$. The local feasible points and upper bounds of $\widetilde{\mathcal{M}}_1, \widetilde{\mathcal{M}}_2$ can be selected as follows.

---

**Lemma 2.18.** The local optimal performance in the new boxes can be lower bounded by the operating points

$$\begin{aligned} \mathbf{g}_{\widetilde{\mathcal{M}}_1} &= \begin{cases} \mathbf{g}_{\mathcal{M}_{\max}} - [\mathbf{g}_{\mathcal{M}_{\max}} - (\mathbf{b}_{\max} - d\mathbf{e}_{\dim})]_+, & \mathbf{g}_{\mathcal{M}_{\max}} \geq \mathbf{a}_{\max} + d\mathbf{e}_{\dim}, \\ \mathbf{g}_{\mathcal{M}_{\max}}, & \text{otherwise}, \end{cases} \\ \mathbf{g}_{\widetilde{\mathcal{M}}_2} &= \mathbf{g}_{\mathcal{M}_{\max}}, \end{aligned} \tag{2.58}$$

and the local upper bounds can be selected as

$$\begin{aligned} \beta(\widetilde{\mathcal{M}}_1) &= \min\left(\beta(\mathcal{M}_{\max}), f(\mathbf{b}_{\max} - d\mathbf{e}_{\dim})\right), \\ \beta(\widetilde{\mathcal{M}}_2) &= \beta(\mathcal{M}_{\max}). \end{aligned} \tag{2.59}$$

---

*Proof.* The feasible point $\mathbf{g}_{\mathcal{M}_{\max}}$ is either in $\widetilde{\mathcal{M}}_1$ or $\widetilde{\mathcal{M}}_2$. In the latter case, a feasible point in $\widetilde{\mathcal{M}}_1$ is achieved by projecting the point onto $\widetilde{\mathcal{M}}_1$ as $\mathbf{g}_{\mathcal{M}_{\max}} - [\mathbf{g}_{\mathcal{M}_{\max}} - \mathbf{b}_{\max} + d\mathbf{e}_{\dim}]_+$. The feasible performance in both boxes is upper bounded by $\beta(\mathcal{M}_{\max})$, but the upper corner of $\widetilde{\mathcal{M}}_1$ provides an alternative upper bound. □

The local feasible point $\mathbf{g}_{\widetilde{\mathcal{M}}_\ell}$ given by this lemma might be dominated by all points in $\widetilde{\mathcal{M}}_\ell$, which we return to in the bounding step.

**Reduce:** Next, the new boxes $\widetilde{\mathcal{M}}_\ell = [\tilde{\mathbf{a}}_\ell, \tilde{\mathbf{b}}_\ell]$ (for $\ell = 1, 2$) are reduced by removing parts that cannot contain the optimal solution; that is, parts that either give performance below the lower bound $f_{\min}$ or above the (local) upper bound $\beta(\widetilde{\mathcal{M}}_\ell)$. The following lemma from [26] shows how to replace $\widetilde{\mathcal{M}}_\ell$ with a (potentially) smaller box $[\tilde{\mathbf{a}}'_\ell, \tilde{\mathbf{b}}'_\ell]$.

---

**Lemma 2.19.** If $f_{\min} > \beta(\widetilde{\mathcal{M}}_\ell)$, then $\widetilde{\mathcal{M}}_\ell$ will not contain the optimal solution and can be removed. Otherwise, all points $\mathbf{g} \in [\tilde{\mathbf{a}}_\ell, \tilde{\mathbf{b}}_\ell]$ satisfying $f_{\min} \leq f(\mathbf{g}) \leq \beta(\widetilde{\mathcal{M}}_\ell)$ are also contained in $[\tilde{\mathbf{a}}'_\ell, \tilde{\mathbf{b}}'_\ell] \subseteq [\tilde{\mathbf{a}}_\ell, \tilde{\mathbf{b}}_\ell]$, where

$$\tilde{\mathbf{a}}'_\ell = \tilde{\mathbf{b}}_\ell - \sum_{k=1}^{K_r} \nu_{\ell k} [\tilde{\mathbf{b}}_\ell - \tilde{\mathbf{a}}_\ell]_k \mathbf{e}_k, \tag{2.60}$$

$$\tilde{\mathbf{b}}'_\ell = \tilde{\mathbf{a}}'_\ell + \sum_{k=1}^{K_r} \mu_{\ell k} [\tilde{\mathbf{b}}_\ell - \tilde{\mathbf{a}}'_\ell]_k \mathbf{e}_k, \tag{2.61}$$

with $\nu_{\ell k}$ and $\mu_{\ell k}$ (for $k = 1, \ldots, K_r$) calculated as

$$\begin{aligned} \nu_{\ell k} &= \max\left\{\nu :\ 0 \leq \nu \leq 1,\, f(\tilde{\mathbf{b}}_\ell - \nu[\tilde{\mathbf{b}}_\ell - \tilde{\mathbf{a}}_\ell]_k \mathbf{e}_k) \geq f_{\min}\right\} \\ \mu_{\ell k} &= \max\left\{\mu :\ 0 \leq \mu \leq 1,\, f(\tilde{\mathbf{a}}'_\ell + \mu[\tilde{\mathbf{b}}_\ell - \tilde{\mathbf{a}}'_\ell]_k \mathbf{e}_k) \leq \beta(\widetilde{\mathcal{M}}_\ell)\right\}. \end{aligned} \tag{2.62}$$

---

*Proof.* Consider the reduction of the box from $[\tilde{\mathbf{a}}_\ell, \tilde{\mathbf{b}}_\ell]$ to $[\tilde{\mathbf{a}}'_\ell, \tilde{\mathbf{b}}_\ell]$. If the boxes are identical, no solutions are lost and we are finished. Otherwise, $\tilde{\mathbf{a}}_\ell \leq \tilde{\mathbf{a}}'_\ell$ with strict inequality in at least one element. For elements with strict inequality we have $\nu_{\ell k} < 1$, thus it exists $\tilde{\nu}$ with $\nu_{\ell k} < \tilde{\nu} \leq 1$. Every such $\tilde{\nu}$ gives a point $\mathbf{g} \leq \tilde{\mathbf{b}}_\ell - \tilde{\nu}[\tilde{\mathbf{b}}_\ell - \tilde{\mathbf{a}}_\ell]_k \mathbf{e}_k$ in $\widetilde{\mathcal{M}}_\ell$ that by the selection of $\nu_{\ell k}$ in (2.62) gives $f(\mathbf{g}) < f_{\min}$. Therefore, the reduction (from below) only removes points with function values strictly below $f_{\min}$. The reduction from above is proved analogously. □

The reduction procedure is illustrated in Figure 2.11. Observe that it is two-step procedure: first, the lower point $\tilde{\mathbf{a}}_\ell$ is updated to $\tilde{\mathbf{a}}'_\ell$ using (2.60) and then $\tilde{\mathbf{a}}'_\ell$ is used to update the upper point $\tilde{\mathbf{b}}_\ell$ using (2.61). The parameters $\nu_{\ell k}, \mu_{\ell k}$ in (2.62) can be calculated using low-complexity line-search (it only involves evaluation $f(\cdot)$ for different parameter values, without caring about feasibility of the points).

Closed-form expressions can be obtained in many cases of practical interest.

---

**Example 2.11 (Reduction for Weighted Arithmetic Mean).** For the weighted arithmetic mean $f(\mathbf{g}) = \sum_{k=1}^{K_r} w_k g_k$, (2.62) is solved by

$$\begin{aligned} \nu_{\ell k} &= \min\left(\frac{\sum_{i=1}^{K_r} w_i [\tilde{\mathbf{b}}_\ell]_k - f_{\min}}{w_k([\tilde{\mathbf{b}}_\ell]_k - [\tilde{\mathbf{a}}_\ell]_k)}, 1\right), \\ \mu_{\ell k} &= \min\left(\frac{\beta(\widetilde{\mathcal{M}}_\ell) - \sum_{i=1}^{K_r} w_i [\tilde{\mathbf{a}}'_\ell]_k}{w_k([\tilde{\mathbf{b}}_\ell]_k - [\tilde{\mathbf{a}}'_\ell]_k)}, 1\right), \end{aligned} \tag{2.63}$$

where the min-operator makes sure that $\nu_{\ell k}, \mu_{\ell k} \leq 1$.

---

**Example 2.12 (Reduction for Weighted Geometric Mean).** For the weighted geometric mean $f(\mathbf{g}) = \prod_{k=1}^{K_r} g_k^{w_k}$, (2.62) is solved by

$$\begin{aligned} \nu_{\ell k} &= \min\left(\frac{[\tilde{\mathbf{b}}_\ell]_k - \left(\frac{f_{\min}}{\prod_{i\neq k}([\tilde{\mathbf{b}}_\ell]_i)^{w_i}}\right)^{\frac{1}{w_k}}}{[\tilde{\mathbf{b}}_\ell]_k - [\tilde{\mathbf{a}}_\ell]_k}, 1\right), \\ \mu_{\ell k} &= \min\left(\frac{\left(\frac{\beta(\widetilde{\mathcal{M}}_\ell)}{\prod_{i\neq k}([\tilde{\mathbf{a}}'_\ell]_i)^{w_i}}\right)^{\frac{1}{w_k}} - [\tilde{\mathbf{a}}'_\ell]_k}{[\tilde{\mathbf{b}}_\ell]_k - [\tilde{\mathbf{a}}'_\ell]_k}, 1\right), \end{aligned} \tag{2.64}$$

where the min-operator makes sure that $\nu_{\ell k}, \mu_{\ell k} \leq 1$.

---

The reduced new boxes are stored in $\mathcal{N}$, while $\mathcal{M}_{\max}$ is removed.

**Bound:** Each iteration ends by a search for better bounds. First, we check if there are any feasible points in $\widetilde{\mathcal{M}}_\ell = [\tilde{\mathbf{a}}'_\ell, \tilde{\mathbf{b}}'_\ell]$, or if $\widetilde{\mathcal{M}}_\ell \cap \mathcal{R} = \emptyset$.

---

**Lemma 2.20.** The intersection $\widetilde{\mathcal{M}}_\ell \cap \mathcal{R} \neq \emptyset$ if $\mathbf{g}_{\widetilde{\mathcal{M}}_\ell} \geq \tilde{\mathbf{a}}'_\ell$. Otherwise, the existence of feasible points in $\widetilde{\mathcal{M}}_\ell$ can be checked by solving the feasibility problem (2.29) with $\tilde{\mathbf{a}}'_\ell$ as the QoS requirements.

---

*Proof.* The first condition follows from that $\mathcal{R}$ is normal, while the second condition checks the feasibility explicitly. □

If the lemma concludes $\widetilde{\mathcal{M}}_\ell \cap \mathcal{R} = \emptyset$, then $\widetilde{\mathcal{M}}_\ell$ is deleted from $\mathcal{N}$.

If $\widetilde{\mathcal{M}}_2 \cap \mathcal{R} \neq \emptyset$, the BRB algorithm applies the bounding procedure in Lemma 2.14 using

$$\mathbf{r}(\tau) = \tilde{\mathbf{a}}'_2 + \tau \frac{(\tilde{\mathbf{b}}'_2 - \tilde{\mathbf{a}}'_2)}{\|\tilde{\mathbf{b}}'_2 - \tilde{\mathbf{a}}'_2\|_1} \quad \tau \in [0, \|\tilde{\mathbf{b}}'_2 - \tilde{\mathbf{a}}'_2\|_1] \tag{2.65}$$

as the search curve and using some line-search accuracy $\delta$. The normalization $\|\tilde{\mathbf{b}}'_2 - \tilde{\mathbf{a}}'_2\|_1$ ensures that the line-search accuracy is a global measure, thus the bounding procedure becomes faster as the boxes get smaller. The bounding procedure produces a feasible point $\mathbf{n} \in \widetilde{\mathcal{M}}_2$ and a local upper bound $f_{\max,\widetilde{\mathcal{M}}_2}$. The point $\mathbf{n}$ replaces the local feasible point if $f(\mathbf{n}) > f(\mathbf{g}_{\widetilde{\mathcal{M}}_\ell})$. Similarly, we set $\beta(\widetilde{\mathcal{M}}_2) = f_{\max,\widetilde{\mathcal{M}}_2}$ if $f_{\max,\widetilde{\mathcal{M}}_2} < \beta(\widetilde{\mathcal{M}}_2)$.

Finally, the global lower bound is updated as $f_{\min} = \max_{\mathcal{M}\in\mathcal{N}} f(\mathbf{g}_{\mathcal{M}})$ and the global upper bound is updated as $f_{\max} = \max_{\mathcal{M}\in\mathcal{N}} \beta(\mathcal{M})$. The stopping criterion $f_{\max} - f_{\min} < \varepsilon$ is checked at the end of each iteration.

The BRB algorithm is summarized in Algorithm 3 and illustrated in Figure 2.10. This formulation of the algorithm is a slight modification of the algorithm in [26], where the generic BRB algorithm from [275] is adapted for multi-cell resource allocation. Other adaptations are available in [123, 292, 293], where another bounding procedure is used. The system model of [123] is less general than [26], while [292, 293] are limited to single-antenna transmitters but can handle multi-cast transmissions (see Section 4.4). The convergence of the BRB algorithm to the global optimum was established in [275] and the following theorem originates from [26].

---

**Theorem 2.21.** For any given accuracy $\varepsilon > 0$, the BRB algorithm finds an interval $[f_{\min}, f_{\max}]$ for the optimal value of (2.1) that satisfies $f_{\max} - f_{\min} \leq \varepsilon$, in a finite number of iterations. The line-search accuracy $\delta > 0$ can be selected arbitrarily.

---

**Algorithm 3**: Branch-Reduce-and-Bound (BRB) Algorithm

**Result**: Solves the monotonic optimization problem in (2.44).
**Input**: Feasible solution $\mathbf{g}_{\text{feasible}}$ on (2.44);
**Input**: Solution accuracy $\varepsilon > 0$ and line-search accuracy $\delta > 0$;
**Input**: Initial box $\mathcal{M}_0 = [\mathbf{0}, \mathbf{b}_0]$ such that $\mathcal{R} \subseteq \mathcal{M}_0$;
Set $\mathcal{N} = \{\mathcal{M}_0\}$, $\mathbf{g}_{\mathcal{M}_0} = \mathbf{g}_{\text{feasible}}$, and $\beta(\mathcal{M}_0) = f(\mathbf{b})$;
Set $f_{\min} = f(\mathbf{g}_{\mathcal{M}_0})$ and $f_{\max} = \beta(\mathcal{M}_0)$;
**while** $f_{\max} - f_{\min} > \varepsilon$ **do**
  Set $\mathcal{M}_{\max} = \operatorname{argmax}_{\mathcal{M} \in \mathcal{N}} \beta(\mathcal{M})$;
  **for** $\ell = 1, 2$ **do**
    Create $\widetilde{\mathcal{M}}_\ell$ using (2.57) with $\mathbf{g}_{\widetilde{\mathcal{M}}_\ell}, \beta(\widetilde{\mathcal{M}}_\ell)$ in Lemma 2.18;
    Reduce $\widetilde{\mathcal{M}}_\ell$ using Lemma 2.19;
    Check feasibility of $\widetilde{\mathcal{M}}_2$ using Lemma 2.20;
    **if** *infeasible* **then**
      Set $\widetilde{\mathcal{M}}_\ell = \emptyset$;
  **if** $\widetilde{\mathcal{M}}_2 \neq \emptyset$ **then**
    Apply bounding procedure in Lemma 2.14 on $\widetilde{\mathcal{M}}_2 = [\tilde{\mathbf{a}}'_2, \tilde{\mathbf{b}}'_2]$ using $\mathbf{r}(\tau) = \tilde{\mathbf{a}}'_2 + \tau \frac{(\tilde{\mathbf{b}}'_2 - \tilde{\mathbf{a}}'_2)}{\|\tilde{\mathbf{b}}'_2 - \tilde{\mathbf{a}}'_2\|_1}$ and $\tau \in [0, \|\tilde{\mathbf{b}}'_2 - \tilde{\mathbf{a}}'_2\|_1]$;
    Obtain feasible point $\mathbf{n}$ and upper bound $f_{\max, \widetilde{\mathcal{M}}_2}$;
    **if** $f(\mathbf{n}) > f(\mathbf{g}_{\widetilde{\mathcal{M}}_2})$ **then**
      Set $\mathbf{g}_{\widetilde{\mathcal{M}}_2} = \mathbf{n}$;
    Set $\beta(\widetilde{\mathcal{M}}_2) = \min(\beta(\widetilde{\mathcal{M}}_2), f_{\max, \widetilde{\mathcal{M}}_2})$;
  Set $\mathcal{N} = (\mathcal{N} \setminus \mathcal{M}_{\max}) \cup \{\widetilde{\mathcal{M}}_1, \widetilde{\mathcal{M}}_2\}$;
  Set $f_{\min} = \max_{\mathcal{M} \in \mathcal{N}} f(\mathbf{g}_{\mathcal{M}})$;
  Set $f_{\max} = \max_{\mathcal{M} \in \mathcal{N}} \beta(\mathcal{M})$;
**Output**: Final interval $[f_{\min}, f_{\max}]$ on optimal value;
**Output**: Feasible point $\mathbf{g}^*_\varepsilon = \underset{\mathbf{g}_{\mathcal{M}}: \mathcal{M} \in \mathcal{N}}{\arg\max} f(\mathbf{g}_{\mathcal{M}})$;

*Proof.* The BRB algorithm can be treated as a standard branch-and-bound algorithm where the reduction step (which does not remove the solution) is part of the bounding step. Two sufficient conditions for achieving an $\varepsilon$-approximate solution in a finite number of iterations are stated in the appendix of [7]: (a) The bounding step truly calculates lower and upper bounds on the optimal value; and (b) The difference $f_{\max} - f_{\min}$ converges (uniformly) to zero. The first condition was proved in Lemma 2.14, and even the trivial bounds in (2.47) are sufficient so any $\delta > 0$ can be used. The second condition follows from the exhaustiveness of bisection and the Lipschitz continuity of $f(\cdot)$. □

Just as for the PA algorithm, the BRB algorithm converges to the global optimum in the sense of finding an interval $[f_{\min}, f_{\max}]$, with $f_{\max} - f_{\min} \leq \varepsilon$, in finitely many iterations (for any $\varepsilon > 0$). The important difference is that the BRB algorithm puts no requirements on the line-search accuracy $\delta$ to achieve convergence, thus $\delta$ can be selected solely on the basis of convergence speed. This is essentially because the BRB algorithm approximates the Pareto boundary from both below and above, while the PA algorithm approximates the whole performance region from above. Accordingly, the BRB algorithm has been claimed to have faster convergence than the PA algorithm [26, 275], or at least a better scaling with the number of users. On the other hand, both algorithms have a worst-case complexity that increases exponentially with the number of users $K_r$; thus, both algorithm are unsuitable for real-time applications and only practically useful for solving problems with a small number of users. The practical convergence of the two algorithms will be compared in the next section.

---

**Remark 2.13 (Variations).** The BRB algorithm can be modified in different ways that might improve the convergence speed: (a) The box $\mathcal{M}_{\max}$ can be branched into more than two boxes and the division rule can be adapted to the problem formulation [215]; (b) the line-search accuracy $\delta$ can be varied; and (c) the bounding procedure can be redesigned to find other (better) feasible points in the box [123, 292, 293].

---

To summarize, multi-cell resource allocation is generally a monotonic problem that can be solved to global optimality by the PA and BRB algorithms. These algorithms utilize that the performance region is normal and approximate the set of candidate solutions in an iteratively refined manner. The enabling factor of both algorithms is a bounding procedure that is solved efficiently using a curve-search procedure of the type in Subsection 2.2.3. Both algorithms can therefore be applied under whatever system conditions the curve-search is a quasi-convex problem. Several generalizations are provided in Section 4.

## 2.4 Numerical Illustrations of Computational Complexity

We end this section by illustrating the computational complexity of solving single-objective resource allocation problems. This section will emphasize the large difference between the convex problems described in Section 2.2 and the general monotonic problems solved in Section 2.3. For computational simplicity, we consider a simple coordinated beamforming scenario with $K_t = 2$ base stations with $N_1 = N_2 = 3$ antennas (see Example 1.2). Each base station serves two unique users (i.e., $K_r = 4$), while coordinating interference to all users. The average single-user SNR $\frac{\mathbb{E}\{q_j \|\mathbf{h}_{jk}\|_2^2\}}{\sigma_k^2}$ is $q_j N_j$ if user $k \in \mathcal{D}_j$ and $q_j \frac{N_j}{3}$ if $k \notin \mathcal{D}_j$, thus users are closer to their serving base station. Each base station has its own total power constraint with $q_j = 10$ (i.e., 10 dBm) and the information rate $g_k(\mathrm{SINR}_k) = \log_2(1 + \mathrm{SINR}_k)$ is used as user performance function.

The simulations in this tutorial are implemented using the modeling languages `CVX` [95] and `Yalmip` [161], which in turn utilize the convex optimization solvers `SeDuMi` [256] and `SDPT3` [271]. Parts of the Matlab code are available for download in [19].

### 2.4.1 Convergence Comparison

First, we compare and evaluate the convergence of the PA and BRB algorithms when solving two monotonic optimization problems: maximizing the arithmetic and geometric means of the information rates. These algorithms are rather different from each other and it is not

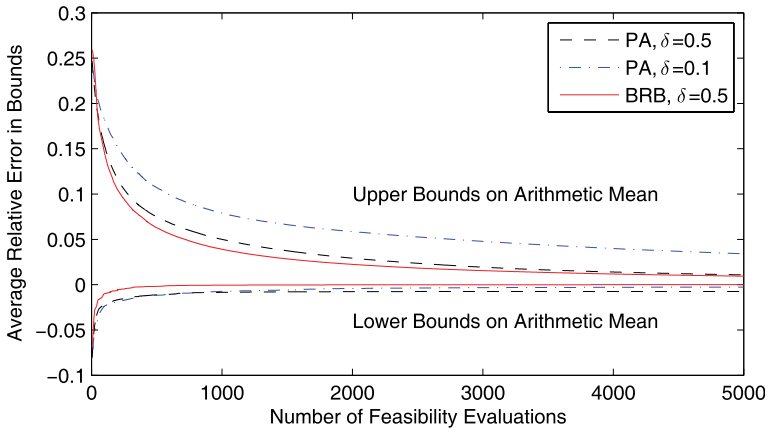
meaningful to compare the number of outer iterations. However, both algorithms are built upon solving a series of convex feasibility problems with QoS requirements (as in Subsection 2.2.2). We will therefore count the average number of such feasibility evaluations that is necessary to achieve certain accuracy on the optimal solution. The accuracy is defined as the relative deviations of the lower and upper bound: $\frac{f_{\min}-f_{\text{opt}}}{f_{\text{opt}}}$ and $\frac{f_{\max}-f_{\text{opt}}}{f_{\text{opt}}}$, respectively, where $f_{\text{opt}}$ is the optimal value.[19]

The arithmetic mean is maximized in Figure 2.12(a). The BRB algorithm is used with the line-search accuracy $\delta = 0.5$, while the PA algorithm is considered for $\delta = 0.1$ and $\delta = 0.5$. Both algorithms quickly find feasible solutions within a few percentages from the optimal value, but many feasibility evaluations are required to achieve a tight upper bound. This is a typical behavior when solving nonconvex problems (see Subsection 2.3.2), thus the search for better upper bounding techniques is an important topic for future research. The BRB algorithm has a clearly faster convergence (particularly in the upper bound) and thus requires fewer evaluations to achieve a certain $\varepsilon$-approximate solution.

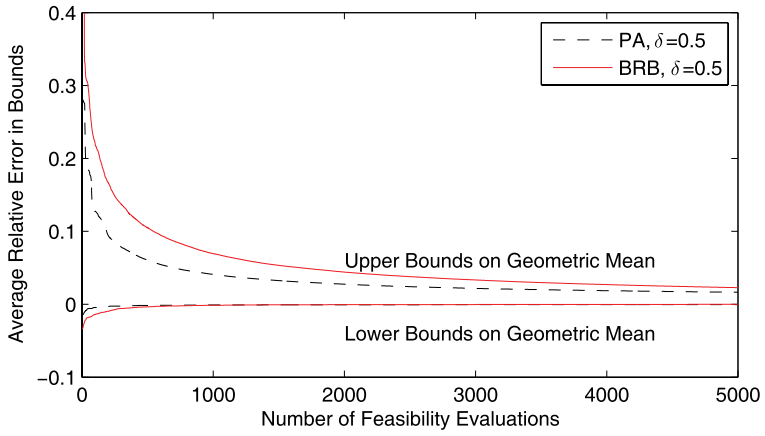
Recall from Theorems 2.17 and 2.21 that the line-search accuracy $\delta$ has a fundamental impact on convergence of the PA algorithm, while the BRB algorithm converges for any $\delta$. This is manifested in Figure 2.12(a) by an accuracy in the lower bound of the PA algorithm that improves as $\delta$ is decreased. Unfortunately, the convergence of the upper bound is improved by having a rougher line-search accuracy, which means more outer iterations that make the volume of the polyblock decrease faster. In other words, there is a tradeoff in the selection of $\delta$.

The geometric mean is maximized in Figure 2.12(b). The PA and BRB algorithms are compared with the line-search accuracy $\delta = 0.5$. Interestingly, the convergence behavior is very different from what we observed for the arithmetic mean; the PA algorithm has a clearly better convergence in the upper bound, while the lower bounds behave very similar. Observe that $\delta$ have been selected to (roughly) optimize the convergence speed, thus the advantage of the PA algorithm is not the

[19] The optimal value is approximately achieved in this simulation by running the algorithms for a very long time.




(a) Maximize Arithmetic Mean




(b) Maximize Geometric Mean

Fig. 2.12 Relative error of the lower and upper bounds on the optimal value as a function of the number of feasibility evaluations. The PA and BRB algorithms are compared for maximizing (a) the arithmetic mean and (b) the geometric mean of the information rates.

result of bad parameter selection. Instead, we believe that the slow convergence for the arithmetic mean depends on the possibility that the solution is very close to an axis (i.e., the scenario when the perturbed problem formulation in (2.49) is required to avoid stalling in the PA algorithm), which is seldom the case for the geometric mean (and other system utility functions that guarantees a nonzero performance level for all users).

To summarize, the BRB algorithm is superior for the arithmetic mean, while the PA algorithm might be the better choice for system utilities that enforces distinctively nonzero performance for all users (and thereby avoids the weaknesses of the PA algorithm).

### 2.4.2 Comparison of System Utility Functions

In addition to the arithmetic and geometric means, max-min fairness is an important system utility function. Figure 2.13 shows the convergence of the lower and upper bounds for max-min fairness, under the same conditions as in Figure 2.12. The difference is really remarkable; the relative deviation after 5000 feasibility evaluations is 0.01–0.02 for the arithmetic and geometric means, while only 14 evaluations are needed to surpass this accuracy under max-min fairness. Furthermore, the lower and upper bounds converge uniformly for max-min fairness, which is not the case for general monotonic problems. The convergence of the BRB and PA algorithms can certainly be improved (see Remarks 2.12 and 2.13), but the polynomial complexity of max-min fairness and exponential complexity of general monotonic problems imply that there is a fundamental and inescapable difference in convergence.

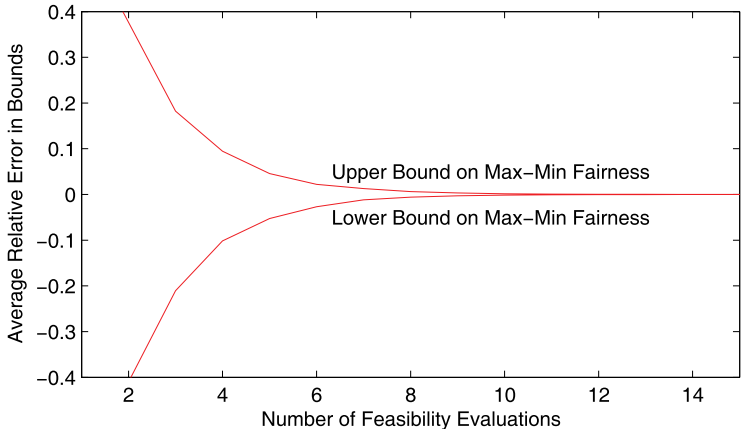



Fig. 2.13 Relative error of the lower and upper bounds on the optimal max-min fairness as a function of the number of feasibility evaluations.

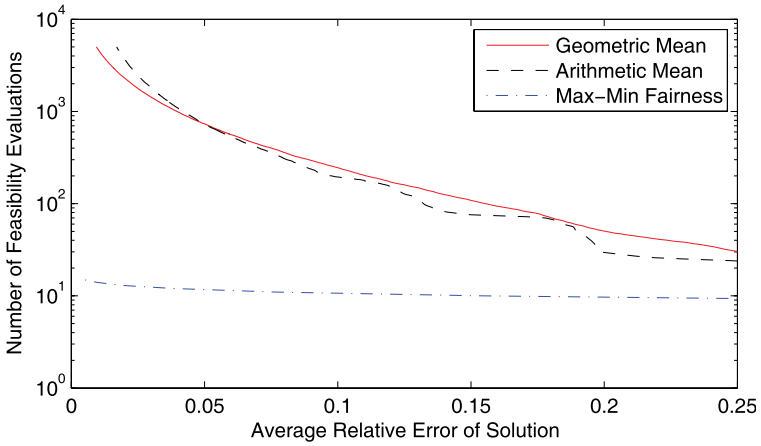



Fig. 2.14 Average relative error on optimal value of the geometric mean, arithmetic mean, and max-min fairness, as a function of the number of feasibility evaluations (in log-scale).

The importance of formulating single-objective resource allocation in a computationally efficient way is further emphasized in Figure 2.14. This figure shows the average relative error on the optimal value, $\frac{f_{\max}-f_{\min}}{f_{\text{opt}}}$, as a function of the number of feasibility evaluations (in logarithmic scale). We summarize the results from Figures 2.12 and 2.13 by showing the convergence of the best scheme (BRB for the arithmetic mean and PA for the geometric mean). The arithmetic mean seems to be somewhat easier to maximize than the geometric mean, at least for most relative errors. However, both system utility functions have much worse convergence than max-min fairness. This is particularly evident in the slope of the curves, which indicate the difference between polynomial and exponential complexity.

## 2.5 Summary

Convex optimization problems can be solved relatively efficiently; the optimal solution is found in polynomial time. It is therefore desirable to identify when single-objective resource allocation problems are convex. In general, these problems are not convex but belong to the wider category of monotonic problems that are more difficult to solve. However, convexity arises when the problem formulation clearly limits

the search-space for optimal solutions (see Section 2.2). This happens when the QoS requirements are fixed or varied over an increasing curve (e.g., for weighted max-min fairness), under zero-forcing constraints, and under single-antenna coordinated beamforming. These special cases are efficiently solved by interior-point methods (e.g., using the implementations `SeDuMi` [256] and `SDPT3` [271]) and are useful in practical applications.

Except from these special cases, most resource allocation problems of practical interest have been proved to be NP-hard [104]. However, these problems can be solved to global optimality using algorithms specifically developed for monotonic problems. The PA and BRB algorithms have been described in Section 2.3. Both algorithms iteratively approximate the performance region and improve bounds on the optimal value. Each bounding procedure is solved efficiently by formulating it as a resource allocation problem that belongs to one of the special convex cases. Convergence to the global optimum is guaranteed in a finite number of iterations, but the computational complexity is unsuitable for real-time applications. The solutions are however useful for offline benchmarking.

# 3

# Structure of Optimal Resource Allocation

This section will devise efficient ways of handling the general multi-objective resource allocation problem in (1.35):

$$\begin{aligned} &\underset{\mathbf{v}_1,\ldots,\mathbf{v}_{K_r}}{\text{maximize}} \ \{g_1(\text{SINR}_1),\ldots,g_{K_r}(\text{SINR}_{K_r})\} \\ &\text{subject to } \text{SINR}_k = \frac{|\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_k \mathbf{v}_k|^2}{\sigma_k^2 + \sum_{i\neq k} |\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_i \mathbf{v}_i|^2} \quad \forall k, \\ &\qquad \sum_{k=1}^{K_r} \mathbf{v}_k^H \mathbf{Q}_{lk} \mathbf{v}_k \leq q_l \quad \forall l. \end{aligned} \tag{3.1}$$

There are many reasons why (3.1) is difficult to solve. The most important might be: (a) conflicting interests of users; (b) strong inter-user coupling caused by interference; (c) performance region is generally nonconvex; (d) a large set of feasible beamforming vectors $\mathbf{v}_1,\ldots,\mathbf{v}_{K_r}$; and (e) the nonconvexity of most scalarizations of (3.1), as shown in Section 2. These factors are all associated with having too many degrees-of-freedom available to optimize a fuzzy performance objective.

To tackle these troubles, Section 3.1 measures the size of the search-space for beamforming vectors, while Section 3.2 presents some

state-of-the-art beamforming parametrizations that reduce the search-space and provide valuable insight on the structure of optimal resource allocation. The set of tentative solutions to (3.1), the Pareto boundary, is parameterized in Section 3.3, based on either weighted max-min fairness or beamforming parametrizations. Both approaches have inherent benefits and drawbacks. The beamforming parameterizations are then utilized in Section 3.4 to explain when and why heuristic approaches based on maximum ratio transmission (MRT), zero-forcing beamforming (ZFBF), and signal-to-leakage-and-noise ratio (SLNR) maximization are close-to-optimal. The section is concluded in Section 3.5 by returning to the four general methods for solving (3.1) that were outlined in Section 1.6. By combining the beamforming parametrizations (from this section) and the experience on which single-objective problems that are efficiently solvable (from Section 2), we provide general guidelines for solving (3.1) in practice.

Matlab codes for some of the examples that are given in this section are available for download in [19].

## 3.1 Limiting the Search-Space

From the problem formulation in (3.1), it seems that the search-space for optimal resource allocation consists of all feasible combinations of beamforming vectors $\mathbf{v}_1, \ldots, \mathbf{v}_{K_r}$. As each vector is $N$-dimensional, this corresponds to $K_r N$ complex-valued parameters. This is much less than selecting $K_r$ full $N \times N$ signal correlation matrices of arbitrary rank, thus showing the importance of utilizing the sufficiency of single-stream beamforming (proved in Section 1.5). The search-space can however be further reduced by utilizing the structure of the dynamic cooperation clusters. Observe that the actual beamforming is given by $\mathbf{D}_k \mathbf{v}_k$ and therefore only $\sum_{k=1}^{K_r} \mathrm{rank}(\mathbf{D}_k)$ complex-valued parameters are needed (and the rest can be set to zero).[1]

The beamforming vectors are also fundamentally connected with the channel vectors, as shown in [23, 119] under different conditions.

[1] The nonzero elements in $\mathbf{v}_k$ correspond to antennas at base stations that serve $\mathrm{MS}_k$.

**Lemma 3.1.** Under the per-transmitter power constraints in (1.10), it is sufficient to consider beamforming vectors $\mathbf{v}_k = [\mathbf{v}_{1k}^T \dots \mathbf{v}_{K_t k}^T]^T$ with

$$\mathbf{v}_{jk} \in \operatorname{span}\Big( \underbrace{\mathbf{D}_{jk}^H [\mathbf{C}_{j1}^H \mathbf{h}_{j1} \dots \mathbf{C}_{jK_r}^H \mathbf{h}_{jK_r}]}_{=\mathbf{F}_{jk}} \Big) \quad \forall j,k. \tag{3.2}$$

In particular, $\mathbf{v}_{jk} = \mathbf{0}_{N_j \times 1}$ for all $j,k$ such that $k \notin \mathcal{D}_j$. The operator $\operatorname{span}(\cdot)$ denotes the column space of a matrix.

*Proof.* The vector $\mathbf{v}_{jk}$ only appears in the SINR expressions as an inner product with the channels $\mathbf{D}_{jk}^H \mathbf{C}_{ji}^H \mathbf{h}_{ji}$ for all $i$, thus any power allocated outside $\operatorname{span}(\mathbf{F}_{jk})$ is wasted from a performance perspective. Under per-transmitter power constraints, power allocated outside $\operatorname{span}(\mathbf{F}_{jk})$ is also wasted from a power usage perspective. For $k \notin \mathcal{D}_j$, we have $\mathbf{D}_{jk} = \mathbf{0}_{N_j}$ and $\operatorname{span}(\mathbf{F}_{jk}) = \emptyset$, therefore $\mathbf{v}_{jk}$ is zero. □

This lemma states that every component $\mathbf{v}_{jk}$ of the beamforming vector $\mathbf{v}_k$ can be written as a linear combination of the channels that the signal passes through: $\mathbf{v}_{jk} = \sum_{m=1}^{\operatorname{rank}(\mathbf{F}_{jk})} \psi_{jkm} \boldsymbol{\upsilon}_{jkm}$ for some complex-valued coordinates $\psi_{jkm}$ and basis vectors $\boldsymbol{\upsilon}_{jkm}$ of the column space of $\mathbf{F}_{jk}$. This is very natural, since signal power transmitted in other directions is not received at neither the intended user nor the co-users.

Using Lemma 3.1, the beamforming vectors now depend on $\sum_{j=1}^{K_t} \sum_{k=1}^{K_r} \operatorname{rank}(\mathbf{F}_{jk})$ complex-valued parameters. This is strictly less than $\sum_{k=1}^{K_r} \operatorname{rank}(\mathbf{D}_k)$ whenever $|\mathcal{C}_j| < N_j$ for any of the base stations.[2] The basis vectors $\boldsymbol{\upsilon}_{jkm}$ can be selected arbitrarily in $\operatorname{span}(\mathbf{F}_{jk})$, but some intuition can be achieved using projection matrices.

**Definition 3.1 (Orthogonal Projection).** The orthogonal projection matrix $\boldsymbol{\Pi}_{\mathbf{X}}$ onto the column space of $\mathbf{X}$ is defined as $\boldsymbol{\Pi}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}^H\mathbf{X})^\dagger \mathbf{X}^H$. The orthogonal projection matrix onto the orthogonal complement of the column space of $\mathbf{X}$ is denoted $\boldsymbol{\Pi}_{\mathbf{X}}^\perp = \mathbf{I} - \boldsymbol{\Pi}_{\mathbf{X}}$.

[2] In fact, $\operatorname{rank}(\mathbf{F}_{jk}) \leq |\mathcal{C}_j|$ for $k \in \mathcal{D}_j$ with equality under very mild conditions on the stochastic generation of the channel vectors.

The basis vectors can be taken as the intended channel $\mathbf{D}_{jk}^H \mathbf{C}_{jk}^H \mathbf{h}_{jk}$ and the projection of it onto the orthogonal complement of the channel of co-user $i$, given by $\mathbf{\Pi}_{\mathbf{D}_{jk}^H \mathbf{C}_{ji}^H \mathbf{h}_{ji}}^{\perp} \mathbf{D}_{jk}^H \mathbf{C}_{jk}^H \mathbf{h}_{jk}$. The latter is a beamforming direction that will cause zero interference to co-user $i \neq k$ [23, 119]. This choice of basis vectors emphasizes that beamforming is a balance between selfishness (maximizing signal power) and altruism (minimizing the interference generated at co-users), which has important implications when a game theory perspective is applied to multi-cell systems [117, 140, 148]. This structure is particularly strong and intuitive in the two-user case, as shown by the following example.

---

**Example 3.1.** For the two-user MISO interference channel (i.e., $K_t = K_r = 2$ and $N_j \geq 2$) with per-transmitter constraints of $q_j = 1$, a simple and intuitive parametrization of all Pareto optimal beamforming vectors is provided in [117, 180]. Assuming that $\mathrm{BS}_j$ transmits to $\mathrm{MS}_j$ for $j = 1, 2$, the beamforming vector for $\mathrm{MS}_1$ is

$$\mathbf{v}_1(\lambda_1) = \sqrt{\lambda_1} \frac{\mathbf{\Pi}_{\mathbf{h}_{12}} \mathbf{h}_{11}}{\|\mathbf{\Pi}_{\mathbf{h}_{12}} \mathbf{h}_{11}\|} + \sqrt{1-\lambda_1} \frac{\mathbf{\Pi}_{\mathbf{h}_{12}}^{\perp} \mathbf{h}_{11}}{\|\mathbf{\Pi}_{\mathbf{h}_{12}}^{\perp} \mathbf{h}_{11}\|}, \tag{3.3}$$

where $\mathbf{h}_{jk}$ is the channel from $\mathrm{BS}_j$ to $\mathrm{MS}_k$. The projection matrices are defined in Definition 3.1 and the parametrization of $\mathbf{v}_2(\lambda_2)$ is analogous. The range of the parameters $\lambda_1$ and $\lambda_2$ are between zero and $\lambda_k^{(\mathrm{MRT})} = \frac{\|\mathbf{\Pi}_{\mathbf{h}_{ki}} \mathbf{h}_{kk}\|}{\|\mathbf{h}_{kk}\|}$ for $i \neq k$ (i.e., $\lambda_k \in [0, \lambda_k^{(\mathrm{MRT})}]$).

More intuition is achieved by rephrasing (3.3) in terms of maximum ratio transmission (MRT) and zero-forcing beamforming (ZFBF). We refer to Section 3.4 for the general definitions, but these beamforming directions are $\bar{\mathbf{v}}_k^{(\mathrm{MRT})} = \frac{\mathbf{h}_{kk}}{\|\mathbf{h}_{kk}\|}$ and $\bar{\mathbf{v}}_k^{(\mathrm{ZFBF})} = \frac{\mathbf{\Pi}_{\mathbf{h}_{ki}}^{\perp} \mathbf{h}_{kk}}{\|\mathbf{\Pi}_{\mathbf{h}_{ki}}^{\perp} \mathbf{h}_{kk}\|}$ in the two-user case ($k$ is the intended user and $i$ is the co-user). The parametrization in (3.3) can be reparameterized as

$$\mathbf{v}_1(\eta_1) = \frac{\sqrt{\eta_1} \bar{\mathbf{v}}_1^{(\mathrm{MRT})} + \sqrt{1-\eta_1} \bar{\mathbf{v}}_1^{(\mathrm{ZFBF})}}{\| \sqrt{\eta_1} \bar{\mathbf{v}}_1^{(\mathrm{MRT})} + \sqrt{1-\eta_1} \bar{\mathbf{v}}_1^{(\mathrm{ZFBF})} \|}, \tag{3.4}$$

while the parametrization of beamforming vector $\mathbf{v}_2(\eta_2)$ is analogous. The range of the parameters $\eta_1$ and $\eta_2$ is between zero and one (i.e., $\eta_1, \eta_2 \in [0,1]$). Beamforming is thus a balance between selfish MRT and altruistic ZFBF.

Based on the parametrization in (3.3), a closed-form expression for the performance region $\mathcal{R}$ was derived independently in [149] and [181, Theorem 2]. The achievable SINR values for both users can be written as a function of the parameters $\lambda_1$ and $\lambda_2$: $\mathrm{SINR}_1(\lambda_1,\lambda_2)$, $\mathrm{SINR}_2(\lambda_1,\lambda_2)$. The Pareto optimal points are given by some $\lambda_1^*, \lambda_2^*$ that satisfy the condition

$$\frac{\partial \mathrm{SINR}_1(\lambda_1,\lambda_2)}{\partial \lambda_1}\frac{\partial \mathrm{SINR}_2(\lambda_1,\lambda_2)}{\partial \lambda_2} = \frac{\partial \mathrm{SINR}_2(\lambda_1,\lambda_2)}{\partial \lambda_1}\frac{\partial \mathrm{SINR}_1(\lambda_1,\lambda_2)}{\partial \lambda_2}. \tag{3.5}$$

For any $\lambda_2$ in the feasible range, the corresponding $\lambda_1$ that satisfies this condition is obtained by solving (3.5) as an equation. The one-dimensional weak Pareto boundary is thus described by a function $p : [0, \lambda_2^{(\mathrm{MRT})}] \to [0, \lambda_1^{(\mathrm{MRT})}]$ that is obtained from (3.5). It is shown in [149, 181] that $p$ is a solution to the cubic polynomial equation

$$a\lambda_1^3 + b\lambda_1^2 + c\lambda_1 + d = 0 \tag{3.6}$$

with

$$\begin{aligned}
a &= -(\|\mathbf{\Pi}_{\mathbf{h}_{12}}\mathbf{h}_{11}\|^2 + \|\mathbf{\Pi}^{\perp}_{\mathbf{h}_{12}}\mathbf{h}_{11}\|^2)(C - \|\mathbf{h}_{12}\|^2)^2,\\
b &= (C - \|\mathbf{h}_{12}\|^2)(2\|\mathbf{\Pi}^{\perp}_{\mathbf{h}_{12}}\mathbf{h}_{11}\|^2(C+\sigma_1^2)\\
&\quad + \|\mathbf{\Pi}_{\mathbf{h}_{12}}\mathbf{h}_{11}\|^2(2\sigma_1^2 + C - \|\mathbf{h}_{12}\|^2)),\\
c &= -\|\mathbf{\Pi}^{\perp}_{\mathbf{h}_{12}}\mathbf{h}_{11}\|^2(C+\sigma_1^2)^2 + \sigma_1^2\|\mathbf{\Pi}_{\mathbf{h}_{12}}\mathbf{h}_{11}\|^2(2\|\mathbf{h}_{12}\|^2 - 2C - \sigma_1^2),\\
d &= \sigma_1^4\|\mathbf{\Pi}_{\mathbf{h}_{12}}\mathbf{h}_{11}\|^2,
\end{aligned}$$

and the constant $C$ as a function of $\lambda_2$ is given by

$$C(\lambda_2) = \frac{\sqrt{\lambda_2\|\mathbf{\Pi}_{\mathbf{h}_{21}}\mathbf{h}_{22}\|^2} + \sqrt{(1-\lambda_2)\|\mathbf{\Pi}^{\perp}_{\mathbf{h}_{21}}\mathbf{h}_{22}\|^2}}{\left(\sqrt{\frac{\|\mathbf{\Pi}_{\mathbf{h}_{21}}\mathbf{h}_{22}\|^2}{\lambda_2}} - \sqrt{\frac{\|\mathbf{\Pi}^{\perp}_{\mathbf{h}_{21}}\mathbf{h}_{22}\|^2}{1-\lambda_2}}\right)\left(\frac{\sigma_2^2}{\|\mathbf{h}_{21}\|^2} + \lambda_2^{(\mathrm{MRT})} - \lambda_2\right)}. \tag{3.7}$$

There are closed-form root expressions for cubic polynomials [110] and the root of interest in (3.6) lies in the interval $[0, \lambda_1^{(\mathrm{MRT})}]$ and satisfies

$$\begin{aligned}&\operatorname{sign}\left(\frac{\sigma_1^2}{\|\mathbf{h}_{12}\|^2} + \lambda_1 - C\lambda_1\right)\\ &\quad = \operatorname{sign}\left(\frac{\sigma_1^2}{\|\mathbf{h}_{12}\|^2} + \lambda_1 + C(1-\lambda_1)\right).\end{aligned} \tag{3.8}$$

For the two-user special case, all interesting operating points on the Pareto boundary can be found by traversing the closed-form characterization above. In particular, solving any scalarization of the MOP in (3.1) reduces to a one-dimensional line search.

---

This example shows that there are cases when the number of parameters that characterizes the beamforming vectors can be reduced far below what is stated in Lemma 3.1. The generality of this observation is further explored in the next section, where we derive necessary properties of the optimal beamforming.

## 3.2 Efficient Beamforming Parametrizations

The previous section showed that the search-space for beamforming vectors consists of at most $\sum_{k=1}^{K_r} \operatorname{rank}(\mathbf{D}_k)$ complex-valued parameters, thus the number of parameters depends strongly on the number of transmit antennas $N$. In this section, we present three state-of-the-art parameterizations that use the problem structure to substantially reduce the search-space for optimal beamforming. In particular, the parameters are positive real-valued (instead of complex-valued) and the number does not increase with $N$. The parametrizations also provide important structural insights that are utilized later in this tutorial to achieve both optimal and suboptimal beamforming.

### 3.2.1 Parametrization Based on Interference-Temperature

The first parametrization is based on adding new constraints to (3.1) that dictate how much interference power the transmission to $\mathrm{MS}_k$ is allowed to cause to $\mathrm{MS}_i$, for each $i \neq k$. These interference temperature constraints have the form $|\mathbf{h}_i^H \mathbf{C}_i \mathbf{D}_k \mathbf{v}_k|^2 \leq \Gamma_{ki}$ for some parameters

$\Gamma_{ki} \geq 0$ that are called *interference-temperature limits*. This terminology originates from underlay cognitive radio [102], where the interference-temperatures might be specified by a regulatory agency (at least for secondary systems). This topic is further described in Section 4.8.

In this subsection we allow for arbitrary selection of the parameters $\Gamma_{ki}$. In addition, we include the per-user power constraints in (1.5), where each power constraint is decomposed as $\mathbf{v}_k^H \mathbf{Q}_{lk} \mathbf{v}_k \leq q_{lk} \, \forall l, k$ and the parameters $q_{lk}$ satisfy $\sum_{k=1}^{K_r} q_{lk} \leq q_l \, \forall l$. The interference-temperature and per-user power constraints can transform (3.1) into

$$\begin{aligned} &\underset{\mathbf{v}_1,\ldots,\mathbf{v}_{K_r}}{\text{maximize}} \; \{g_1(\text{SINR}_1), \ldots, g_{K_r}(\text{SINR}_{K_r})\} \\ &\text{subject to } \text{SINR}_k = \frac{|\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_k \mathbf{v}_k|^2}{\sigma_k^2 + \sum\limits_{i \neq k} \Gamma_{ik}} \quad \forall k, \\ &\qquad\qquad |\mathbf{h}_i^H \mathbf{C}_i \mathbf{D}_k \mathbf{v}_k|^2 \leq \Gamma_{ki} \quad \forall k, i, \; i \neq k, \\ &\qquad\qquad \mathbf{v}_k^H \mathbf{Q}_{lk} \mathbf{v}_k \leq q_{lk} \quad \forall l, k, \end{aligned} \tag{3.9}$$

where $\text{SINR}_k$ is the exact SINR when all interference-temperature constraints are active — it is otherwise a lower bound.

This amended multi-objective optimization problem allows for decomposition into $K_r$ independent single-objective problems [235, 325].

---

**Theorem 3.2.** The unique optimum of the multi-objective problem in (3.9) is achieved by independently solving the $K_r$ convex problems

$$\begin{aligned} &\underset{\mathbf{v}_k}{\text{maximize}} \; |\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_k \mathbf{v}_k|^2 \\ &\text{subject to } |\mathbf{h}_i^H \mathbf{C}_i \mathbf{D}_k \mathbf{v}_k|^2 \leq \Gamma_{ki} \quad \forall i \neq k, \\ &\qquad\qquad \mathbf{v}_k^H \mathbf{Q}_{lk} \mathbf{v}_k \leq q_{lk} \quad \forall l. \end{aligned} \tag{3.10}$$

Furthermore, every strong Pareto optimal point $\mathbf{g} \in \partial \mathcal{R}$ is achieved by solving (3.9) in this manner for some nonnegative parameters $\{q_{lk}\}_{l=1,k=1}^{L,K_r}$ and $\{\Gamma_{ki}\}_{k=1,i=1,i\neq k}^{K_r,K_r}$.

---

*Proof.* The decomposition into single-user problems follows immediately, since each objective and constraint in (3.10) is only affected

by one of the beamforming vectors. Note that maximizing $g_k(\mathrm{SINR}_k)$ is equivalent to maximizing $|\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_k \mathbf{v}_k|^2$ in this case. Furthermore, assume that $\mathbf{v}_1^*, \ldots, \mathbf{v}_{K_r}^*$ achieves a certain strong Pareto optimal point $\mathbf{g}$. If we select the parameters as $\Gamma_{ki} = |\mathbf{h}_i^H \mathbf{C}_i \mathbf{D}_k \mathbf{v}_k^*|^2$ and $q_{lk} = (\mathbf{v}_k^*)^H \mathbf{Q}_{lk} \mathbf{v}_k^*$, then $\mathbf{v}_k^*$ must be a feasible and optimal solution to (3.10) (otherwise we contradict Definition 1.10 of strong Pareto optimality). □

Theorem 3.2 provides a parametrical characterization of the Pareto boundary of $\mathcal{R}$ and generalizes the prior work in [235, 325] that only considered interference channels. The number of parameters is clarified by the following corollary.

---

**Corollary 3.3.** The parametrization in Theorem 3.2 requires selecting $K_r(K_r - 1)$ nonnegative[3] interference-temperature limits and $\sum_{l=1}^{L}(\nu_l - 1)$ per-user power limits, where $\nu_l$ is the number of users affected by the $l$th power constraint (i.e., those with $\mathbf{Q}_{lk} \neq \mathbf{0}_N$).

---

*Proof.* Without loss of generality, we can set $q_{lk} = 0$ whenever $\mathbf{Q}_{lk} = \mathbf{0}_N$ and calculate one parameter as $q_{li} = q_l - \sum_{k \neq i} q_{lk}$ for each $l$. □

If all power constraints affect all users (i.e., $\nu_l = K_r \, \forall l$), then Corollary 3.3 shows that we need to select $(L + K_r)(K_r - 1)$ parameters in total. The other extreme is when each power constraint only affects one user (e.g., the interference channel [325] with $\nu_l = 1 \, \forall l$), then we only need to select $K_r(K_r - 1)$ parameters.

For each parameter selection, the corresponding beamforming vectors are calculated by solving $K_r$ single-user problems. These convex optimization problems will in general not have closed-form solutions (but can be solved by interior-point methods), thus Theorem 3.2 provides an indirect beamforming parametrization. A

[3] There is an upper bound on how much interference power that can be generated at a given user under the power constraints, but observe that the sets $[0, \infty)$ and $[0, c]$ for $0 < c < \infty$ are equal in terms of complexity (i.e., there are bijective functions between the sets).

closed-form parametrization can however be achieved by replacing the interference temperature constraints with an equal number of angles that geometrically specify the location of the beamforming vectors; we refer to [235] for details as it is hard to describe this approach mathematically.

It should also be noted that Theorem 3.2 only provides necessary conditions for achieving the Pareto boundary; it is unlikely to find exact Pareto optimal points by random parameter selection. The strength of the parametrization is that it decouples the multi-objective resource allocation into independent and convex single-user problems, which enables distributed algorithms where the parameters are iteratively updated to move toward the Pareto boundary. This is further described in Section 4.2 and in [325].

### 3.2.2 Parametrization Based on Channel Gain Regions

The components $\mathbf{v}_{jk}$ of the optimal beamforming vector $\mathbf{v}_k$ for user $k$ belong to subspaces only spanned by local CSI (i.e., channel vectors from $\text{BS}_j$ to users $k \in \mathcal{C}_j$) according to Lemma 3.1. The optimal choices within these subspaces depend however on the decisions taken by the other base stations. In other words, the main difficulty in multi-cell resource allocation is not the lack of global CSI, but the need for coordinated parameter selection and decision making.

A closed-form beamforming parametrization that simplifies coordination can be obtained directly from the channel gain regions (and the approach taken in Lemma 1.7).

---

**Theorem 3.4.** Each Pareto optimal point is achieved by beamforming vectors $\mathbf{v}_k(\boldsymbol{\lambda}_k, \boldsymbol{\mu}_k) = \sqrt{p_k}\bar{\mathbf{v}}_k$ with

$$\bar{\mathbf{v}}_k = \boldsymbol{v}_{\max}\left(\sum_{i=1}^{K_r} \lambda_{ki} e_{ki} \mathbf{D}_k^H \mathbf{C}_i^H \mathbf{h}_i \mathbf{h}_i^H \mathbf{C}_i \mathbf{D}_k - \sum_{l=1}^{L} \mu_{lk} \mathbf{Q}_{lk}\right), \tag{3.11}$$

$$p_k = \frac{\sum_{l=1}^{L} \mu_{lk} q_{lk}}{\sum_{l=1}^{L} \mu_{lk} \bar{\mathbf{v}}_k^H \mathbf{Q}_{lk} \bar{\mathbf{v}}_k}, \tag{3.12}$$

where $e_{ki} = \begin{cases} +1, & k = i, \\ -1, & k \neq i, \end{cases}$ and the operator $\boldsymbol{v}_{\max}$ gives the dominating unit-norm eigenvector.[4] The parameters $\lambda_{k1}, \ldots, \lambda_{kK_r}, \mu_{1k}, \ldots, \mu_{Lk} \geq 0$ are selected to satisfy $\sum_{i=1}^{K_r} \lambda_{ki} = 1$.

*Proof.* It is shown in Lemma 1.5 that the beamforming vectors which attain the Pareto boundary of the performance region also attain the boundary of the channel gain regions $\Omega_k$ in directions $\mathbf{e}_1, \ldots, \mathbf{e}_{K_r}$. In the proof of Lemma 1.7 it is shown that the boundary of $\Omega_k$ is achieved by a beamforming vector $\mathbf{v}_k$ which solves

$$\underset{\mathbf{v}_k}{\text{maximize}} \sum_{i=1}^{K_r} \lambda_i |\mathbf{h}_i^H \mathbf{C}_i \mathbf{D}_k \mathbf{v}_k|^2 \quad \text{subject to} \quad \mathbf{v}_k^H \mathbf{Q}_{lk} \mathbf{v}_k \leq q_{lk} \; \forall l. \quad (3.13)$$

In order to achieve the boundary of the channel gain region $\Omega_k$ in direction $\mathbf{e}_k = [-1 \ldots -1 \; +1 \; -1 \ldots -1]^T$ (with a plus one at element $k$), the weights $\lambda_1, \ldots, \lambda_{K_r}$ need to have the following signs

$$\text{sign}(\lambda_i) = \begin{cases} +1, & i = k, \\ -1, & \text{otherwise.} \end{cases} \quad (3.14)$$

The stationarity KKT condition (2.13) for (3.13) implies that $\bar{\mathbf{v}}_k$ is the eigenvector corresponding to the largest eigenvalue (this eigenvalue is zero) of $\sum_{i=1}^{K_r} \lambda_{ki} e_{ki} \mathbf{D}_k^H \mathbf{C}_i^H \mathbf{h}_i \mathbf{h}_i^H \mathbf{C}_i \mathbf{D}_k - \sum_{l=1}^{L} \mu_{lk} \mathbf{Q}_{lk}$, where $\mu_{lk}$ is the Lagrange multiplier associated with the $l$th per-user power constraint. Strong duality finally implies $\sum_{l=1}^{L} \mu_{lk} q_{lk} = p_k \sum_{i=1}^{K_r} \lambda_i |\mathbf{h}_i^H \mathbf{C}_i \mathbf{D}_k \bar{\mathbf{v}}_k|^2 = p_k \sum_{l=1}^{L} \mu_{lk} \bar{\mathbf{v}}_k^H \mathbf{Q}_{lk} \bar{\mathbf{v}}_k$, from which (3.12) follows. □

The advantage of the explicit parametrization in Theorem 3.4 is that the operational meaning of the weights at each transmitter is clear. The larger the $\lambda_{ki}$-weight, the more important the corresponding $\text{MS}_i$ is. Either the positive impact on the signal power is increased (if $i = k$) or the negative impact on the interference power is reduced (if $i \neq k$). In addition, the larger the $\mu_{lk}$-weight, the more

[4] The dominating eigenvector in Theorem 3.4 can be computed as $\bar{\mathbf{v}}_k = (\sum_{l=1}^{L} \mu_{lk} \mathbf{Q}_{lk} + \sum_{i=1}^{K_r} \lambda_{ki} \mathbf{D}_k^H \mathbf{C}_i^H \mathbf{h}_i \mathbf{h}_i^H \mathbf{C}_i \mathbf{D}_k)^{\dagger} \mathbf{D}_k \mathbf{h}_k$.

the beamforming direction $\bar{\mathbf{v}}_k$ is shaped by the corresponding power constraint. The disadvantage of the parametrization is that it has in total $K_r(L + K_r - 1)$ parameters to describe the $K_r - 1$ dimensional Pareto boundary. The number of parameters reduces to $K_r(K_r - 1)$ when there is only a total power constraint per user [180], but it still suggests that more efficient parametrizations with less parameters exist.

---

**Remark 3.1 (Extensions).** Another advantage of the parametrization in Theorem 3.4 is that it can be extended to scenarios in which multiple users are interested in the same data; for example, a multicast scenario in which the data stream $i$ from $\mathrm{BS}_j$ is intended for two receivers $\mathrm{MS}_k$ and $\mathrm{MS}_{k+1}$. The parametrization in (3.11) can then be reused with $e_{ki} = e_{k+1\,i} = +1$ and all other $e_{\ell i} = -1$ for $\ell \neq k, k+1$. This scenario is further discussed in Section 4.4.

---

### 3.2.3 Parametrization Based on Uplink–Downlink Duality

A very compact beamforming parametrization can be achieved by exploiting the uplink–downlink duality described in Subsection 2.2.2. In particular, recall from Corollary 2.8 that the optimal beamforming vectors $\mathbf{v}_k^*$ are equal (up to a scaling factor) to

$$\bar{\mathbf{v}}_k^* = \left( \sum_{l=1}^{L} \frac{\mu_l}{q_l} \mathbf{Q}_{lk} + \sum_{i=1}^{K_r} \frac{\lambda_i}{\sigma_i^2} \mathbf{D}_k^H \mathbf{C}_i^H \mathbf{h}_i \mathbf{h}_i^H \mathbf{C}_i \mathbf{D}_k \right)^{\dagger} \mathbf{D}_k^H \mathbf{C}_k^H \mathbf{h}_k, \qquad (3.15)$$

for any resource allocation problem with fixed QoS requirements. Observe that the QoS constraints themselves are not present in (3.15), but only implicitly represented by the optimal Lagrange multipliers $\lambda_k, \mu_l$. Corollary 2.8 therefore provides a necessary condition: *every* feasible point in $\mathcal{R}$ can be achieved using beamforming directions of the type (3.15) for *some* choice of Lagrange multipliers. The parametrization in this subsection utilizes this relation in the opposite direction: for *different* choices of Lagrange multipliers, using beamforming directions of the type (3.15) (along with optimal power allocation) can achieve any point in $\mathcal{R}$. The following theorem originates from [16].

**Theorem 3.5.** Every feasible point $\mathbf{g} \in \mathcal{R}$ is achieved by beamforming vectors $\mathbf{v}_k = \sqrt{p_k}\bar{\mathbf{v}}_k$ for all $k$, where

$$\bar{\mathbf{v}}_k = \frac{\boldsymbol{\Psi}_k^\dagger \mathbf{D}_k^H \mathbf{h}_k}{\|\boldsymbol{\Psi}_k^\dagger \mathbf{D}_k^H \mathbf{h}_k\|}, \tag{3.16}$$

$$\begin{bmatrix} p_1 & \dots & p_{K_r} \end{bmatrix} = \begin{bmatrix} \gamma_1\sigma_1^2 & \dots & \gamma_{K_r}\sigma_{K_r}^2 \end{bmatrix} \mathbf{M}^\dagger, \tag{3.17}$$

$$\boldsymbol{\Psi}_k = \left( \sum_{l=1}^{L} \frac{\mu_l}{q_l} \mathbf{Q}_{lk} + \sum_{i=1}^{K_r} \frac{\lambda_i}{\sigma_i^2} \mathbf{D}_k^H \mathbf{C}_i^H \mathbf{h}_i \mathbf{h}_i^H \mathbf{C}_i \mathbf{D}_k \right), \tag{3.18}$$

$$\gamma_k = \frac{\lambda_k}{\sigma_k^2} \mathbf{h}_k^H \mathbf{D}_k \big( \boldsymbol{\Psi}_k - \frac{\lambda_k}{\sigma_k^2} \mathbf{D}_k^H \mathbf{C}_k^H \mathbf{h}_k \mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_k \big)^\dagger \mathbf{D}_k^H \mathbf{h}_k, \tag{3.19}$$

$$[\mathbf{M}]_{ik} = \begin{cases} |\mathbf{h}_i^H \mathbf{C}_i \mathbf{D}_i \bar{\mathbf{v}}_i|^2, & i = k, \\ -\gamma_k |\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_i \bar{\mathbf{v}}_i|^2, & i \neq k, \end{cases} \tag{3.20}$$

for some nonnegative parameters $\{\lambda_k\}_{k=1}^{K_r}$ and $\{\mu_l\}_{l=1}^{L}$. The Moore–Penrose pseudo-inverse is denoted with $(\cdot)^\dagger$ and $[\mathbf{M}]_{ik}$ is the $ik$th element of $\mathbf{M} \in \mathbb{R}^{K_r \times K_r}$.

*Proof.* The normalized direction $\bar{\mathbf{v}}_k = \boldsymbol{\Psi}_k^\dagger \mathbf{D}_k^H \mathbf{h}_k / \|\boldsymbol{\Psi}_k^\dagger \mathbf{D}_k^H \mathbf{h}_k\|$ of $\mathbf{v}_k$ is given by Corollary 2.8. By exploiting the uplink–downlink duality, the normalized receive combining vectors $\{\bar{\mathbf{v}}_k\}_{k=1}^{K_r}$ achieve the uplink SINRs $\gamma_k$ in (3.19), whose expression is achieved by multiplying (2.39) with $\mathbf{h}_k^H \mathbf{D}_k$ from the left and then dividing by $\mathbf{h}_k^H \mathbf{D}_k \bar{\mathbf{v}}_k$.

To determine $p_k$ for $k = 1, \dots, K_r$, observe that $\mathrm{SINR}_k = \gamma_k$ is also fulfilled in the downlink. Plugging $\mathbf{v}_k = \sqrt{p_k}\bar{\mathbf{v}}_k$ into the expression for $\mathrm{SINR}_k$ gives the system of linear equations

$$p_k |\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_k \bar{\mathbf{v}}_k|^2 = \gamma_k \left( \sigma_k^2 + \sum_{i \neq k} p_i |\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_i \bar{\mathbf{v}}_i|^2 \right) \quad \forall k \tag{3.21}$$

that can be expressed and solved as in (3.17). □

This theorem provides an explicit beamforming parametrization that can achieve any point in the performance region by selecting $K_r + L$ nonnegative parameters. The number of parameters can be reduced if we are only interested in the Pareto boundary.

---

**Corollary 3.6.** Every Pareto optimal point $\mathbf{g} \in \partial^+\mathcal{R}$ is achieved by the parametrization in Theorem 3.5 for some nonnegative parameters $\{\lambda_k\}_{k=1}^{K_r}$ and $\{\mu_l\}_{l=1}^{L}$ satisfying $\sum_{k=1}^{K_r} \lambda_k = 1$ and $\sum_{l=1}^{L} \mu_l = 1$.

The modified beamforming vectors $\tilde{\mathbf{v}}_k = \mathbf{v}_k/\sqrt{\varsigma}\ \forall k$, with $\varsigma = \max_l(\sum_k \frac{\mathbf{v}_k^H \mathbf{Q}_{lk} \mathbf{v}_k}{q_l})$, will always be feasible (i.e., satisfy all constraints).

---

*Proof.* The modified beamforming $\{\tilde{\mathbf{v}}_k\}$ is feasible by construction and only modifies suboptimal and infeasible strategies since Theorem 1.9 shows that at least one power constraint is satisfied with equality at the optimum. Furthermore, we always have $\lambda_k, \mu_l > 0$ for some $k, l$ as all nonzero SINR constraints and at least one power constraint are active. We can thus rescale the Lagrange multipliers to satisfy $\sum_{k=1}^{K_r} \lambda_k + \sum_{l=1}^{L} \mu_l = 2$, which will not remove any solutions since all expressions in Theorem 3.5 are unaffected by a common scaling of all Lagrange multipliers. In addition, the dual function in (2.33) is $\sum_{k=1}^{K_r} \lambda_k - \sum_{l=1}^{L} \mu_l$ and strong duality implies that it is zero at the optimum (as it is dual to a feasibility problem). The combination of these two constraints implies $\sum_{k=1}^{K_r} \lambda_k = 1$ and $\sum_{l=1}^{L} \mu_l = 1$. □

This corollary strengthen the initial parametrization in Theorem 3.5 by showing that only $K_r + L - 2$ parameters between zero and one need to be selected — the remaining two parameters are uniquely determined by the two sum constraints in Corollary 3.6. The number of parameters only scales linearly with the number of users $K_r$ and power constraints $L$, thus it generally includes much fewer parameters than those in the previous two subsections (where the number of parameters generally scales as $K_r L + K_r^2$). On the other hand, the parametrization does not exhibit the same distributed property as the previous ones — all parameters essentially affect the beamforming to all users.

The parametrization only provides a necessary condition for Pareto optimality, but there exist special cases when it is also sufficient. An

example is single-cell transmission with a total power constraint [39], or any multi-cell scenario with only one power constraint.

---

**Corollary 3.7.** Suppose there is only one power constraint (i.e., $L = 1$), then every parameter selection that satisfies the sum constraints in Corollary 3.6 will use full power and achieve a Pareto optimal point.

---

*Proof.* For any given set of parameters, the transmit strategy in Theorem 3.5 solves (2.33) for $g_k^{-1}(r_k^*) = \gamma_k$, which is the dual problem to problem (2.29). If $L = 1$, it is also the dual problem to (with $\mu_1 = 1$)

$$\underset{\mathbf{v}_k \, \forall k}{\text{minimize}} \; \sum_{k=1}^{K_r} \frac{1}{q_1} \mathbf{v}_k^H \mathbf{Q}_{1k} \mathbf{v}_k - \mu_1 \tag{3.22}$$

$$\text{subject to } \frac{1}{\sigma_k^2 \gamma_k} |\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_k \mathbf{v}_k|^2 \geq \left( 1 + \sum_{i \neq k} \frac{1}{\sigma_k^2} |\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_i \mathbf{v}_i|^2 \right) \quad \forall k.$$

Therefore, $\sum_{k=1}^{K_r} \lambda_k - \mu_1 = 0$ implies $\sum_{k=1}^{K_r} \mathbf{v}_k^H \mathbf{Q}_{lk} \mathbf{v}_k = q_l$ due to strong duality. In other words, Theorem 3.5 always produces an operating point $\boldsymbol{\gamma} = [\gamma_1 \ldots \gamma_{K_r}]^T \in \mathcal{R}$ that can only be attained using full power.

Furthermore, suppose for the purpose of contradiction that $\boldsymbol{\gamma} \notin \partial^+ \mathcal{R}$, thus there exists $\mathbf{r} \in \mathcal{R}$ with $\mathbf{r} > \boldsymbol{\gamma}$. Based on the feasible beamforming vectors $\{\tilde{\mathbf{v}}_k\}$ that attain $\mathbf{r}$, we can find $\varsigma < 1$ such that $\{\sqrt{\varsigma} \tilde{\mathbf{v}}_k\}$ also yields strictly better performance than $\boldsymbol{\gamma}$. This transmit strategy would achieve a strictly smaller value in (3.22) than the transmit strategy in Theorem 3.5, which is a contradiction. Consequently, every parameter selection gives a point on the weak Pareto boundary. □

As the parameters equal the Lagrange multipliers for resource allocation with fixed QoS requirements, they have the same interpretation: $\lambda_k$ is the (normalized) transmit power from $\text{MS}_k$ in the dual uplink and $\mu_l$ is the (normalized) uplink noise variance. Intuitively, this means that a relative increase in $\lambda_k$ should improve for $\text{MS}_k$ and degrade for the other users, while a relative increase in $\mu_l$ should degrade for all users. As the parameter $\mu_l$ is associated with the $l$th power constraint, the KKT conditions imply that it should be zero for all inactive constraints. The intuition is confirmed by the following corollary.

**Corollary 3.8.** The parameters in Theorem 3.5 have the following impact on the performance of $\mathrm{MS}_k$:

$$\begin{aligned} &\frac{\partial}{\partial \lambda_i} g_k(\mathrm{SINR}_k) \begin{cases} \geq 0, & k = i, \\ \leq 0, & k \neq i, \end{cases} \\ &\frac{\partial}{\partial \mu_l} g_k(\mathrm{SINR}_k) \leq 0 \quad \forall l. \end{aligned} \tag{3.23}$$

*Proof.* The corollary follows from differentiating the expression for $\mathrm{SINR}_k$ in (3.19), in conjunction with the monotonicity of $g_k(\cdot)$. □

To summarize, Theorem 3.5 provides a very compact beamforming parametrization: only $K_r + L - 2$ parameters are required to explicitly characterize beamforming vectors that can attain any point on the Pareto boundary. The characterization does generally not provide a sufficient condition for Pareto optimality, but numerical evaluation has shown very good performance under heuristic parameter selection [18] — there is a strong connection to the heuristic SLNR maximizing beamforming described in Subsection 3.4.3. The parametrization also shows that the optimal beamforming directions are achieved by taking the channel direction $\mathbf{D}_k^H \mathbf{h}_k$ and rotating it using a matrix $\mathbf{\Psi}_k$ whose terms determine to which extent power constraints and inter-user interference are taken into account.

## 3.3 Necessary and Sufficient Pareto Boundary Parametrization

Recall that the Pareto boundary represents all optimal solutions to (3.1). The beamforming parameterizations in the previous section provides necessary conditions for Pareto optimality, but not sufficient conditions (except for $L = 1$, see Corollary 3.7). This means that each Pareto optimal point is achieved by some set of parameters, but each set of parameters will not give a Pareto optimal point. A necessary *and* sufficient characterization of the Pareto boundary can however be achieved as follows, based on the line of work in [16, 126, 185, 325].

**Theorem 3.9.** Each point on the weak Pareto boundary of $\mathcal{R}$ is achieved by solving a weighted max-min fairness problem (e.g., an FPO problem with $a_k = 0\ \forall k$, see Example 2.8) for some *unique* weighting vector $\mathbf{w} = [w_1 \dots w_{K_r}]^T \in \mathbb{R}_+^{K_r}$ with $\sum_{k=1}^{K_r} w_k = 1$. Furthermore, every such weighting vector gives a point on the weak Pareto boundary.

*Proof.* The first part follows from Lemma 1.10, where fairness-profile optimization was used to constructively prove that every weak Pareto optimal point is achieved by some system utility function. The uniqueness follows from that $\mathcal{R}$ is normal. The second part is trivial since the weighted max-min fairness problem is a scalarization of the multi-objective optimization problem in (3.1). □

This characterization has $K_r$ parameters between zero and one, but we only need to select $K_r - 1$ parameters due to the unit sum constraint. Therefore, Theorem 3.9 is more powerful than the beamforming parameterizations in Section 3.2, both in terms of guaranteeing a Pareto optimal point for any parameter selection and by having fewer parameters. The drawback is that a fairness-profile optimization problem needs to be solved for every parameter selection, which is a quasi-convex problem and has polynomial computational complexity to reach an accuracy of $\delta > 0$ (see Subsection 2.2.3). On the contrary, the beamforming parametrizations in Section 3.2 are based on closed-form expressions that enables immediate calculation of the beamforming vectors and user performance achieved by any parameter selection.

In other words, there are two main approaches to generate/approximate the Pareto boundary of $\mathcal{R}$ (e.g., for visualization):

(1) Calculate sample points on the Pareto boundary by applying Theorem 3.9 on a fine grid of weighting vectors $\mathbf{w}$;
(2) Calculate sample points of the whole performance region using any of the beamforming parametrizations in Section 3.2 over a fine grid of parameters. Then generate the Pareto boundary by taking the convex hull of these points.

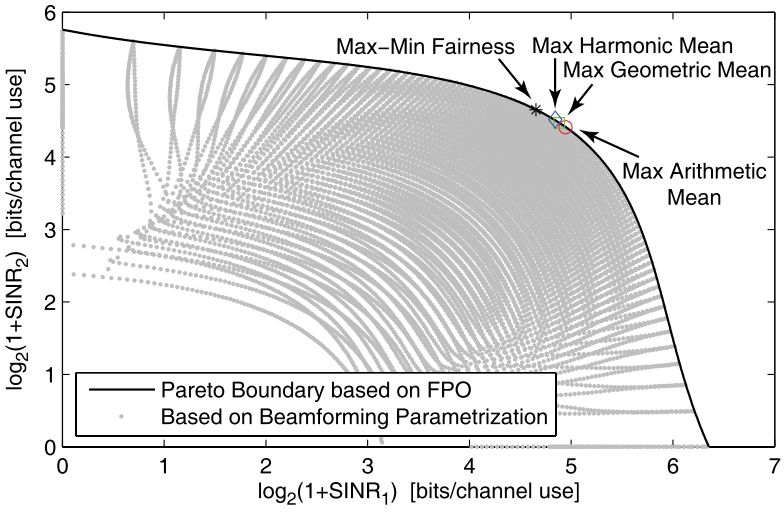
We can typically afford many more sample points with the second approach than with the first approach, because the point generation is based on closed-form expressions. In fact, the accuracy of the second approach is essentially limited by the storage capability and the computational complexity of the convex hull operation, and not by the computation of feasible points. The beamforming parametrization in Subsection 3.2.3 is recommended when using the second approach, because it has the smallest number of parameters and thus less redundancy in the search-space.

To complete the picture, note that there are special cases when the Pareto boundary can be characterized in a necessary and sufficient manner without the need for numerically solving an optimization problem for each point. Example 3.1 showed that this is possible for the two-user interference channel with per-transmitter power constraints [149, 181] and Corollary 3.7 showed it for $L = 1$.

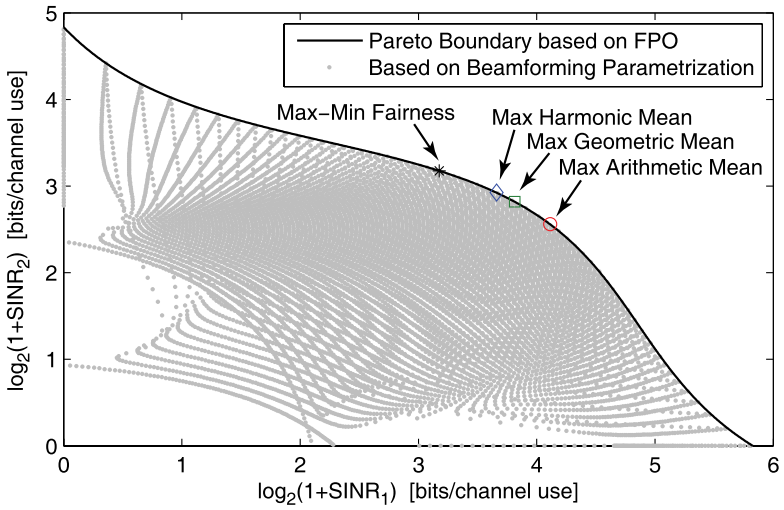
### 3.3.1 Numerical Illustrations

Next, we visualize the two approaches for generating the Pareto boundary of $\mathcal{R}$. We consider a simple scenario with $K_t$ base stations and per-transmitter power constraints with $q_l = 10$ (i.e., 10 dBm). Each base station has one user in its vicinity, but all $K_r = K_t$ users are served by global joint transmission (see Example 1.3). The channels are generated as uncorrelated Rayleigh fading and the average single-user SNR $\frac{\mathbb{E}\{q_l \|\mathbf{h}_{jk}\|_2^2\}}{\sigma_k^2}$ is $q_l N_j$ for the user close to $\mathrm{BS}_j$ and $q_l \frac{N_j}{3}$ for other users. The information rate $g_k(\mathrm{SINR}_k) = \log_2(1 + \mathrm{SINR}_k)$ is considered.

Figure 3.1 considers the case with $K_t = K_r = 2$ and shows two independent channel realizations. The solid curve shows the (approximate) Pareto boundary generated by FPO using Theorem 3.9 with 1001 equally spaced weighting vectors. The shaded area shows the sample points generated by the beamforming parametrization in Subsection 3.2.3 when the $K_r + L - 2 = 2$ parameters were varied in steps of 0.01. The optimal points with different system utility functions are also indicated. These points are close together for the convex region in Figure 3.1(a) and spread out for the nonconvex region in Figure 3.1(b). Observe that the sample points generated by the




(a) Channel Realization 1

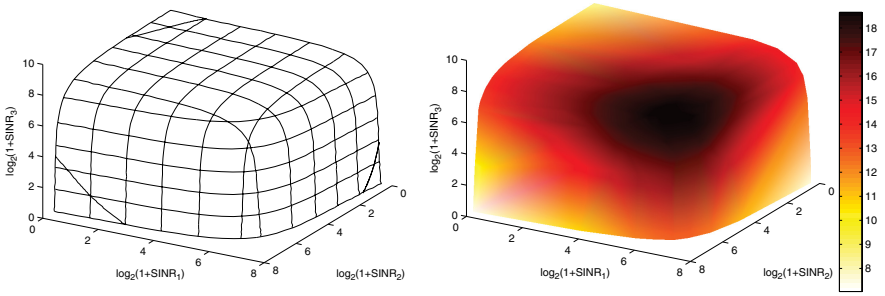



(b) Channel Realization 2

Fig. 3.1 Performance regions for two different channel realizations under global joint transmission with two antennas per base station. The Pareto boundaries are approximated in two ways: (1) Solving FPO problems over a grid of weighting vectors; and (2) Generating sample points using a beamforming parametrization.

beamforming parametrization are concentrated in the area where these system utility functions are located, but the convex hull of the points will closely approximate the shape of the whole region. We will return to this example in Section 4 to visualize the effect of different system model generalizations.

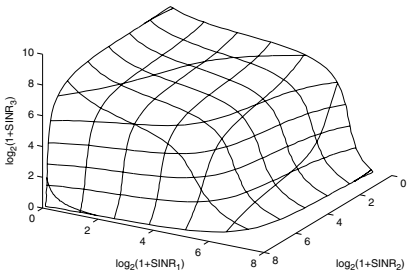
Figure 3.2 considers the case with $K_t = K_r = 3$ and shows two independent channel realizations. (a) and (c) show the (approximate) Pareto boundary generated by FPO using Theorem 3.9 and a fine grid of weighting vectors. (b) and (d) show the convex hull of the sample points generated by the beamforming parametrization in Subsection 3.2.3 when the $K_r + L - 2 = 4$ parameters were varied in steps of 0.03. On the one hand, the grid achieved by the first approach is somewhat easier to interpret and also provides a more accurate visualization



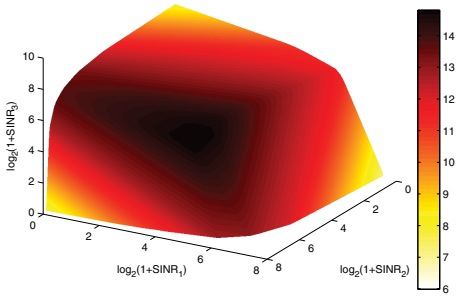

(a) Channel Realization 1, Approach 1




(b) Channel Realization 1, Approach 2




(c) Channel Realization 2, Approach 1




(d) Channel Realization 2, Approach 2

Fig. 3.2 Performance regions for two different channel realizations under global joint transmission with three antennas per base station. The Pareto boundaries are approximated in two ways: (1) Solving FPO problems over a grid of weighting vectors; and (2) Generating sample points using a beamforming parametrization. The color bar shows the sum utility.

when the performance region is nonconvex, which is the case for the second channel realization. On the other hand, the second approach requires much less computational efforts (and roughly half the running time was spent on generating the convex hull).

## 3.4 Heuristic Coordinated Beamforming

This section will discuss heuristic approaches for solving (3.1), for the purpose of achieving practical low-complexity algorithms. To bring some perspective, we begin with describing a related problem:

A classic scenario in signal processing is the detection of a scalar data symbol $s_k$ which is observed under channel distortion, additive interference, and white noise [281]. If multiple channel observations are available for a certain data symbol (e.g., from multiple base station antennas in the multi-cell uplink), this scenario can be modeled as

$$\mathbf{y} = \sum_{k=1}^{K_r} \mathbf{h}_k s_k + \mathbf{n}, \tag{3.24}$$

where $\mathbf{h}_k$ is the channel for symbol $s_k$, $\mathbb{E}\{s_k\} = 0$, $\mathbb{E}\{|s_k|^2\} = 1$, and $\mathbb{E}\{\mathbf{n}\mathbf{n}^H\} = \sigma^2\mathbf{I}$. The symbol $s_k$ can be estimated from the vector-valued observation $\mathbf{y}$ as $\hat{s}_k = \bar{\mathbf{v}}^H\mathbf{y}$ using a linear receive combining filter $\bar{\mathbf{v}}$. Three classic (coherent) receive combining techniques are:

(1) *Maximum ratio combining* (or matched filtering): Weighs and aligns the observations as $\bar{\mathbf{v}} = \frac{1}{\|\mathbf{h}_k\|_2^2+\sigma^2}\mathbf{h}_k$ to maximize the ratio between received signal power and noise power.
(2) *Zero-forcing filtering*: Removes interference by projecting the observations as $\bar{\mathbf{v}} = \left(\sum_{i=1}^{K_r} \mathbf{h}_i\mathbf{h}_i^H\right)^{\dagger}\mathbf{h}_k$, which is the orthogonal complement of the interfering signals. This maximizes the ratio between received signal power and interference power.
(3) *Wiener filtering* (or linear MMSE filtering): The MSE-minimizing $\bar{\mathbf{v}} = (\sum_{i=1}^{K_r} \mathbf{h}_i\mathbf{h}_i^H + \sigma^2\mathbf{I})^{-1}\mathbf{h}_k$ that balances between maximizing signal power and suppressing interference.

For fixed and known channel and noise characteristics, the properties of these combining techniques are relatively easy to analyze. This holds even in large and complex systems, because the filtering

is an internal signal processing procedure at the receiver and thus independent of the processing at other receivers. The Wiener filter is derived as the one maximizing the SINR (and minimizing the MSE) in the filtered signal. This is equivalent to maximum ratio combining in noise-limited scenarios (i.e., when the noise is very strong compared with the interference) and equivalent to zero-forcing in interference-limited scenarios (i.e., when the interference is very strong).

This section explores the relationship between the aforementioned linear receive combining techniques and the linear beamforming vectors in the downlink resource allocation problem (3.1). Beamforming is basically a linear *transmit* filtering, but it is more difficult to optimize than receive filtering. The main reason is that the beamforming affects the channel characteristics (directivity and gain) of the intended and interfering signals, while these are fixed under receive combining. Nevertheless, there are important connections established by the uplink–downlink duality in [30, 226, 282, 283, 315]; see Subsection 2.2.2. The counterparts to the three classic receive combining techniques are *maximum ratio transmission (MRT)*, *zero-forcing beamforming (ZFBF)*, and *signal-to-leakage-and-noise ratio maximizing (SLNR-MAX) beamforming* (also known as transmit Wiener filter [115]). These are described in the remainder of this section and the beamforming parametrization in Subsection 3.2.3 is used to prove under which conditions these heuristic strategies are actually optimal.

For clarity, we mainly consider a multi-cell scenario where the power constraints and cooperation clusters enable derivation of closed-form expressions: coordinated beamforming with per-transmitter power constraints (see Example 1.2 and (1.10)). The multi-objective resource allocation problem in (3.1) then becomes

$$\begin{aligned} \underset{\{\bar{\mathbf{v}}_{j_k k}\}_{k=1}^{K_r}, \{p_{j_k k}\}_{k=1}^{K_r}}{\text{maximize}} \quad & \{g_1(\mathrm{SINR}_1), \ldots, g_{K_r}(\mathrm{SINR}_{K_r})\} \\ \text{subject to} \quad & p_{j_k k} \geq 0, \quad \|\bar{\mathbf{v}}_{j_k k}\|_2 = 1 \quad \forall k, \\ & \mathrm{SINR}_k = \frac{|\sqrt{p_{j_k k}} \mathbf{h}_{j_k k}^H \mathbf{C}_{j_k k} \bar{\mathbf{v}}_{j_k k}|^2}{\sigma_k^2 + \sum_{i \neq k} |\sqrt{p_{j_i i}} \mathbf{h}_{j_i k}^H \mathbf{C}_{j_i k} \bar{\mathbf{v}}_{j_i i}|^2} \quad \forall k, \\ & \sum_{k \in \mathcal{D}_j} p_{j_k k} \leq q_j \quad \forall j, \end{aligned} \tag{3.25}$$

where $j_k$ is the index of the base station that serves $\mathrm{MS}_k$ and the beamforming vectors are decomposed as $\mathbf{v}_k = [\mathbf{0} \dots \mathbf{0} \ \sqrt{p_{j_k k}} \bar{\mathbf{v}}_{j_k k}^T \ \mathbf{0} \dots \mathbf{0}]^T$. Here, $\bar{\mathbf{v}}_{j_k k} \in \mathbb{C}^{N_{j_k} \times 1}$ is the unit-norm beamforming direction and $p_{j_k k} \geq 0$ is the power allocated by $\mathrm{BS}_{j_k}$ for transmission to $\mathrm{MS}_k$. Theorem 3.5 gives the following parametrization of the beamforming directions.

---

**Corollary 3.10.** In coordinated beamforming with per-transmitter constraints, each feasible point $\mathbf{g} \in \mathcal{R}$ is achieved by beamforming directions $\bar{\mathbf{v}}_{j_k k} = \boldsymbol{\Psi}_{j_k k}^{-1} \mathbf{h}_{j_k k} / \| \boldsymbol{\Psi}_{j_k k}^{-1} \mathbf{h}_{j_k k} \|_2$, where

$$\boldsymbol{\Psi}_{j_k k} = \left( \frac{\mu_{j_k}}{q_{j_k}} \mathbf{I}_{N_{j_k}} + \sum_{i=1}^{K_r} \frac{\lambda_i}{\sigma_i^2} \mathbf{C}_{j_k i}^H \mathbf{h}_{j_k i} \mathbf{h}_{j_k i}^H \mathbf{C}_{j_k i} \right) \quad \forall k, \tag{3.26}$$

for some parameters $\{\lambda_k\}_{k=1}^{K_r}$ and $\{\mu_j\}_{j=1}^{K_t}$ between zero and one.

---

This corollary will be useful when analyzing the optimality of heuristic beamforming strategies.

### 3.4.1 Maximum Ratio Transmission (MRT)

The beamforming concept of maximum ratio transmission was introduced in [159] to maximize the SNR $\frac{p_{j_k k}}{\sigma_k^2} |\mathbf{h}_{j_k k}^H \bar{\mathbf{v}}_{j_k k}|^2$ at $\mathrm{MS}_k$ in multi-antenna transmission. Variations on this concept have appeared even earlier; see [115] for an overview.

---

**Definition 3.2 (Maximum Ratio Transmission).** The beamforming directions

$$\bar{\mathbf{v}}_{j_k k}^{(\mathrm{MRT})} = \frac{\mathbf{h}_{j_k k}}{\| \mathbf{h}_{j_k k} \|_2} \quad \forall k \tag{3.27}$$

are called *maximum ratio transmission (MRT).*

---

MRT is the counterpart of maximum ratio combining in receive processing; in fact, the latter name is sometimes used to describe both techniques, although it may cause confusion. MRT can be viewed as a matched filter where the gain of each entry in $\bar{\mathbf{v}}_{j_k k}^{(\mathrm{MRT})}$ equals the relative strength of the corresponding channel coefficient in $\mathbf{h}_{j_k k}$ and the phase

makes the signal contribution from each channel coefficient add up constructively. The inner product $|\mathbf{h}_{j_k k}^H \bar{\mathbf{v}}_{j_k k}^{(\mathrm{MRT})}|$ is therefore maximized, which protects the useful signal against channel fading. The direction of the MRT vector is illustrated in Figure 3.3, while Figure 3.4 shows

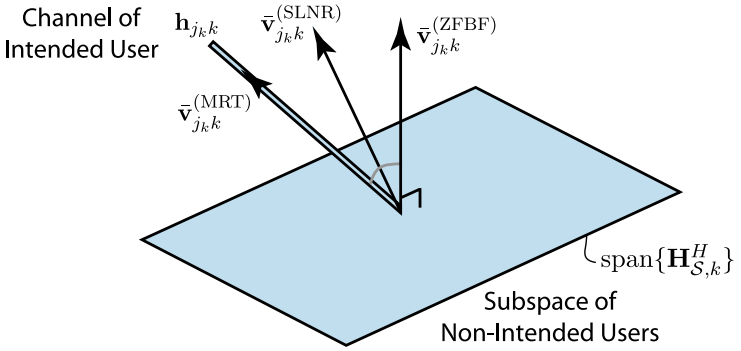



Fig. 3.3 Illustration of the beamforming directions with maximum ratio transmission (MRT), zero-forcing beamforming (ZFBF), and signal-to-leakage-and-noise ratio maximizing (SLNR-MAX) beamforming. MRT follows the channel of the intended user, ZFBF is orthogonal to the channel of nonintended users, and SLNR-MAX balances between these extremes (and moves between them depending on the SNR).

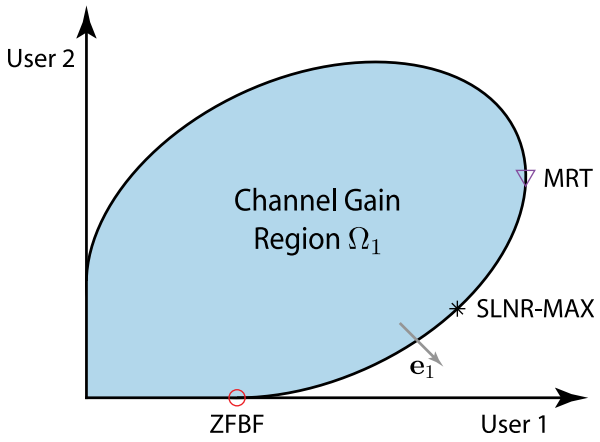



Fig. 3.4 Illustration of the channel gain region $\Omega_1$ of the signal intended for User 1. The boundary points represent different beamforming directions. In particular, the upper boundary in direction $\mathbf{e}_1 = [+1 \ -1]^T$ contains maximum ratio transmission (MRT), zero-forcing beamforming (ZFBF), and signal-to-leakage-and-noise ratio maximizing (SLNR-MAX) beamforming. MRT maximizes the channel gain of User 1, ZFBF causes zero interference to User 2, and SLNR-MAX balances between these extremes (and moves between them depending on the SNR).

that MRT equals the boundary point in the channel gain region that maximizes the gain of the intended user.

Analytic expressions for the average performance with MRT can be derived, for example, for point-to-point systems [63, 178, 317]. By observing that $\bar{\mathbf{v}}_{j_k k}^{(\mathrm{MRT})}$ is the dominating eigenvector of $\mathbf{h}_{j_k k}\mathbf{h}_{j_k k}^H$, MRT can easily be extended to scenarios where this outer product is not known perfectly at the transmitter; the average SNR can be maximized by using the dominating eigenvector of $\mathbb{E}\{\mathbf{h}_{j_k k}\mathbf{h}_{j_k k}^H\}$ instead [214].

We have the following result regarding the optimality of MRT.

---

**Corollary 3.11.** In coordinated beamforming with per-transmitter constraints, each feasible point $\mathbf{g} \in \mathcal{R}$ is asymptotically achieved by $\bar{\mathbf{v}}_{j_k k}^{(\mathrm{MRT})}$ as $\frac{q_{j_k}}{\sigma_i^2} \to 0 \, \forall k, i$ (for some feasible power allocation $\{p_{j_k k}\}_{k=1}^{K_r}$).

---

*Proof.* The beamforming direction in Corollary 3.10 can be equally expressed as $\bar{\mathbf{v}}_{j_k k} = \widetilde{\mathbf{\Psi}}_{j_k k}^{-1}\mathbf{h}_{j_k k}/\|\widetilde{\mathbf{\Psi}}_{j_k k}^{-1}\mathbf{h}_{j_k k}\|_2$, where $\widetilde{\mathbf{\Psi}}_{j_k k} = (\mu_{j_k}\mathbf{I}_{N_{j_k}} + \sum_{i=1}^{K_r} \frac{q_{j_k}\lambda_i}{\sigma_i^2}\mathbf{C}_{j_k i}^H\mathbf{h}_{j_k i}\mathbf{h}_{j_k i}^H\mathbf{C}_{j_k i}) \to \mu_{j_k}\mathbf{I}_{N_{j_k}}$ as $\frac{q_{j_k}}{\sigma_i^2} \to 0$ (since $0 \leq \lambda_i \leq 1 \, \forall k$). This is a scaled identity matrix and will not affect the beamforming direction, thus $\bar{\mathbf{v}}_{j_k k} \to \frac{\mathbf{h}_{j_k k}}{\|\mathbf{h}_{j_k k}\|_2} = \bar{\mathbf{v}}_{j_k k}^{(\mathrm{MRT})}$. □

The corollary shows that MRT provides the optimal beamforming directions for (3.25) in the low-SNR regime, irrespectively of which point in the performance region that we are interested in (or which single-objective problem that we want to solve). The exact operating point is determined by the power allocation, which is further discussed in Subsection 3.4.4. Furthermore, MRT achieves the corner points of the Pareto boundary, $[0 \ldots 0 \, u_k \, 0 \ldots 0]^T$, where the system only transmits to $\mathrm{MS}_k$ and uses full power $p_{j_k k} = q_{j_k}$ (this holds in any SNR regime).

### 3.4.2 Zero-Forcing Beamforming (ZFBF)

Zero-forcing refers to signal processing that completely eliminates interference. This can be achieved at the transmitter-side by selecting beamforming vectors that are orthogonal to the channels of nonintended users. This idea has been used in wireless communications for at least two decades, under alternative names such as pre-decorrelation,

pre-equalization, channel inversion, and interference nulling. Some early works are [151, 186, 285] and many more references are available in [115]. A theoretical motivation is that zero-forcing simultaneously minimizes the MSE between the received signal and the transmitted symbol $s_k$,

$$\mathrm{MSE}_k = \mathbb{E}\left\{ \left| \underbrace{\sum_{i=1}^{K_r} \sqrt{p_{j_i i}} \mathbf{h}_{j_i k}^H \mathbf{C}_{j_i k} \bar{\mathbf{v}}_{j_i i} s_i + n_k}_{=y_k \text{ (received signal at MS}_k\text{)}} - s_k \right|^2 \right\} \geq \mathbb{E}\{|n_k|^2\}, \quad (3.28)$$

in the ideal case without any transmit power constraints. Zero-forcing is only applied for the *active* users, which are defined as follows.

---

**Definition 3.3 (Active Users).** The set $\mathcal{S} \subseteq \{1, \ldots, K_r\}$ is called a *scheduling set* if $p_{j_k k} = 0$ for all users $k \notin \mathcal{S}$. Users with indices in $\mathcal{S}$ are *active*, while all other users are *inactive*.

---

The definition of zero-forcing beamforming is based on [23, 203].

---

**Definition 3.4 (Zero-Forcing Beamforming).** The beamforming directions

$$\bar{\mathbf{v}}_{j_k k}^{(\mathrm{ZFBF})} = \frac{\begin{bmatrix} \mathbf{h}_{j_k k} & \mathbf{H}_{\mathcal{S},k}^H \end{bmatrix} \left( \begin{bmatrix} \mathbf{h}_{j_k k}^H \\ \mathbf{H}_{\mathcal{S},k} \end{bmatrix} \begin{bmatrix} \mathbf{h}_{j_k k} & \mathbf{H}_{\mathcal{S},k}^H \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 \\ 0 \\ \vdots \end{bmatrix}}{\left\| \begin{bmatrix} \mathbf{h}_{j_k k} & \mathbf{H}_{\mathcal{S},k}^H \end{bmatrix} \left( \begin{bmatrix} \mathbf{h}_{j_k k}^H \\ \mathbf{H}_{\mathcal{S},k} \end{bmatrix} \begin{bmatrix} \mathbf{h}_{j_k k} & \mathbf{H}_{\mathcal{S},k}^H \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 \\ 0 \\ \vdots \end{bmatrix} \right\|_2} \quad \forall k \in \mathcal{S} \quad (3.29)$$

are called *zero-forcing beamforming (ZFBF)* toward the users in the scheduling set $\mathcal{S}$. The matrix $\mathbf{H}_{\mathcal{S},k} \in \mathbb{C}^{(|\mathcal{S} \cap \mathcal{C}_{j_k}|-1) \times N}$ of user $k \in \mathcal{S}$ contains the channels $\mathbf{h}_{j_k^i}^H \mathbf{C}_{j_k i}$ for co-users $i \in \mathcal{S} \cap \mathcal{C}_{j_k} \setminus \{k\}$ that $\mathrm{BS}_{j_k}$ coordinates interference toward. ZFBF can be equivalently defined as

$$\bar{\mathbf{v}}_{j_k k}^{(\mathrm{ZFBF})} = \frac{\boldsymbol{\Pi}_{\mathbf{H}_{\mathcal{S},k}^H}^{\perp} \mathbf{h}_{j_k k}}{\| \boldsymbol{\Pi}_{\mathbf{H}_{\mathcal{S},k}^H}^{\perp} \mathbf{h}_{j_k k} \|_2} = \frac{\left( \mathbf{I}_{N_{j_k}} - \mathbf{H}_{\mathcal{S},k}^H (\mathbf{H}_{\mathcal{S},k} \mathbf{H}_{\mathcal{S},k}^H)^{-1} \mathbf{H}_{\mathcal{S},k} \right) \mathbf{h}_{j_k k}}{\left\| \left( \mathbf{I}_{N_{j_k}} - \mathbf{H}_{\mathcal{S},k}^H (\mathbf{H}_{\mathcal{S},k} \mathbf{H}_{\mathcal{S},k}^H)^{-1} \mathbf{H}_{\mathcal{S},k} \right) \mathbf{h}_{j_k k} \right\|_2}. \quad (3.30)$$

---

ZFBF is the counterpart of zero-forcing filtering in receive processing. To cancel all inter-user interference, the beamforming directions $\bar{\mathbf{v}}_{j_k k}^{(\mathrm{ZFBF})}$ are achieved by projecting the channel vector $\mathbf{h}_{j_k k}$ of the intended user onto the orthogonal complement $\mathbf{\Pi}_{\mathbf{H}_{\mathcal{S},k}^H}^{\perp}$ of the subspace spanned by rows of $\mathbf{H}_{\mathcal{S},k}$ (the channels of the nonintended users); see Figure 3.3. The orthogonal complement is only nonempty if the nonintended users are fewer than $N_{j_k}$, which is the dimension of the beamforming vector. This explains the need to consider the scheduling set $\mathcal{S}$ in Definition 3.4; interference need only to be canceled for active users. The existence of ZFBF can be guaranteed as follows.

---

**Lemma 3.12.** ZFBF exists if the channel vectors $\mathbf{h}_{j_k k}$ of the users $k \in \mathcal{S} \cap \mathcal{C}_j$ are linearly independent for all base stations $j$. This is typically satisfied whenever $|\mathcal{S} \cap \mathcal{C}_j| \leq N_j \; \forall j$.

---

*Proof.* This follows directly from the fact that linear independence means that no channel vector lies in the span of the other channel vectors. □

As noted in the lemma, this condition is satisfied whenever the base station is not coordinating interference to more (active) users than its number of transmit antennas. In this case, one can argue that channel vectors generated independently from (perhaps unknown) stochastic distributions with high-rank covariance matrices are nonzero and linearly independent with probability one.

ZFBF has a practically appealing structure as the interference cancelation transforms the SINR of each user into an SNR; this is illustrated in Figure 3.4 where ZFBF equals the boundary point in the channel gain region that maximizes the channel gain while causing zero interference to nonintended users. Recall from Subsection 2.2.1 that even very difficult multi-cell resource allocation problems become solvable under zero-forcing assumptions — although closed-form expressions as the one in Definition 3.4 are difficult to obtain under arbitrary power constraints [297]. If zero-forcing constraints are part of the original problem formulation, then zero-forcing transmission is certainly

optimal; however, the optimality of ZFBF can be shown under other conditions.

---

**Corollary 3.13.** In coordinated beamforming with per-transmitter constraints, consider any feasible point $\mathbf{g} \in \mathcal{R}$ where only a subset $\mathcal{S}$ of the users is active (i.e., $g_k > 0$ for $k \in \mathcal{S}$ and $g_k = 0$ for $k \notin \mathcal{S}$). If $\mathcal{S}$ satisfies Lemma 3.12, then $\mathbf{g}$ is asymptotically achieved by $\bar{\mathbf{v}}_{j_k k}^{(\mathrm{ZFBF})}$ as $q_j \to \infty \, \forall j$ (for some feasible power allocation $\{p_{j_k k}\}_{k=1}^{K_r}$).

---

*Proof.* The beamforming direction in Corollary 3.10 can be equally expressed as $\bar{\mathbf{v}}_{j_k k} = \widetilde{\mathbf{\Psi}}_{j_k k}^{-1} \mathbf{h}_{j_k k} / \| \widetilde{\mathbf{\Psi}}_{j_k k}^{-1} \mathbf{h}_{j_k k} \|_2$ for $\widetilde{\mathbf{\Psi}}_{j_k k} = (\mathbf{I}_{N_{j_k}} + q_{j_k} \mathbf{U} \mathbf{\Lambda} \mathbf{U}^H)$, where $\mathbf{U} \mathbf{\Lambda} \mathbf{U}^H = \sum_{i=1}^{K_r} \frac{\lambda_i}{\mu_{j_k} \sigma_i^2} \mathbf{C}_{j_k i}^H \mathbf{h}_{j_k i} \mathbf{h}_{j_k i}^H \mathbf{C}_{j_k i}$ denotes the eigen decomposition. Let $\tilde{\lambda}_m$ be the $m$th largest eigenvalue and $\mathbf{u}_m$ be the corresponding eigenvector, then observe that

$$\begin{aligned}
\bar{\mathbf{v}}_{j_k k} &= \frac{(\mathbf{I}_{N_{j_k}} + q_{j_k} \mathbf{U} \mathbf{\Lambda} \mathbf{U}^H)^{-1} \mathbf{h}_{j_k k}}{\| (\mathbf{I}_{N_{j_k}} + q_{j_k} \mathbf{U} \mathbf{\Lambda} \mathbf{U}^H)^{-1} \mathbf{h}_{j_k k} \|_2} \\
&= \frac{\left( \left( \mathbf{I}_{N_{j_k}} - \sum_{m=1}^{\mathrm{rank}(\mathbf{\Lambda})} \mathbf{u}_m \mathbf{u}_m^H \right) + \sum_{m=1}^{\mathrm{rank}(\mathbf{\Lambda})} \frac{1}{\tilde{\lambda}_m q_{j_k} + 1} \mathbf{u}_m \mathbf{u}_m^H \right) \mathbf{h}_{j_k k}}{\left\| \left( \left( \mathbf{I}_{N_{j_k}} - \sum_{m=1}^{\mathrm{rank}(\mathbf{\Lambda})} \mathbf{u}_m \mathbf{u}_m^H \right) + \sum_{m=1}^{\mathrm{rank}(\mathbf{\Lambda})} \frac{1}{\tilde{\lambda}_m q_{j_k} + 1} \mathbf{u}_m \mathbf{u}_m^H \right) \mathbf{h}_{j_k k} \right\|_2} \\
&\to \frac{\left( \mathbf{I}_{N_{j_k}} - \sum_{m=1}^{\mathrm{rank}(\mathbf{\Lambda})} \mathbf{u}_m \mathbf{u}_m^H \right) \mathbf{h}_{j_k k}}{\left\| \left( \mathbf{I}_{N_{j_k}} - \sum_{m=1}^{\mathrm{rank}(\mathbf{\Lambda})} \mathbf{u}_m \mathbf{u}_m^H \right) \mathbf{h}_{j_k k} \right\|_2} = \bar{\mathbf{v}}_{j_k k}^{(\mathrm{ZFBF})}
\end{aligned} \tag{3.31}$$

as $q_j \to \infty \, \forall j$ while the noise power is fixed.[5] This expression is well-defined as $\mathrm{rank}(\mathbf{\Lambda}) < N_{j_k}$ whenever $\mathcal{S}$ satisfies Lemma 3.12. □

[5] This is not necessarily the case when the noise term includes uncoordinated interference that is amplified when the transmit power is increased. We can still expect ZFBF to have an approximately optimal structure in practice, but one should be careful when simulating the performance in the high-SNR regime as large multi-cell systems are fundamentally interference-limited; see [164].

This means that ZFBF provides the optimal beamforming directions for (3.25) in the high-SNR regime, if we limit ourselves to those parts of the performance region where ZFBF actually exists. The asymptotic optimality is expected since ZFBF (with proper power allocation) minimizes the MSE in (3.28) when the power constraints are ignored. Moreover, the loss in signal power due to interference cancelation typically diminishes as the number of transmit antennas is increased (due to the increased spatial beamforming resolution [220]).

As it is desirable to deploy multi-cell systems that operate in the high-SNR regime (for spectral efficiency reasons), many researchers have proposed ZFBF-based transmit strategies for practical use; see [75, 97, 319] among others. Although perfect interference nulling requires perfect CSI, ZFBF can be implemented under imperfect CSI by avoiding inter-user interference along some of the strongest eigenvectors of $\mathbb{E}\{\mathbf{h}_{j_k k}\mathbf{h}_{j_k k}^H\}$ of each user $k$ [23, 99]. If we consider estimated and quantized CSI, this naive but simple approach is robust in the sense of only giving a limited average performance loss[6] in the high-SNR regime [15, 44, 113].

### 3.4.3 Signal-to-Leakage-and-Noise Ratio Maximizing (SLNR-MAX) Beamforming

The heuristic MRT and ZFBF in the previous two subsections follow from straightforward extensions of the corresponding criteria for receive combining: maximize SNR and minimize interference power, respectively. These criteria decouple the selection of beamforming directions by either ignoring or canceling interference. Wiener filtering balances between signal power maximization and interference power minimization, making it is less obvious how to define a transmission counterpart; the transmit beamforming direction $\bar{\mathbf{v}}_{j_k k}$ for $\mathrm{MS}_k$ affects the SINRs of all co-users $i \in \mathcal{C}_{j_k} \setminus \{k\}$. A heuristic way to balance between signal and interference is to maximize the *signal-to-leakage-and-noise ratio*

[6] The CSI quality needs to increase with the SNR to bound the performance loss, but this happens naturally when the uplink SNR increases linearly with the downlink SNR [44].

*(SLNR)*, which we define as

$$\mathrm{SLNR}_k = \frac{\frac{1}{\sigma_k^2}|\mathbf{h}_{j_k k}^H \mathbf{C}_{j_k k} \bar{\mathbf{v}}_{j_k k}|^2}{\frac{1}{\eta_{j_k}} + \sum_{i \neq k} \frac{1}{\sigma_i^2}|\mathbf{h}_{j_k i}^H \mathbf{C}_{j_k i} \bar{\mathbf{v}}_{j_k k}|^2} \quad \forall k \tag{3.32}$$

for some parameters $\eta_{j_k} \geq 0$. This expression is slightly different from the original definition in [221, 259], where the noise powers are handled inconsistently.[7] If the parameters $\{\eta_j\}_{j=1}^{K_t}$ represent equal power allocation from each base station, $\eta_{j_k} = \frac{q_{j_k}}{|\mathcal{D}_{j_k}|}$, then $\mathrm{SLNR}_k$ is the ratio between the signal power at the intended user and the (normalized) noise plus the total interference power that leaks to nonintended users. Other values on $\eta_{j_k}$ are also possible; see Remark 3.2. If $\{\eta_j\}_{j=1}^{K_t}$ are fixed, the following heuristic beamforming directions maximize (3.32).

---

**Definition 3.5 (SLNR Maximizing Beamforming).** The beamforming directions

$$\bar{\mathbf{v}}_{j_k k}^{(\mathrm{SLNR})} = \frac{\left(\frac{1}{\eta_{j_k}}\mathbf{I}_{N_{j_k}} + \sum_{i=1}^{K_r} \frac{1}{\sigma_i^2}\mathbf{C}_{j_k i}^H \mathbf{h}_{j_k i}\mathbf{h}_{j_k i}^H \mathbf{C}_{j_k i}\right)^{-1} \mathbf{h}_{j_k k}}{\left\|\left(\frac{1}{\eta_{j_k}}\mathbf{I}_{N_{j_k}} + \sum_{i=1}^{K_r} \frac{1}{\sigma_i^2}\mathbf{C}_{j_k i}^H \mathbf{h}_{j_k i}\mathbf{h}_{j_k i}^H \mathbf{C}_{j_k i}\right)^{-1} \mathbf{h}_{j_k k}\right\|_2} \quad \forall k \tag{3.33}$$

are called *signal-to-leakage-and-noise ratio maximizing (SLNR-MAX) beamforming.*

---

The expression in (3.33) resembles that of Wiener filtering, which is natural since $\mathrm{SLNR}_k$ is very similar to the uplink SINR of an multi-antenna receiver. By recognizing (3.32) as a generalized Rayleigh quotient, we have the following results.

---

**Lemma 3.14.** The beamforming direction $\bar{\mathbf{v}}_{j_k k}^{(\mathrm{SLNR})}$ maximizes $\mathrm{SLNR}_k$.

---

[7] The SLNR criterion is only well-defined if it is invariant to noise normalizations, because the SINRs are invariant in this respect. In [221, 259], $\mathrm{SLNR}_k$ was defined with $\sigma_k^2$ replacing all $\sigma_i^2$ in the denominator of (3.32). This has the strange consequence that scaling the noise power $\sigma_k^2$ and all the channels $\mathbf{h}_{jk}^H$ of $\mathrm{MS}_k$ by a common factor $c > 1$ will increase $\mathrm{SLNR}_k$ and decrease $\mathrm{SLNR}_i$ for $i \neq k$, although all SINRs are unaffected.

*Proof.* As the beamforming direction $\bar{\mathbf{v}}_{j_k k}$ satisfies $\|\bar{\mathbf{v}}_{j_k k}\|_2 = 1$, we can rewrite (3.32) as $\frac{\bar{\mathbf{v}}_{j_k k}^H \mathbf{a}\mathbf{a}^H \bar{\mathbf{v}}_{j_k k}}{\bar{\mathbf{v}}_{j_k k}^H \mathbf{B} \bar{\mathbf{v}}_{j_k k}}$, where $\mathbf{a} = \mathbf{C}_{j_k k}^H \mathbf{h}_{j_k k}$ and $\mathbf{B} = \frac{1}{\eta_{j_k}} \mathbf{I}_{N_{j_k}} + \sum_{i \neq k} \frac{1}{\sigma_i^2} \mathbf{C}_{j_k i}^H \mathbf{h}_{j_k i} \mathbf{h}_{j_k i}^H \mathbf{C}_{j_k i}$. The optimal direction can be found by minimizing $\bar{\mathbf{v}}_{j_k k}^H \mathbf{B} \bar{\mathbf{v}}_{j_k k}$ subject to $\mathbf{a}^H \bar{\mathbf{v}}_{j_k k} = 1$ (where the unit-norm constraint has been dropped and the phase has been specified). This is equivalent to minimizing $\bar{\mathbf{v}}_{j_k k}^H (\mathbf{B} + \mathbf{a}\mathbf{a}^H) \bar{\mathbf{v}}_{j_k k}$ subject to $\bar{\mathbf{v}}_{j_k k}^H \mathbf{a} = 1$ (as $\bar{\mathbf{v}}_{j_k k}^H \mathbf{a}\mathbf{a}^H \bar{\mathbf{v}}_{j_k k} = 1$). The stationarity KKT condition (2.13) implies that $\bar{\mathbf{v}}_{j_k k} = (\mathbf{B} + \mathbf{a}\mathbf{a}^H)^{-1} \mathbf{a}$, which is just a scaled version of $\bar{\mathbf{v}}_{j_k k}^{(\mathrm{SLNR})}$ in (3.33) since $\mathbf{C}_{j_k k} = \mathbf{I}_{N_{j_k}}$. □

Lemma 3.14 motivates the terminology SLNR-MAX beamforming, but there are many other names and alternative motivations for this heuristic transmit direction.

---

**Remark 3.2 (Many Terms for Essentially the Same Thing).** The principle of SLNR-MAX beamforming has been reinvented and remotivated many times by different authors in the past two decades. Some of the earliest works are [320] that suggests selecting the transmit weighting vector "such that the quotient of the mean power of the desired contribution to the undesired contributions is maximized," [79] that finds (3.33) by maximizing the harmonic mean of SINRs, and [88] that minimizes the average interference power subject to a desired received signal power constraint (as in the proof of Lemma 3.14).

By selecting the parameters $\{\eta_j\}_{j=1}^{K_t}$ based on power constraints and minimization of certain sum MSEs, (3.33) was derived as the *constrained MMSE transmit filter* in [285] and the *transmit Wiener filter* in [115]. The parameter $\eta_j$ can also be selected to achieve numerical stability, robustness to channel uncertainty, and avoid performance saturation in the large-system regime as in *regularized channel inversion* (zero-forcing) [203, 287].

The SLNR terminology is coined in [221, 259] as an optimization criterion that decouples the beamforming selection for interference channels and enables closed-form expressions. A similar motivation is used in [142], where the equivalent *signal-to-generating-interference-plus*

*noise-ratio (SGINR)* criterion is used. The authors of [142] also showed that the approach maximizes the product of two SINRs, which maximizes a high-SINR approximation of the sum information rate. The maximization of the SLNR/SGINR is a generalized eigenvalue problem, leading to the name *generalized eigenvalue-based beamformer* in [253].

Recently, many motivations have appeared based on uplink–downlink duality (see Subsection 2.2.2). The *virtual-uplink MVDR beamforming* in [99] is derived by assuming equal power allocation in the virtual uplink, while the similar *virtual SINR maximizing beamforming* is considered in [310] and shown to obtain a Pareto optimal point for the two-user MISO interference channel. Generalizations such as the *layered virtual SINR beamforming* in [21, 311] and the *centralized/distributed virtual SINR beamforming* in [23, 18] utilize the duality to achieve heuristic solutions in more complicated multi-cell scenarios.

---

This remark shows that SLNR-MAX beamforming solves certain optimization problems, enables decoupled optimization, and has the beamforming structure suggested by uplink–downlink duality. These are evidence explaining why the approach provides remarkably good performance and can be derived in many different ways, although it is generally a suboptimal transmit strategy. We have the following result on the optimality in coordinated beamforming systems.

---

**Corollary 3.15.** In coordinated beamforming with per-transmitter constraints, $\bar{\mathbf{v}}_{j_k k}^{(\mathrm{SLNR})} \to \bar{\mathbf{v}}_{j_k k}^{(\mathrm{MRT})}$ as $\frac{q_{j_k}}{\sigma_i^2} \to 0 \, \forall k, i$. For a scheduling set $\mathcal{S}$ such that ZFBF exists, $\bar{\mathbf{v}}_{j_k k}^{(\mathrm{SLNR})} \to \bar{\mathbf{v}}_{j_k k}^{(\mathrm{ZFBF})}$ as $q_j \to \infty \, \forall j$ (when only users in $\mathcal{S}$ are active).

---

This corollary is proved by observing that SLNR-MAX beamforming is a special case of the beamforming parametrization in Corollary 3.10 (with $\lambda_k = 1$ and $\mu_j = \frac{q_j}{\eta_j}$), which was used to prove the asymptotic optimalities of MRT and ZFBF. This connection has two important implications. First, SLNR-MAX beamforming has the optimal beamforming structure. Second, it combines the benefit of MRT at low SNR (when the interference terms in (3.33) are negligible)

with the benefit of ZFBF at high SNR (when the interference terms in (3.33) dominates over the identity matrix). The balancing between these extremes is illustrated geometrically in Figure 3.3 and in terms of channel gain regions in Figure 3.4. Furthermore, SLNR-MAX always exists while ZFBF is only possible under certain conditions (see Lemma 3.12). In other words, it is more versatile than MRT and ZFBF and should always be used instead of these — except perhaps for asymptotic performance analysis.

The definition of SLNR-MAX beamforming assumes perfect CSI, but heuristic extensions are possible (see [21, 23]) by taking $\bar{\mathbf{v}}_{j_k k}^{(\mathrm{SLNR})}$ as the dominating eigenvector of

$$\left(\frac{1}{\eta_{j_k}}\mathbf{I}_{N_{j_k}} + \sum_{i=1}^{K_r} \frac{1}{\sigma_i^2}\mathbf{E}_{j_k i}\right)^{-1/2} \mathbf{E}_{j_k k} \left(\frac{1}{\eta_{j_k}}\mathbf{I}_{N_{j_k}} + \sum_{i=1}^{K_r} \frac{1}{\sigma_i^2}\mathbf{E}_{j_k i}\right)^{-1/2}, \tag{3.34}$$

where $\mathbf{E}_{j_k i} = \mathbb{E}\{\mathbf{C}_{j_k i}^H \mathbf{h}_{j_k i} \mathbf{h}_{j_k i}^H \mathbf{C}_{j_k i}\}$ average over the CSI uncertainty.

### 3.4.4 Power Allocation

The preceding subsections defined MRT, ZFBF, and SLNR-MAX as heuristic ways of selecting the beamforming directions $\{\bar{\mathbf{v}}_{j_k k}\}_{k=1}^{K_r}$. When these have been selected, the power allocation $\{p_{j_k k}\}_{k=1}^{K_r}$ will ultimately determine the operating point in the performance region that is achieved by the heuristic transmit strategy. For given $\{\bar{\mathbf{v}}_{j_k k}\}_{k=1}^{K_r}$, the SINRs in (3.25) become

$$\mathrm{SINR}_k = \frac{p_{j_k k}\rho_{kk}}{\sigma_k^2 + \sum_{i \neq k} p_{j_i i}\rho_{ik}} \tag{3.35}$$

with fixed $\rho_{ik} = |\mathbf{h}_{j_i k}^H \mathbf{C}_{j_i k} \bar{\mathbf{v}}_{j_i i}|^2$ for all $k, i$. This SINR expression has the same structure as the SINRs with single-antenna transmitters; in fact, single-stream beamforming effectively transforms all MISO channels into SISO channels (which hopefully have good properties). Consequently, the power allocation can be optimized as in Subsection 2.2.4, where single-objective optimization with single-antenna transmitters was considered. Recall that most scalarizations of (3.25) lead to convex problem formulations in this scenario, with the weighted arithmetic mean as a notable exception.

The power allocation can be solved explicitly under ZFBF, because all the effective interfering channels are zero: $\rho_{ik} = 0$ for $i \neq k$. As shown in Subsection 2.2.1 (for concave user performance functions), even the weighted arithmetic mean gives convex problem formulations in this special case. The solution is given by the following theorem.

---

**Theorem 3.16.** Suppose the user performance functions $g_k(\cdot)$ are concave functions with invertible derivatives. For a given $\mathrm{BS}_j$ and coefficients $\rho_{kk} > 0 \; \forall k \in \mathcal{D}_j$, the power allocation problem

$$\begin{aligned} &\underset{p_{jk} \geq 0 \; \forall k \in \mathcal{D}_j}{\text{maximize}} \sum_{k \in \mathcal{D}_j} w_k g_k \left( p_{jk} \frac{\rho_{kk}}{\sigma_k^2} \right) \\ &\text{subject to} \sum_{k \in \mathcal{D}_j} p_{jk} \leq q_j \end{aligned} \tag{3.36}$$

is solved by

$$p_{jk} = \left[ \frac{\sigma_k^2}{\rho_{kk}} g_k'^{-1} \left( \frac{\sigma_k^2}{\nu_j w_k \rho_{kk}} \right) \right]_+ \quad \forall k \in \mathcal{D}_j, \tag{3.37}$$

where $\frac{d}{dx} g_k(x) = g_k'(x)$, $[\cdot]_+$ replaces negative values with zero, and the parameter $\nu_j \geq 0$ is selected to use full power.

---

*Proof.* The Lagrangian function of the convex problem in (3.36) is

$$\mathcal{L} = - \sum_{k \in \mathcal{D}_j} w_k g_k \left( p_{jk} \frac{\rho_{kk}}{\sigma_k^2} \right) + \frac{1}{\nu_j} \left( \sum_{k \in \mathcal{D}_j} p_{jk} - q_j \right), \tag{3.38}$$

where $\frac{1}{\nu_j}$ is the Lagrange multiplier (for notational convenience). The solution (3.37) is obtained from the stationarity KKT condition in (2.13). Finally, Theorem 1.9 requires that $\nu_j$ is scaled to satisfy the power constraint with equality. □

The optimal power allocation under ZFBF depends on the first derivative of the user performance function $g_k(\cdot)$ and on the weighting factors. To exemplify the structure, the power allocations for the

information rate and MSE are based on

$$g_k(x) = \log_2(1+x) \qquad \Rightarrow\ g_k'^{-1}(y) = \frac{1}{y\log_e(2)} - 1, \tag{3.39}$$

$$g_k(x) = \frac{x}{1+x} \qquad \Rightarrow\ g_k'^{-1}(y) = \frac{1}{\sqrt{y}} - 1, \tag{3.40}$$

respectively. The resulting power allocation in (3.37) becomes

$$p_{j_k k} = \left[\nu_j \varrho_k - \frac{\sigma_k^2}{\rho_{kk}}\right]_+ \text{ where } \begin{cases} \varrho_k = w_k & \text{for information rate,}^8 \\ \varrho_k = \sqrt{\frac{w_k \sigma_k^2}{\rho_{kk}}} & \text{for MSE.} \end{cases}$$

Both power allocations result in so-called *waterfilling* solutions [37] with the characteristics: (a) power is allocated according to some user-dependent factor $\varrho_k > 0$; and (b) zero power might be allocated to users with the weakest channels/weights. The water terminology originates from viewing the power allocation as pouring water into a tank with an uneven bottom. Each user is represented by a column of width $\varrho_k$ and height $\frac{\sigma_k^2}{\rho_{kk}\varrho_k}$. The water area above the column equals the power allocated to this user, and water might not reach up to this level. The water-level $\nu_j$ is selected to make the total water area equal the total transmit power. The waterfilling interpretation is visualized in Figure 3.5.

The power is allocated proportionally to $\varrho_k$ when $\nu_j$ is large (i.e., $q_j$ is large), while zero power is allocated to $\text{MS}_k$ when $\nu_j \leq \frac{\sigma_k^2}{\rho_{kk}\varrho_k}$. These asymptotic properties have different consequences for the information rate and MSE, since $\varrho_k$ is different. In case of equal user weights (i.e., $w_k = \frac{1}{K_r} \forall k$), the information rate activates a user when the waterlevel is $\nu_j \geq \frac{K_r \sigma_k^2}{\rho_{kk}}$ and performs uniform power allocation when $\nu_j$ is large. The waterfilling when using the MSE activates $\text{MS}_k$ when $\nu_j \geq \sqrt{\frac{K_r \sigma_k^2}{\rho_{kk}}}$ and allocates power proportional to $\sqrt{\frac{\sigma_k^2}{\rho_{kk}}}$ when $\nu_j$ is large — the MSE is therefore activating weak users earlier and asymptotically allocates more power to these users (to even out the user conditions). The user weights will however have a similar impact on both user performance functions: increasing $w_k$ will increase the power allocated to $\text{MS}_k$.

[8] The constant factor $\log_e(2)$ has been included in $\nu_j$ for notational convenience.

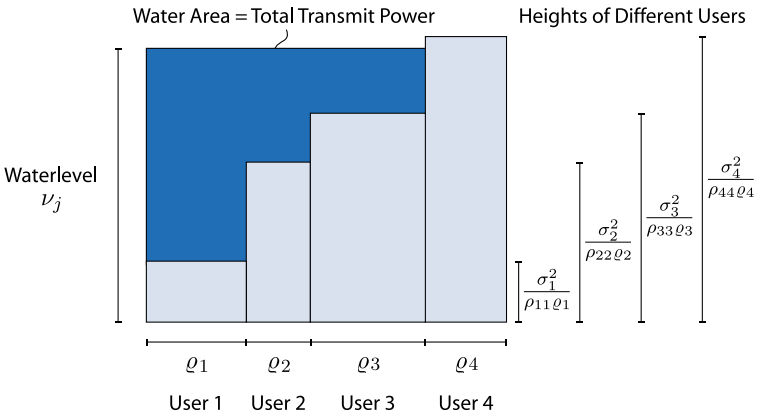



Fig. 3.5 Illustration of the power allocation $p_{j_k k} = \left[\nu_j \varrho_k - \frac{\sigma_k^2}{\rho_{kk}}\right]_+$ for $k = 1, \ldots, 4$. This formula is called *waterfilling* as it can be interpreted as pouring water into a tank with an uneven bottom. Each column represents a user: the width $\varrho_k$ depends on the user performance function and the area $\frac{\sigma_k^2}{\rho_{kk}}$ is inversely proportional to its SNR $\frac{\rho_{kk}}{\sigma_k^2}$. The water area above the column equals the power allocated to this user, and the waterlevel $\nu_j$ makes the total water area equal the total transmit power.

Under coordinated ZFBF beamforming with per-transmitter constraints, Theorem 3.16 solves the power allocation problem explicitly. For other heuristic beamforming strategies such as SLNR-MAX and MRT, the corresponding power allocation is solved optimally using the techniques in Section 2. However, Theorem 3.16 can be used for heuristic power allocation, by either pretending that ZFBF is used during power allocation or by ignoring the inter-user interference [18]. We will return to heuristic beamforming and power allocation in Section 4.2 in the context of distributed resource allocation.

### 3.4.5 Numerical Comparison

To illustrate the behavior of different heuristic beamforming directions, we consider a 4-user MISO interference channel with $N_j = 4$ antennas per base station and global interference coordination. The channel vectors $\mathbf{h}_{jk}$ are generated as uncorrelated Rayleigh fading and the average channel gains $\frac{\mathbb{E}\{\|\mathbf{h}_{jk}\|_2^2\}}{\sigma_k^2}$ equal $N_j$ for the serving base station and $\frac{N_j}{2}$ for all interfering base stations.

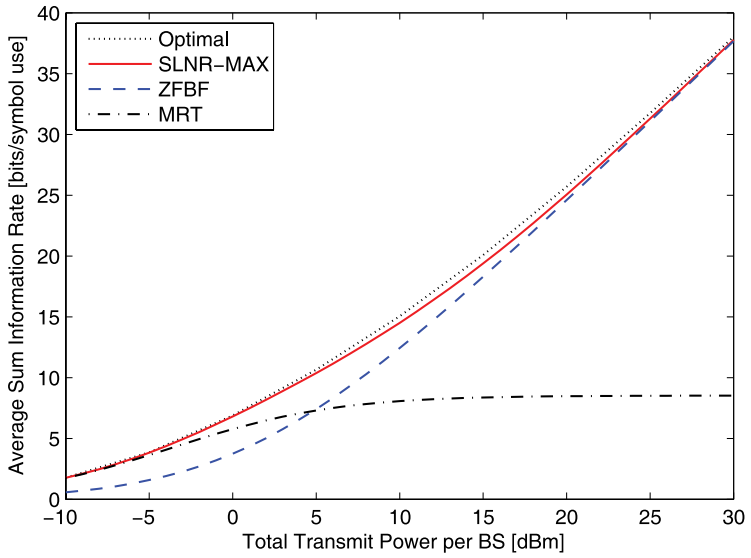



Fig. 3.6 Average sum information rate (over channel realizations) with SLNR-MAX, MRT, and ZFBF in a 4-user MISO interference channel, as a function of the transmit power. MRT and ZFBF are good beamforming directions at low and high SNR, respectively, while SLNR-MAX shows good performance in the entire SNR range.

The average achievable sum information rate is shown in Figure 3.6 as a function of the total transmit power (per base station). The optimal transmit strategy is computed using the BRB algorithm in Section 2.3. As expected from Corollaries 3.11, 3.13, and 3.15, MRT is good at very low SNR and ZFBF is good at high SNR. However, SLNR-MAX is a more versatile strategy as it combines the respective asymptotic benefits of MRT and ZF and clearly outperforms them at intermediate SNRs by being remarkably close to the optimal solution.

## 3.5 General Guidelines for Solving Multi-Objective Resource Allocation Problems

The multi-objective resource allocation problem in (3.1) provides a mathematical formulation for the conflicting interests of users. The Pareto boundary $\partial^+\mathcal{R}$ of the performance region $\mathcal{R}$ represents all tentative solutions and $\partial^+\mathcal{R}$ is a surface of dimension $K_r - 1$. To identify and select a final operating point $\mathbf{g}^* \in \mathcal{R}$, the system designer needs to

formulate its subjective preference of different $\mathbf{g} \in \mathcal{R}$. This subjective input might be invariant and known *a priori*, but it can also be refined during the optimization procedure as partial knowledge on the shape of $\mathcal{R}$ is obtained. This can be described as a *psychological convergence* as it is hard (if not impossible) to exactly formulate the subjective preference without knowing all the alternatives.

There is certainly a tradeoff between computational complexity and the possibility for the system designer to iteratively refine its subjective preference. However, it is even more important to formulate the preference in a way that facilitates numerical optimization. Section 2 showed that scalarizations of the MOP are generally NP-hard, but we identified some cases that lead to convex problem formulations and thus are solvable in polynomial time; see Table 2.1. A *pragmatic approach* to resource allocation would therefore be to select one of these cases and then let the weighting factors be adapted to the subjective preference.

We will now bring these insights into practical use by providing general guidelines for solving (3.1) efficiently. We will differentiate between the four categories suggested in [38, 324]: no-preference methods, *a priori* methods, *a posteriori* methods, and interactive methods. These categories represent different types of input from the system designers.

### 3.5.1 No-Preference Methods

If the system designer has no subjective preference on the final solution and is satisfied with any Pareto optimal point, it makes sense to optimize some single-objective problem that allocates resources proportionally to the user conditions. The proportionality can be achieved by selecting weighting factors based on the utopia point as $w_k = \frac{u_k}{\sum_{i=1}^{K_r} u_i}$ (see Lemma 1.3), which represents the fraction of the aggregate performance that $\mathrm{MS}_k$ achieves under TDMA. To achieve a convex problem formulation, we recommend the system utility function $f(\mathbf{g}) = \min_k \frac{g_k - a_k}{w_k}$, which is called fairness-profile optimization (see Subsection 2.2.3 and Example 2.8). This function gives the so-called Kalai–Smorodinsky bargaining solution that provides a type of relative fairness [193] — it is obtained in game theory by defining a set of axioms

on what would be a reasonable bargaining solution. The start-point $\mathbf{a} = [a_1 \dots a_{K_r}]^T$ can, for example, be the origin or be achieved from the beamforming parametrization in Theorem 3.5 by using the weighting factors as user priorities (i.e., $\lambda_k = w_k \, \forall k$) and equal enforcement of all power constraints (i.e., $\mu_l = \frac{1}{L} \, \forall l$).

### 3.5.2 A Priori Methods

If the system designer knows in advance which system utility function $f(\cdot)$ that should be optimized, then Table 2.1 in Section 2 shows whether the corresponding resource allocation is a convex or monotonic problem. If the problem is convex (or quasi-convex), it can be solved to global optimality in polynomial time, thus the optimal beamforming solution can be obtained and used in practice (at least if the coherence time of the channels is sufficiently long). Some important convex examples are:

- Quality-of-service requirements: If we want to achieve a point $\mathbf{r}^* = [r_1^* \dots r_{K_r}^*]^T \in \mathcal{R}$, a feasible beamforming solution is obtained by solving the convex feasibility problem in (2.29). Details are given in Subsection 2.2.2.
- Weighted Chebyshev compromise: If we want to achieve a point $\mathbf{r}^* = [r_1^* \dots r_{K_r}^*]^T \notin \mathcal{R}$, then we can find an alternative point $\mathbf{g} \in \mathcal{R}$ that is as close to $\mathbf{r}^*$ as possible. This problem is quasi-convex if the $L_\infty$-norm (also known as Chebyshev metric) is used: $f(\mathbf{g}) = -\max_k w_k(r_k^* - g_k)$. Details are given in Example 2.9.
- Fairness-profile optimization: If we want to guarantee $g_k \geq a_k$ for each user and divide remaining resources so that each user gets a predefined fraction $w_k > 0$, then $f(\mathbf{g}) = \min_k \frac{g_k - a_k}{w_k}$. This problem is quasi-convex and details are given in Example 2.7.
- General zero-forcing beamforming: If the power constraints dictate zero inter-user interference, then the resource allocation problem is convex whenever the system utility and user performance functions are concave. This can sometimes be generalized to nonzero interference constraints, details are

given in Subsection 2.2.1. Closed-form solutions are also possible in special cases; see Section 3.4.

- Single-antenna transmission: Many resource allocation problems are convex when only a single antenna transmits to each user. Details are given in Subsection 2.2.4.

If the system designer is allowed to select the system utility function, we recommend a pragmatic approach: select one of the convex examples above. Fairness-profile optimization, $f(\mathbf{g}) = \min_k \frac{g_k - a_k}{w_k}$, is a good choice because (a) it can be solved in polynomial time and (b) any Pareto optimal point can be achieved by some set of weighting factors $\{w_k\}_{k=1}^{K_r}$. The weights are then used to balance between user fairness and aggregate throughput — the former is represented by identical weights for all users and the latter by giving larger weights to users with strong channel conditions.

There are practical scenarios when the system designer is stuck with a system utility function that gives a monotonic problem. For example, we might want to optimize a utility function defined on the long-term average user performances (instead of instantaneous values). This can be achieved by stochastic network optimization [78, 244], which essentially consists of a sequence of weighted sum performance optimizations. The weights are updated between each time slot using virtual queuing techniques. As the computational complexity of general monotonic problems scales exponentially with the number of users, some kind of approximation is necessary in practice. Either the system utility function is approximated using one of the convex formulations listed above, or we can search for an approximate solution. For example, the beamforming directions can be approximated by plugging heuristic parameter values into the beamforming parameterizations in Section 3.2. The remaining power allocation problem is convex in many cases (see Remark 2.7), but can also be approximated using Theorem 3.16. There are also iterative algorithms for finding locally optimal points; see [104, 150, 201, 225, 243, 257, 280, 291, 293] and Section 4.2. The performance of any approximate beamforming strategy can be evaluated by solving the original problem using the PA or BRB algorithms.

### 3.5.3 *A Posteriori* Methods

As the performance region is multi-dimensional and not explicitly known, it is difficult to foresee what would be a good outcome of the resource allocation. As an example, suppose that users have near-orthogonal channel directions and homogeneous channel conditions, then we can expect good fairness even if the system utility function ignores fairness. On the other hand, it might be important to emphasize fairness when the users have strongly heterogeneous channel conditions. This illustrates the difficulty in selecting $f(\cdot)$ in advance.

If the system designer is unable to formalize its preference in advance, we can compute a set of sample point that roughly describes the shape of the performance region. These points are then analyzed by the system designer and a suitable operating point is selected — this is called an *a posteriori* method as the preference is defined after the numerical procedures. A computationally efficient approach is to generate sample points using the beamforming parametrization of the Pareto boundary in Section 3.3.

### 3.5.4 Interactive Methods

The PA and BRB algorithms in Section 2.3 are designed to solve any monotonic resource allocation problem in an iterative manner. Although both algorithms assume that the system utility function is fixed, it can actually be updated in every iteration based on the new knowledge that has been obtained (e.g., the new feasible point $\mathbf{n}$). The system designer can thereby achieve psychological convergence since both the choice of system utility function and the best feasible solution converges. The enabling factor is that the algorithms approximate the Pareto boundary (where all tentative solutions to the MOP are located) and $f(\cdot)$ only determine which part of the approximation that should be refined in the next iteration. It is however important not to remove vertices in the PA algorithm (see Step 10 in Algorithm 2) and not to reduce boxes in the BRB algorithm (see Step 7 in Algorithm 3), since the removed parts might contain operating point of later interest.

Alternatively, the interactive method can be based on the beamforming parametrization of the Pareto boundary in Section 3.3.

Initially, a set of sample points is generated by evaluating the performance over a grid of parameter values. The system designer selects one or a few promising points and then new sample points are generated by varying the parameters around the values that generated the selected points. This procedure is iterated until psychological convergence is achieved.

Interactive methods are more computational expensive than the three other methods and thus less suitable for practical applications.

## 3.6 Summary

Resource allocation is generally a multi-objective optimization problem where the performance maximization of the $K_r$ users constitutes the $K_r$ conflicting objectives. The $N$-dimensional beamforming vectors are the optimization variables and should satisfy the $L$ power constraints. The search-space initially contains $K_r N$ complex-valued parameters, but this can be greatly reduced by exploiting the structure of Pareto optimal beamforming solutions. This section has presented some state-of-the-art beamforming parametrizations only requiring $K_r + L - 2$ or $K_r(K_r - 1)$ numbers between zero and one — the number of parameters depends on whether they are selected centrally or distributively (and on the distributiveness of the power constraints). The strength of the parametrizations is easily observed by their tight connection to many heuristic beamforming strategies that have been developed through the years. They also provide a foundation for obtaining sample points that illustrate the shape of the Pareto boundary, thus clarifying the available options in the resource allocation.

There are certain main categories of methods that solve multi-objective optimization problems. These represent different types of involvement of the system designer when selecting an appropriate operating point on the Pareto boundary. If the system designer can formulate its preferences as a system utility function, the corresponding single-objective resource allocation problem can be solved as described in Section 2. The problem can be solved to global optimality if it is convex (or quasi-convex), while heuristic approximations (e.g., based on beamforming parametrizations) are otherwise necessary for practical

feasibility. A pragmatic approach would therefore be to always formulate the resource allocation within the category of convex problems. Alternatively, a set of sample points in the performance region can be generated as described in this section. The system designer can then analyze these points and make an informed decision. Finally, the monotonic optimization algorithms from Section 2 can be utilized in an interactive manner, meaning that the system utility function is iteratively refined to achieve both psychological and numerical convergence to the most suitable point on the Pareto boundary.

# 4

# Extensions and Generalizations

The multi-cell system model analyzed in the previous sections was based on three simplifying assumptions: perfect CSI, unlimited backhaul capacity, and ideal transceiver hardware. These assumptions are generalized in this section to more realistic conditions, and we will show that many of the previous results are readily generalizable. Furthermore, we describe how the system model can be extended to also incorporate multi-cast transmission, multi-carrier systems, and multi-antenna receivers. Finally, we discuss some recent work on the design of dynamic cooperation clusters and show that the framework of this tutorial is applicable in cognitive radio systems and for physical layer security.

Although many of the generalizations and extensions in this section are mutually feasible, the sections describe each of them independently. The combinations of multiple generalizations can be viewed as interesting topics for future research. Matlab code for some of the algorithms developed in this section is available for download in [19].

## 4.1 Robustness to Channel Uncertainty

The analysis in previous sections was based on the assumption that the channel vectors $\mathbf{h}_k$ are perfectly known at the base stations. This simplified the presentation of optimization algorithms for resource allocation, but it is clearly an ideal model as practical transmitters have to operate under uncertain CSI. The uncertainty originates from a variety of sources; for example, imperfect channel estimation, feedback quantization, inadequate channel reciprocity, and delays in CSI acquisition on fading channels. It is common to have an additive error model [9, 17, 26, 50, 239, 241, 242, 257, 286, 289, 328] with

$$\mathbf{h}_k = \widehat{\mathbf{h}}_k + \tilde{\boldsymbol{\epsilon}}_k \quad \forall k, \tag{4.1}$$

where $\widehat{\mathbf{h}}_k = [\widehat{\mathbf{h}}_{1k}^T \dots \widehat{\mathbf{h}}_{K_t k}^T]^T \in \mathbb{C}^{N \times 1}$ is the nominal value of the CSI available at the base stations and $\tilde{\boldsymbol{\epsilon}}_k \in \mathbb{C}^{N \times 1}$ is the combined error vector. This model can, for instance, be motivated by viewing channel estimation as the main source of uncertainty [13, 22, 138]; see Example 4.1 below. Observe that the nominal channel and the error should both be set to zero for all $\mathbf{h}_{jk}$ with $k \notin \mathcal{C}_j$, because CSI is not acquired for these channels and their impact are included in the noise terms.

The purpose of this section is to present a framework for handling CSI uncertainty in a robust manner, while enabling generalizations of the results from previous sections. Robustness refers to ensuring a certain level of performance under the error model in (4.1). The system cannot account for any error; the stochastic error vector $\tilde{\boldsymbol{\epsilon}}_k$ could potentially cancel out the nominal vector as $\tilde{\boldsymbol{\epsilon}}_k = -\widehat{\mathbf{h}}_k$ or be very large (the distribution is even unbounded for Rayleigh fading channels). This is often handled by only considering a subset of error vectors, *the uncertainty set*, that has high probability of containing the error [9, 17, 26, 50, 239, 241, 242, 257, 286, 289, 328]. If the design of these sets is explicitly included in the resource allocation (e.g., optimization with acceptable outage probabilities), it seems that conservative approximations[1] of each user's performance are required to achieve tractable problem formulations [50, 241, 289]. The alternative is to have fixed uncertainty sets and maximizing the worst-case performance, which is

[1] Conservative approximation means optimizing lower bounds on the user's performance.

mathematically more convenient as it can provide convex problem formulations [17, 26, 242, 257]. Therefore, this section describes a set of worst-case robustness approaches for multi-cell resource allocation.

Worst-case robustness is sometimes accused of giving conservative performance results [80], because the worst case could have very low probability in practice. However, this is not a fundamental weakness of the approach but the result of using ill-structured uncertainty sets and can be avoided by proper selection of them.[2] Furthermore, one can argue that we choose between solving a good problem formulation in a conservative way and solving a conservative problem formulation in an efficient way — the choice is thus a matter of taste.

For analytic convenience and motivated by channel estimation [26, 242, 257], we consider (compact) ellipsoidal channel uncertainty sets. These sets can either be defined jointly for all base stations to a certain user,

$$\mathcal{U}_k(\widehat{\mathbf{h}}_k, \mathbf{B}_k) = \left\{ \mathbf{h}_k : \mathbf{h}_k = \widehat{\mathbf{h}}_k + \mathbf{B}_k \boldsymbol{\epsilon}_k,\ \|\boldsymbol{\epsilon}_k\|_2 \leq 1 \right\} \quad \forall k, \tag{4.2}$$

or separately for the channel from each $\mathrm{BS}_j$ to each user $k \in \mathcal{C}_j$,

$$\mathcal{U}_{jk}(\widehat{\mathbf{h}}_{jk}, \mathbf{B}_{jk}) = \left\{ \mathbf{h}_{jk} : \mathbf{h}_{jk} = \widehat{\mathbf{h}}_{jk} + \mathbf{B}_{jk} \boldsymbol{\epsilon}_{jk},\ \|\boldsymbol{\epsilon}_{jk}\|_2 \leq 1 \right\} \quad \forall j, k \in \mathcal{C}_j. \tag{4.3}$$

Looking at the original additive model in (4.1), $\tilde{\boldsymbol{\epsilon}}_k = \mathbf{B}_k \boldsymbol{\epsilon}_k$ and $\tilde{\boldsymbol{\epsilon}}_{jk} = \mathbf{B}_{jk} \boldsymbol{\epsilon}_{jk}$, respectively. The matrices $\mathbf{B}_k \succeq \mathbf{0}_N$ and $\mathbf{B}_{jk} \succeq \mathbf{0}_{N_j}$ define the shapes of the ellipsoids, as illustrated in Figure 4.1. The uncertain CSI at the transmitters is now characterized by $\{\mathcal{U}_k\}_{k=1}^{K_r}$ and $\{\mathcal{U}_{jk}\}_{j=1,k=1}^{K_t,K_r}$, respectively, along with the corresponding nominal vectors and ellipsoid shaping matrices.

To elaborate the difference between (4.2) and (4.3), we consider the special case of $\mathbf{B}_k = \sqrt{K_t}\mathrm{diag}(\mathbf{B}_{1k}, \ldots, \mathbf{B}_{K_t k})$, where the block-diagonal structure is motivated by separate estimation/quantization between the transmitters. The uncertainty then have the same general shape,

[2] In the probabilistic approach, the guaranteed performance is maximized under a given outage probability. For an optimal transmit strategy, we can create a set $\mathcal{U}$ of all error vectors that gives exactly the optimal guaranteed performance (or better). If $\mathcal{U}$ is used as the uncertainty set in the worst-case approach, it will provide the same optimal transmit strategy and will not be any more conservative.

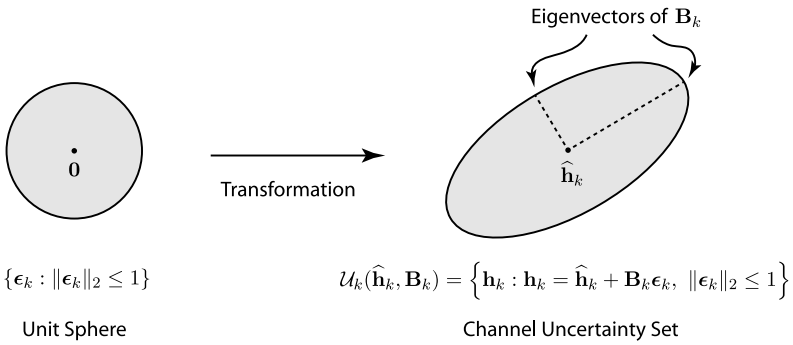



Fig. 4.1 Illustration of the ellipsoidal uncertainty set $\mathcal{U}_k(\widehat{\mathbf{h}}_k, \mathbf{B}_k)$ in (4.2) and the unit sphere $\{\boldsymbol{\epsilon}_k : \|\boldsymbol{\epsilon}_k\|_2 \leq 1\}$ that it is created from.

but (4.2) is more generous as the error in $\mathbf{h}_{jk}$ can be increased by decreasing the error in some other channel component $\mathbf{h}_{ji}$ with $i \neq k$. The uncertainty in (4.3) is independent between each channel component $\mathbf{h}_{jk}$. In other words, $\mathcal{U}_k$ and $\mathcal{U}_{jk}$ represent two different ways of cutting out uncertainty sets from the underlying CSI uncertainty and the choice is a matter of uncertainty modeling.

The uncertainty sets can either be designed experimentally or by modeling the underlying uncertainty sources. The latter approach is illustrated by the following example.

---

**Example 4.1 (Channel Estimation Uncertainty).** If the channel vectors are modeled as Rayleigh fading, $\mathbf{h}_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}_k)$, and estimated using training signaling, then the error vector will be distributed as $\tilde{\boldsymbol{\epsilon}}_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{E}_k)$. The shape of the error covariance matrix $\mathbf{E}_k$ depends on $\mathbf{R}_k$ and on the type of channel estimation (e.g., least-squares estimation [13] or minimum mean square error estimation [22, 138]) and will be block-diagonal as $\mathbf{E}_k = \text{diag}(\mathbf{E}_{1k}, \ldots, \mathbf{E}_{K_t k})$ if the channel from each base station to each user is estimated separately.

The estimation error $\tilde{\boldsymbol{\epsilon}}_k$ belongs with probability $\tilde{p}_k$ to the ellipsoidal set $\{\tilde{\boldsymbol{\epsilon}}_k : 2\tilde{\boldsymbol{\epsilon}}_k^H \mathbf{E}_k^{-1} \tilde{\boldsymbol{\epsilon}}_k \leq \chi^2_{\tilde{p}_k}(2N)\}$, where $\chi^2_{\tilde{p}_k}(n)$ denotes the $\tilde{p}_k$-percentile of the chi-squared distribution with $n$ degrees-of-freedom. If we limit the robustness to error vectors in this set, the channel uncertainty is given by (4.2) using $\mathbf{B}_k = \sqrt{\chi^2_{\tilde{p}_k}(2N)/2}\,\mathbf{E}_k^{1/2}$. To enforce higher or lower

robustness to errors on channels from some base stations (e.g., channel components that are expected to carry strong interference), we can include additional weighting factors on the diagonal blocks of $\mathbf{B}_k$.

Alternatively, the estimation error $\tilde{\boldsymbol{\epsilon}}_{jk}$ of the channel component $\mathbf{h}_{jk}$ belongs with probability $\tilde{p}_{jk}$ to the ellipsoidal set $\{\tilde{\boldsymbol{\epsilon}}_{jk} : 2\tilde{\boldsymbol{\epsilon}}_{jk}^H \mathbf{E}_{jk}^{-1} \tilde{\boldsymbol{\epsilon}}_{jk} \leq \chi^2_{\tilde{p}_{jk}}(2N_j)\}$. Limiting the robustness to error vectors in this set corresponds to (4.3) with $\mathbf{B}_{jk} = \sqrt{\chi^2_{\tilde{p}_{jk}}(2N_j)/2}\,\mathbf{E}_{jk}^{1/2}$.

### 4.1.1 Worst-Case Robustness under Joint Uncertainty

In this subsection, we show how to guarantee QoS requirements under worst-case robustness to CSI uncertainty. This is both a subproblem of various generalizations of max-min fairness (see Subsection 2.2.3) and of the PA and BRB algorithms for solving arbitrary monotonic problems (see Section 2.3), thus we provide a foundation for solving any resource allocation problem under CSI uncertainty. The approach builds on standard results in robust optimization and can be applied to other problem formulations as well.

The sufficiency of using single-stream beamforming to obtain any point in $\mathcal{R}$ was proved in Theorem 1.8 under perfect CSI. It is not obvious whether this will also hold under worst-case robustness. We therefore formulate the SOP with QoS requirements $g_k(\mathrm{SINR}_k) \geq r_k^*$ in (2.29) as

$$\begin{aligned} &\text{find } \mathbf{S}_1 \ldots, \mathbf{S}_{K_r} \\ &\text{subject to } \frac{\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_k \mathbf{S}_k \mathbf{D}_k^H \mathbf{C}_k^H \mathbf{h}_k}{\sigma_k^2 + \mathbf{h}_k^H \mathbf{C}_k (\sum_{i \neq k} \mathbf{D}_i \mathbf{S}_i \mathbf{D}_i^H) \mathbf{C}_k^H \mathbf{h}_k} \geq g_k^{-1}(r_k^*) \quad \forall \mathbf{h}_k \in \mathcal{U}_k, \forall k, \\ &\sum_{k=1}^{K_r} \mathrm{tr}(\mathbf{Q}_{lk} \mathbf{S}_k) \leq q_l \quad \forall l \end{aligned} \tag{4.4}$$

where the beamforming vectors have been replaced by signal correlation matrices $\mathbf{S}_k$. Joint ellipsoidal channel uncertainty sets are used in (4.4) and consequently there are infinitely many SINR constraints — one for each $\mathbf{h}_k \in \mathcal{U}_k$. This makes the problem fundamentally different from those considered in Section 2, but fortunately there is a way to obtain

a finite number constraints. The following is known as the S-procedure and is an important tool in robust optimization [35, Section 2.6.3].

---

**Lemma 4.1 (S-Procedure).** Let $\theta_m(\boldsymbol{\epsilon}) = \boldsymbol{\epsilon}^H \mathbf{Z}_m \boldsymbol{\epsilon} + \mathbf{z}_m^H \boldsymbol{\epsilon} + \boldsymbol{\epsilon}^H \mathbf{z}_m + \tilde{z}_m$, $m = 1, 2$, be two quadratic functions in $\boldsymbol{\epsilon}$ and let $\mathbf{Z}_m$ be Hermitian. Suppose it exists $\hat{\boldsymbol{\epsilon}}$ such that $\theta_1(\hat{\boldsymbol{\epsilon}}) > 0$, then the implication

$$\theta_1(\boldsymbol{\epsilon}) \geq 0 \Rightarrow \theta_2(\boldsymbol{\epsilon}) \geq 0 \tag{4.5}$$

holds true if and only if it exists $\lambda \geq 0$ such that

$$\begin{bmatrix} \mathbf{Z}_2 & \mathbf{z}_2 \\ \mathbf{z}_2^H & \tilde{z}_2 \end{bmatrix} - \lambda \begin{bmatrix} \mathbf{Z}_1 & \mathbf{z}_1 \\ \mathbf{z}_1^H & \tilde{z}_1 \end{bmatrix} \succeq \mathbf{0}. \tag{4.6}$$

---

If $\theta_1(\boldsymbol{\epsilon}) \geq 0$ describes the uncertainty set and $\theta_2(\boldsymbol{\epsilon}) \geq 0$ is an SINR constraint, then Lemma 4.1 provides a single condition (4.6) for proving that the SINR constraint holds for every vector in the uncertainty set. This observation is formalized in the following theorem.

---

**Theorem 4.2.** Let $\mathbf{S}_1, \ldots, \mathbf{S}_{K_r}$ be a transmit strategy and let $\gamma_k \geq 0$ be given, then the robust SINR constraint

$$\frac{\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_k \mathbf{S}_k \mathbf{D}_k^H \mathbf{C}_k^H \mathbf{h}_k}{\sigma_k^2 + \mathbf{h}_k^H \mathbf{C}_k (\sum\limits_{i \neq k} \mathbf{D}_i \mathbf{S}_i \mathbf{D}_i^H) \mathbf{C}_k^H \mathbf{h}_k} \geq \gamma_k \quad \forall \mathbf{h}_k \in \mathcal{U}_k(\widehat{\mathbf{h}}_k, \mathbf{B}_k) \tag{4.7}$$

is fulfilled if and only if it exists $\lambda_k \geq 0$ such that

$$\begin{bmatrix} \mathbf{B}_k^H \mathbf{A}_k \mathbf{B}_k & \mathbf{B}_k^H \mathbf{A}_k \widehat{\mathbf{h}}_k \\ \widehat{\mathbf{h}}_k^H \mathbf{A}_k \mathbf{B}_k & \widehat{\mathbf{h}}_k^H \mathbf{A}_k \widehat{\mathbf{h}}_k - \sigma_k^2 \end{bmatrix} + \begin{bmatrix} \lambda_k \mathbf{I}_N & \mathbf{0}_{N \times 1} \\ \mathbf{0}_{1 \times N} & -\lambda_k \end{bmatrix} \succeq \mathbf{0}_{N+1}, \tag{4.8}$$

where $\mathbf{A}_k = \frac{1}{\gamma_k} \mathbf{C}_k \mathbf{D}_k \mathbf{S}_k \mathbf{D}_k^H \mathbf{C}_k^H - \sum\limits_{i \neq k} \mathbf{C}_k \mathbf{D}_i \mathbf{S}_i \mathbf{D}_i^H \mathbf{C}_k^H$.

---

*Proof.* The channel uncertainty and SINR constraints can be expressed as $\theta_1(\boldsymbol{\epsilon}) \geq 0$ and $\theta_2(\boldsymbol{\epsilon}) \geq 0$, respectively, where

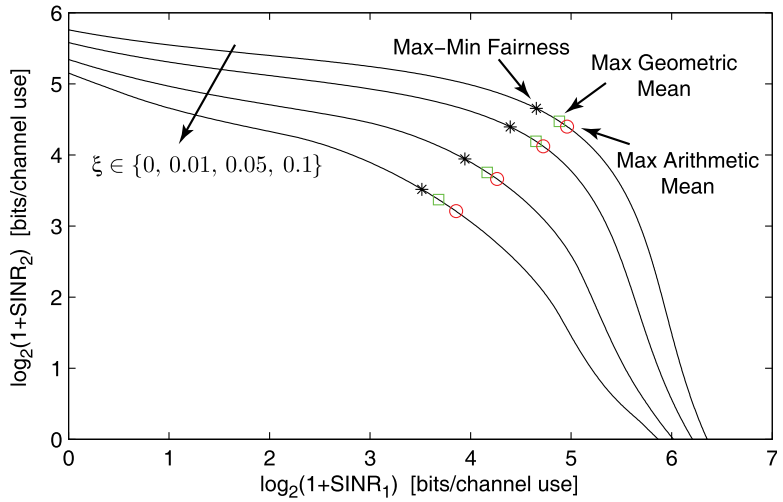
$$\begin{aligned} \theta_1(\boldsymbol{\epsilon}) &= -\boldsymbol{\epsilon}^H \mathbf{I}_N \boldsymbol{\epsilon} + 1 \\ \theta_2(\boldsymbol{\epsilon}) &= \boldsymbol{\epsilon}^H \mathbf{B}_k^H \mathbf{A}_k \mathbf{B}_k \boldsymbol{\epsilon} + \boldsymbol{\epsilon}^H \mathbf{B}_k^H \mathbf{A}_k \widehat{\mathbf{h}}_k \\ &\quad + \widehat{\mathbf{h}}_k^H \mathbf{A}_k \mathbf{B}_k \boldsymbol{\epsilon} + \widehat{\mathbf{h}}_k^H \mathbf{A}_k \widehat{\mathbf{h}}_k - \sigma_k^2. \end{aligned} \tag{4.9}$$

The theorem then follows directly from Lemma 4.1. □

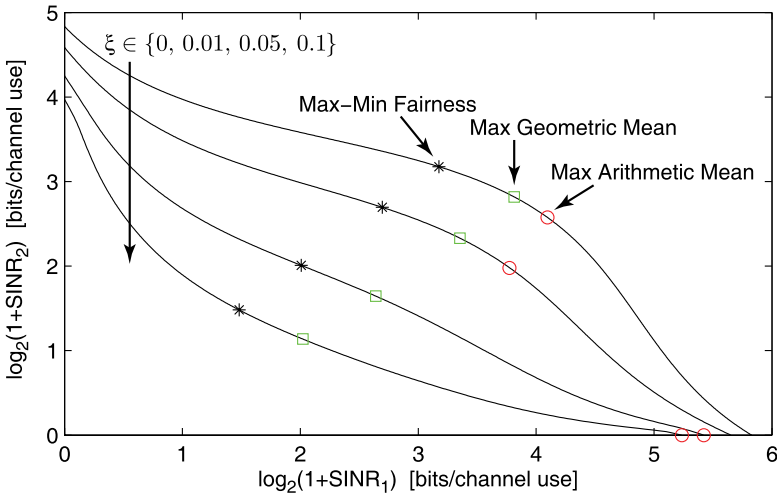
This theorem converts the infinite number of SINR constraints in (4.4) into just one (linear) semi-definite constraint per user — at the cost of adding $K_r$ extra variables $\{\lambda_k\}_{k=1}^{K_r}$ that indirectly represents the worst channel conditions in the uncertainty set; if we can find $\lambda_k \geq 0$ that satisfies the constraint (4.8), then $\text{SINR}_k \geq \gamma_k$ for all $\mathbf{h}_k \in \mathcal{U}_k$. Having channel uncertainty naturally increases the computational complexity (due to the additional optimization variables), but Theorem 4.2 shows that the robust problem in (4.4) is convex. The complexity is thus still polynomial in the number of antennas $N$, users $K_r$, and power constraints $L$ [10, Chapter 6].

The transmit strategy $\mathbf{S}_1^*, \ldots, \mathbf{S}_{K_r}^*$ that solves (4.4) might in general have $\text{rank}(\mathbf{S}_k^*) > 1$ for some users, which is not practically convenient as the computational complexity of decoding such multi-stream beamforming is relatively high [89]. As single-stream beamforming is always sufficient under perfect CSI, we can however expect it to also work well when the uncertainty is small. Recent work in [51, 251, 239] proves that single-stream beamforming is indeed sufficient for single-cell and coordinated beamforming scenarios with per-transmitter power constraints and some minor technical assumptions (e.g., a generous bound on the amount of uncertainty or uniqueness of the solution). Therefore, we expect single-stream beamforming to be sufficient also under channel uncertainty.

These results are confirmed by the simulations leading to Figure 4.2. This figure shows the performance regions for the same two-user global joint transmission scenario and channel realizations as in Figure 4.2 of Section 3.3. The Pareto boundaries are generated by combining Theorems 3.9 and 4.2 under spherical uncertainty sets $\mathcal{U}_k(\widehat{\mathbf{h}}_k, \mathbf{B}_k)$ with $\mathbf{B}_k = \sqrt{\xi}\mathbf{I}_N$ and different squared radius: $\xi \in \{0, 0.01, 0.05, 0.1\}$. All Pareto optimal points were achieved using single-stream beamforming. As expected, channel uncertainty reduces the size of the regions, but it can also affect the shape since the system becomes specifically sensitive to inter-user interference. This can be seen in Figure 4.2(b), where a nonconvex region is transformed into a concave region when the uncertainty is increased. This indicates that SDMA becomes less attractive as the CSI uncertainty grows.




(a) Channel Realization 1




(b) Channel Realization 2

Fig. 4.2 Performance regions for two different channel realizations (same as in Figure 3.1) under global joint transmission and joint spherical uncertainty regions with different radius $\sqrt{\xi}$. The operating points with different system utility functions are also indicated. Uncertainty reduces the area of the region and can also affect the shape; the small non-convexities in (a) become more visible and the clearly nonconvex region in (b) becomes increasingly concave-like.

To summarize, we can solve any robust multi-cell resource allocation problem as described in this subsection and we can expect a single-stream beamforming solution — which is a practically implementable solution. If we anyway would achieve a high-rank solution (although these have not been found in simulations), an approximate solution can be achieved by, for example, taking the dominating eigenvector of $\mathbf{S}_k$ as the beamforming direction $\bar{\mathbf{v}}_k$ or taking a realization from a zero-mean Gaussian distribution with the high-rank $\mathbf{S}_k$ as correlation matrix [166].

---

**Remark 4.1 (MSE-Based Single-Stream Beamforming).** This subsection has considered selection of transmit strategies to achieve robustness to channel uncertainty, but there are other system processes affected by uncertainty. In particular, measuring user performance as a function of the SINR implicitly assumes ideal receive processing. To relax this assumption, we consider a system with single-stream beamforming (for practical reasons). $\text{MS}_k$ processes the received signal $y_k = \mathbf{h}_k^H \mathbf{C}_k \sum_{i=1}^{K_r} \mathbf{D}_i \mathbf{v}_i s_i + n_k$ using an equalizing coefficient $\zeta_k$. The purpose is to achieve an estimate $\hat{s}_k = \zeta_k y_k$ of the transmitted signal $s_k$ that minimizes the mean square error $\text{MSE}_k = \mathbb{E}\{|\hat{s}_k - s_k|^2\}$.

By guaranteeing a certain $\text{MSE}_k$ under CSI uncertainty, we also guarantee that $\text{SINR}_k \geq \frac{1}{\text{MSE}_k} - 1$ [242], but equality can only be ensured under perfect CSI. The single-cell works in [242, 286] and multi-cell work in [26] measure the performance of $\text{MS}_k$ by a strictly monotonically *decreasing* function $\tilde{g}_k : \mathbb{R}_+ \to \mathbb{R}$ of $\text{MSE}_k$. We would like to minimize the MSEs of all users and the counterpart to the multi-objective optimization problem in (1.35) is

$$\underset{\mathbf{v}_k, \zeta_k \, \forall k}{\text{maximize}} \; \{\tilde{g}_1(\tilde{\gamma}_1), \ldots, \tilde{g}_{K_r}(\tilde{\gamma}_{K_r})\} \tag{4.10}$$

$$\text{subject to } \tilde{\gamma}_k \geq \min_{\mathbf{h}_k \in \mathcal{U}_k} \|\zeta_k \mathbf{h}_k^H \mathbf{C}_k \mathbf{V}_{\text{tot}} - \mathbf{e}_k^T\|_2^2 + |\zeta_k|^2 \sigma_k^2 \quad \forall k, \tag{4.11}$$

$$\sum_{k=1}^{K_r} \mathbf{v}_k^H \mathbf{Q}_{lk} \mathbf{v}_k \leq q_l \quad \forall l, \tag{4.12}$$

where $\tilde{\gamma}_k$ denotes $\text{MSE}_k$ and $\mathbf{V}_{\text{tot}} = [\mathbf{D}_1 \mathbf{v}_1 \ldots \mathbf{D}_{K_r} \mathbf{v}_{K_r}]$ and $\mathbf{e}_k$ denotes the $k$th column of $\mathbf{I}_{K_r}$. The corresponding robust performance region

is compact and normal (see [26, Lemma 1]) and the robust MSE constraint (4.11) can be reformulated as

$$\begin{bmatrix} \sqrt{\tilde{\gamma}_k}\tilde{\zeta}_k - \lambda_k & \widehat{\mathbf{h}}_k^H \mathbf{C}_k \bar{\mathbf{V}} - \tilde{\zeta}_k \mathbf{e}_k^T & \sigma_k & \mathbf{0} \\ \bar{\mathbf{V}}^H \mathbf{C}_k^H \widehat{\mathbf{h}}_k - \tilde{\zeta}_k \mathbf{e}_k & \sqrt{\tilde{\gamma}_k}\tilde{\zeta}_k \mathbf{I}_{K_r} & \mathbf{0} & -\bar{\mathbf{V}}^H \mathbf{C}_k^H \mathbf{B}_k \\ \sigma_k & \mathbf{0} & \sqrt{\tilde{\gamma}_k}\tilde{\zeta}_k & \mathbf{0} \\ \mathbf{0} & -\mathbf{B}_k^H \mathbf{C}_k \bar{\mathbf{V}} & \mathbf{0} & \lambda_k \mathbf{I}_N \end{bmatrix} \succeq \mathbf{0}_{N+K_r+2}, \tag{4.13}$$

where $\lambda_k \geq 0$ is an auxiliary variable and $\tilde{\zeta}_k = \zeta_k^{-1}$ can be taken as positive (because any complex phase can be included in the beamforming vector $\mathbf{v}_k$ without affecting the feasibility of (4.13)). The reformulation of (4.11) into (4.13) is based on an extension of the S-procedure from [66, Proposition 2]. Observe that (4.13) is convex in $\tilde{\zeta}_k, \lambda_k, \mathbf{v}_i \, \forall i$ (see [26, 66, 242, 286] for details). Consequently, many of the convexity results in Section 2 can be extended with only a minor increase in computational complexity (i.e., semi-definite constraints as (4.13) are more demanding than the second-order cone constraints under perfect CSI [10]). For example, [26] builds a resource allocation framework where weighted max-min fairness is shown to be a quasi-convex problem also under worst-case robustness and the BRB algorithm is used to solve any robust monotonic problem.

### 4.1.2 Worst-Case Robustness Under Separate Uncertainty

The scenario when the uncertainty in $\mathbf{h}_k$ is modeled separately for each channel component $\mathbf{h}_{jk}$, $j = 1, \ldots, K_t$, is more analytically involved. The mutual uncertainty in $\mathbf{h}_k$ can be viewed as the intersection of (general-type) ellipsoids, which generally leads to problems which are NP-hard [9]. The hardness can however be avoided if we limit the scope to problems where the impact of each channel component can also be separated. Coordinated beamforming is such an example, because the transmitted signals over each channel component are independent. This is also fulfilled when multiple base stations send multiple independent signals to a given user [180, 239], but for simplicity this subsection concentrates on coordinated beamforming.

We let $j_k$ denote the index of the base station that serves $\text{MS}_k$ and consider a signal correlation matrix $\mathbf{S}_{j_k k} \in \mathbb{C}^{N_{j_k} \times N_{j_k}}$ of arbitrary rank. The SOP with QoS requirements $g_k(\text{SINR}_k) \geq r_k^*$ in (2.29) becomes

$$\begin{aligned} &\text{find } \mathbf{S}_{j_k k} \succeq \mathbf{0}_{N_{j_k}} \quad \forall k \\ &\text{subject to } \frac{\mathbf{h}_{j_k k}^H \mathbf{C}_{j_k k} \mathbf{S}_{j_k k} \mathbf{C}_{j_k k}^H \mathbf{h}_{j_k k}}{\sigma_k^2 + \sum_{i \neq k} \mathbf{h}_{j_i k}^H \mathbf{C}_{j_i k} \mathbf{S}_{j_i i} \mathbf{C}_{j_i k}^H \mathbf{h}_{j_i k}} \geq g_k^{-1}(r_k^*) \quad \forall \mathbf{h}_{jk} \in \mathcal{U}_{jk}, \forall j, k, \\ &\sum_{k=1}^{K_r} \text{tr}(\mathbf{Q}_{l j_k k} \mathbf{S}_{j_k k}) \leq q_l \quad \forall l \end{aligned} \tag{4.14}$$

under coordinated beamforming with the separate ellipsoidal channel uncertainty sets in (4.3). A main difference from the coordinated beamforming considered in Section 3.4 is that we now have general power constraints, defined by $\mathbf{Q}_{l j_k k} \succeq \mathbf{0}_{N_{j_k}}$. The S-procedure in Lemma 4.1 can be used to rewrite the infinitely many SINR constraints in (4.14) as a finite number of constraints — once again at the expense of adding auxiliary variables.

---

**Theorem 4.3.** Let $\mathbf{S}_{j_1 1}, \ldots, \mathbf{S}_{j_{K_r} K_r}$ be a transmit strategy and let $\gamma_k \geq 0$ be given, then the robust SINR constraint

$$\frac{\mathbf{h}_{j_k k}^H \mathbf{C}_{j_k k} \mathbf{S}_{j_k k} \mathbf{C}_{j_k k}^H \mathbf{h}_{j_k k}}{\sigma_k^2 + \sum_{i \neq k} \mathbf{h}_{j_i k}^H \mathbf{C}_{j_i k} \mathbf{S}_{j_i i} \mathbf{C}_{j_i k}^H \mathbf{h}_{j_i k}} \geq \gamma_k \quad \forall \mathbf{h}_{jk} \in \mathcal{U}_{jk}(\widehat{\mathbf{h}}_{jk}, \mathbf{B}_{jk}), \forall j \tag{4.15}$$

is fulfilled if and only if it exists $\lambda_{jk} \geq 0 \forall j$ and $\vartheta_{jk} \geq 0 \forall j \neq j_k$ such that

$$\begin{bmatrix} \mathbf{B}_{jk}^H \mathbf{A}_{jk} \mathbf{B}_{jk} & \mathbf{B}_{jk}^H \mathbf{A}_{jk} \widehat{\mathbf{h}}_{jk} \\ \widehat{\mathbf{h}}_{jk}^H \mathbf{A}_{jk} \mathbf{B}_{jk} & \widehat{\mathbf{h}}_{jk}^H \mathbf{A}_{jk} \widehat{\mathbf{h}}_{jk} - a_{jk} \end{bmatrix} + \begin{bmatrix} \lambda_{jk} \mathbf{I}_{N_j} & \mathbf{0}_{N_j \times 1} \\ \mathbf{0}_{1 \times N_j} & -\lambda_{jk} \end{bmatrix} \succeq \mathbf{0}_{N_j + 1} \tag{4.16}$$

for $j = 1, \ldots, K_t$, where

$$\mathbf{A}_{jk} = \begin{cases} \mathbf{C}_{jk} \big( \frac{\gamma_k}{1+\gamma_k} \mathbf{S}_{jk} - \sum_{i \in \mathcal{D}_j} \mathbf{S}_{ji} \big) \mathbf{C}_{jk}^H, & j = j_k, \\ -\mathbf{C}_{jk} \big( \sum_{i \in \mathcal{D}_j} \mathbf{S}_{ji} \big) \mathbf{C}_{jk}^H, & j \neq j_k, \end{cases} \tag{4.17}$$

$$a_{jk} = \begin{cases} \sigma_k^2 + \sum_{m \neq j} \vartheta_{mk}, & j = j_k, \\ -\vartheta_{jk}, & j \neq j_k. \end{cases} \tag{4.18}$$

Furthermore, single-stream beamforming is sufficient when (a) each $\mathrm{BS}_j$ serves at most one user; or (b) each $\mathrm{BS}_j$ has perfect CSI to the users $k \in \mathcal{D}_j$ that it serves.

---

*Proof.* To separate the impact of the different uncertainty sets, we introduce the auxiliary variables $\vartheta_{jk}$ to split the SINR constraint into $\mathbf{h}_{j_k k}^H \mathbf{A}_{j_k k} \mathbf{h}_{j_k k} + \sigma_k^2 + \sum_{m \neq j_k} \vartheta_{mk} \geq 0$ and $\vartheta_{mk} + \mathbf{h}_{mk}^H \mathbf{A}_{mk} \mathbf{h}_{mk} \geq 0$ for $m \neq j_k$. The reformulation (4.16) then follows by applying Lemma 4.1 on each inequality (similar to what was done in Theorem 4.2).

The statement on sufficiency of single-stream beamforming is proved by analyzing the dual problem to (4.14) when the zero-valued cost function has been replaced with $\sum_{k=1}^{K_r} \mathrm{tr}(\mathbf{S}_{j_k k})$. The modified cost function will not affect the feasibility, but simplifies the analysis and explains why we prove the sufficiency of single-stream beamforming and not also necessity. The dual problem is

$$\begin{aligned} \underset{\boldsymbol{\Upsilon}_{jk}, \vartheta_{jk}, \lambda_{jk}, \mu_l \, \forall k,l}{\text{minimize}} \quad & \sum_{k=1}^{K_r} \sigma_k^2 [\boldsymbol{\Upsilon}_{jk}]_{N_j+1 \, N_j+1} - \sum_{l=1}^{L} \mu_l q_l \qquad (4.19) \\ \text{subject to} \quad & \boldsymbol{\Upsilon}_{jk} \succeq \mathbf{0}_{N_j+1}, \vartheta_{jk} \geq 0, \lambda_{jk} \geq 0, \mu_l \geq 0 \quad \forall j,k,l, \\ & \mathrm{tr}\left( \boldsymbol{\Upsilon}_{jk} \begin{bmatrix} \mathbf{I}_{N_j} & \mathbf{0}_{N_j \times 1} \\ \mathbf{0}_{1 \times N_j} & -1 \end{bmatrix} \right) \succeq \mathbf{0}_{N_j+1} \quad \forall j,k, \\ & \mathbf{Y}_k \succeq \mathbf{0}_{N_{j_k}+1} \quad \forall k, \end{aligned}$$

where $\tilde{\mathbf{B}}_{jk} = \mathbf{C}_{jk}^H [\mathbf{B}_{jk} \, \bar{\mathbf{h}}_{jk}]$ and

$$\mathbf{Y}_k = \left( \mathbf{I}_{N_j+1} + \sum_{l=1}^{L} \mu_l \mathbf{Q}_{l j_k k} + \sum_{i \neq k} \tilde{\mathbf{B}}_{j_k i} \boldsymbol{\Upsilon}_{j_k i} \tilde{\mathbf{B}}_{j_k i} - \frac{1}{\gamma_k} \tilde{\mathbf{B}}_{j_k k} \boldsymbol{\Upsilon}_{j_k k} \tilde{\mathbf{B}}_{j_k k} \right). \qquad (4.20)$$

One of the complementary slackness conditions is $\mathrm{tr}(\mathbf{S}_{j_k k} \mathbf{Y}_k) = 0$, which implies that $\mathbf{S}_{j_k k}$ should lie in the null space of $\mathbf{Y}_k$. If we can show that $\mathrm{rank}(\mathbf{Y}_k) \geq N_{j_k} - 1$, then it follows that $\mathrm{rank}(\mathbf{S}_{j_k k}) \leq 1$ [239, 251]. Observe that the first part of $\mathbf{Y}_k$ is positive definite while $-\frac{1}{\gamma_k} \tilde{\mathbf{B}}_{j_k k} \boldsymbol{\Upsilon}_{j_k k} \tilde{\mathbf{B}}_{j_k k}$ is negative semi-definite. We have $\tilde{\mathbf{B}}_{j_k k} = [\mathbf{0}_{N_{j_k}} \, \mathbf{C}_{j_k k}^H \bar{\mathbf{h}}_{j_k k}]$ in case (a), which has rank one. The

negative semi-definite part can then have at most be rank one and $\text{rank}(\mathbf{Y}_k) \geq N_{j_k} - 1$ follows. To prove case (b), we invoke the complementary slackness condition to (4.16),

$$\boldsymbol{\Upsilon}_{j_k k}\left(\frac{1}{\gamma_k}\tilde{\mathbf{B}}_{j_k k}^H \mathbf{S}_{j_k k}\tilde{\mathbf{B}}_{j_k k} + \begin{bmatrix}\lambda_{j_k k}\mathbf{I}_{N_{j_k}} & \mathbf{0}_{N_{j_k}\times 1} \\ \mathbf{0}_{1\times N_{j_k}} & -\lambda_{j_k k} - a_{j_k k}\end{bmatrix}\right) = \mathbf{0}_{N_{j_k}+1}, \tag{4.21}$$

where it must hold that $\lambda_{j_k k} > 0$ (to make (4.16) positive semi-definite). As the bracketed term have at least $N_{j_k}$ in rank, it follows that $\text{rank}(\boldsymbol{\Upsilon}_{j_k k}) \leq 1$ and thus $\text{rank}(\mathbf{Y}_k) \geq N_{j_k} - 1$ from (4.20). □

This theorem shows that (4.14) is a convex problem. It can therefore be used as a subproblem to solve weighted max-min fairness in polynomial time, for applying the PA and BRB algorithms on any robust resource allocation problem, and to parameterize the Pareto boundary of the robust performance region. The second part of Theorem 4.3 is a generalization of recent results in [239]. There is further evidence in [239] indicating that single-stream beamforming might always be optimal, just as for the case of perfect CSI.

The two special cases when single-stream beamforming is provably optimal corresponds to the MISO interference channel and to coordinated beamforming with perfect intra-cell CSI (which might be a reasonable model as inter-cell CSI is harder to obtain [17]). In these cases it is possible to optimize the beamforming vectors $\mathbf{v}_{j_k k}$ directly (instead of optimizing $\mathbf{S}_{j_k k}$ which basically has $N_{j_k}$ times more variables) and achieve convex problem formulations — which is generally not the case even when the final solution is known to be rank one. The following theorem bounds the worst-case performance in these cases, similar to [182].

---

**Theorem 4.4.** The signal term at $\text{MS}_k$ under single-stream beamforming and the separate uncertainty sets in (4.3) satisfies

$$x_{kk}(\mathbf{v}_{j_k k}) \geq \left(\left[|\hat{\mathbf{h}}_{j_k k}^H \mathbf{C}_{j_k k}\mathbf{v}_{j_k k}| - \|\mathbf{B}_{j_k k}^H \mathbf{C}_{j_k k}\mathbf{v}_{j_k k}\|_2\right]_+\right)^2 \tag{4.22}$$

while the interference term from transmission to $\text{MS}_i$, $i \neq k$, satisfies

$$x_{ik}(\mathbf{v}_{j_i i}) \leq \left(|\hat{\mathbf{h}}_{j_i k}^H \mathbf{C}_{j_i k}\mathbf{v}_{j_i i}| + \|\mathbf{B}_{j_i k}\mathbf{C}_{j_i k}\mathbf{v}_{j_i i}\|_2\right)^2. \tag{4.23}$$

The resulting worst-case achievable SINR at $\mathrm{MS}_k$ is

$$\mathrm{SINR}_k \geq \frac{\left(\left[|\hat{\mathbf{h}}_{j_k k}^H \mathbf{C}_{j_k k}\mathbf{v}_{j_k k}| - \|\mathbf{B}_{j_k k}^H \mathbf{C}_{j_k k}\mathbf{v}_{j_k k}\|_2\right]_+\right)^2}{\sigma_k^2 + \sum_{i\neq k}\left(|\hat{\mathbf{h}}_{j_i k}^H \mathbf{C}_{j_i k}\mathbf{v}_{j_i i}| + \|\mathbf{B}_{j_i k}\mathbf{C}_{j_i k}\mathbf{v}_{j_i i}\|_2\right)^2}. \quad (4.24)$$

All inequalities hold with equality for the MISO interference channel.

*Proof.* The bounds in (4.22) and (4.23) are achieved by treating each term in $\mathrm{SINR}_k$ separately. The worst-case signal term satisfies

$$\begin{aligned}
\sqrt{x_{kk}(\mathbf{v}_{j_k k})} &= \min_{\mathbf{h}_{j_k k}\in\mathcal{U}_{j_k k}(\widehat{\mathbf{h}}_{j_k k},\mathbf{B}_{j_k k})} |\mathbf{h}_{j_k k}^H \mathbf{C}_{j_k k}\mathbf{v}_{j_k k}| \\
&= \min_{\boldsymbol{\epsilon}_{j_k k}:\|\boldsymbol{\epsilon}_{j_k k}\|_2\leq 1} |\hat{\mathbf{h}}_{j_k k}^H \mathbf{C}_{j_k k}\mathbf{v}_{j_k k} + \boldsymbol{\epsilon}_{j_k k}^H \mathbf{B}_{j_k k}^H \mathbf{C}_{j_k k}\mathbf{v}_{j_k k}| \\
&\geq \min_{\boldsymbol{\epsilon}_{j_k k}:\|\boldsymbol{\epsilon}_{j_k k}\|_2\leq 1} \left[|\hat{\mathbf{h}}_{j_k k}^H \mathbf{C}_{j_k k}\mathbf{v}_{j_k k}| - |\boldsymbol{\epsilon}_{j_k k}^H \mathbf{B}_{j_k k}^H \mathbf{C}_{j_k k}\mathbf{v}_{j_k k}|\right]_+ \\
&= \left(|\hat{\mathbf{h}}_{j_k k}^H \mathbf{C}_{j_k k}\mathbf{v}_{j_k k}| - \|\mathbf{B}_{j_k k}^H \mathbf{C}_{j_k k}\mathbf{v}_{j_k k}\|_2\right)^+, \qquad (4.25)
\end{aligned}$$

where the inequality follows from the triangle inequality and equality is achieved by $\boldsymbol{\epsilon}_{j_k k} = -\frac{\mathbf{B}_{j_k k}^H \mathbf{C}_{j_k k}\mathbf{v}_{j_k k}}{\|\mathbf{B}_{j_k k}^H \mathbf{C}_{j_k k}\mathbf{v}_{j_k k}\|_2} e^{\imath\angle(\hat{\mathbf{h}}_{j_k k}^H \mathbf{C}_{j_k k}\mathbf{v}_{j_k k})}$. The worst-case interference term is computed similarly as

$$\begin{aligned}
\sqrt{x_{ik}(\mathbf{v}_{j_i i})} &= \max_{\mathbf{h}_{j_i k}\in\mathcal{U}_{j_i k}(\widehat{\mathbf{h}}_{j_i k},\mathbf{B}_{j_i k})} |\mathbf{h}_{j_i k}^H \mathbf{C}_{j_i k}\mathbf{v}_{j_i i}| \\
&= \max_{\boldsymbol{\epsilon}_{j_i k}:\|\boldsymbol{\epsilon}_{j_i k}\|_2\leq 1} |\hat{\mathbf{h}}_{j_i k}^H \mathbf{C}_{j_i k}\mathbf{v}_{j_i i} + \boldsymbol{\epsilon}_{j_i k}^H \mathbf{B}_{j_i k}^H \mathbf{C}_{j_i k}\mathbf{v}_{j_i i}| \\
&\leq \max_{\boldsymbol{\epsilon}_{j_i k}:\|\boldsymbol{\epsilon}_{j_i k}\|_2\leq 1} |\hat{\mathbf{h}}_{j_i k}^H \mathbf{C}_{j_i k}\mathbf{v}_{j_i i}| + |\boldsymbol{\epsilon}_{j_i k}^H \mathbf{B}_{j_i k}^H \mathbf{C}_{j_i k}\mathbf{v}_{j_i i}| \\
&= |\hat{\mathbf{h}}_{j_i k}^H \mathbf{C}_{j_i k}\mathbf{v}_{j_i i}| + \|\mathbf{B}_{j_i k}^H \mathbf{C}_{j_i k}\mathbf{v}_{j_i i}\|, \qquad (4.26)
\end{aligned}$$

where the inequality follows again from the triangle inequality and equality is achieved by $\boldsymbol{\epsilon}_{j_i k} = +\frac{\mathbf{B}_{j_i k}^H \mathbf{C}_{j_i k}\mathbf{v}_{j_i i}}{\|\mathbf{B}_{j_i k}^H \mathbf{C}_{j_i k}\mathbf{v}_{j_i i}\|} e^{\imath\angle(\hat{\mathbf{h}}_{j_i k}^H \mathbf{C}_{j_i k}\mathbf{v}_{j_i})}$. The inequality for the SINR follows since $\mathrm{SINR}_k = \frac{x_{kk}(\mathbf{v}_{j_k k})}{\sigma_k^2 + \sum_{i\neq k} x_{ik}(\mathbf{v}_{j_i i})}$.

Finally, all bounds are simultaneously achievable if $\mathbf{h}_{jk}$ only appears once in $\mathrm{SINR}_k$ for each $j$, which happens for the MISO interference channel where each transmitter only sends one signal. □

This theorem obtains a lower bound on the worst-case SINR under ellipsoidal channel uncertainty, and the bound is tight for interference channels. The SINR expression in (4.24) is attractive since the signal and interference terms both have the structure of second-order cones (see Example 2.2), thus we can parameterize the Pareto boundary of the robust performance region as follows.

---

**Theorem 4.5.** For a MISO interference channel with per-transmitter constraints of $q_j$, all Pareto optimal points on the corresponding robust performance region are achieved by beamforming vectors $\mathbf{v}_{j_k k}(\boldsymbol{\lambda}_k)$ for $\boldsymbol{\lambda}_k \in [0,1]^{K_r - 1}$ for $k = 1, \ldots, K_r$, which solves

$$\begin{aligned} \mathbf{v}_{j_k k}(\boldsymbol{\lambda}_k) = \underset{\mathbf{v}_{j_k k}}{\arg\max}\ & \Re(\hat{\mathbf{h}}_{j_k k}^H \mathbf{C}_{j_k k} \mathbf{v}_{j_k k}) - \|\mathbf{B}_{j_k k}^H \mathbf{C}_{j_k k} \mathbf{v}_{j_k k}\| \\ & \text{subject to} \quad \Im\left(\hat{\mathbf{h}}_{j_k k}^H \mathbf{C}_{j_k k} \mathbf{v}_{j_k k}\right) = 0, \\ & \|\mathbf{v}_{j_k k}\| \leq q_{j_k}, \\ & |\hat{\mathbf{h}}_{j_k i}^H \mathbf{C}_{j_k i} \mathbf{v}_{j_k k}| + \|\mathbf{B}_{j_k i} \mathbf{C}_{j_k i} \mathbf{v}_{j_k k}\|_2 \leq \sqrt{\lambda_{ki} \Gamma_{ki}} \quad \forall i \neq k, \end{aligned} \tag{4.27}$$

for fixed values of $\Gamma_{ki}$. The elements in $\boldsymbol{\lambda}_k$ are denoted $\lambda_{ki}$ for all $i \neq k$.

---

*Proof.* The proof works by contradiction and is analogue to the proof detailed in [182, Theorem 1]. □

This parametrization can be seen as a robust counterpart to Theorem 3.2 and the values $\Gamma_{ki}$ correspond to the interference temperature limits, which are applied in underlay and overlay cognitive radio systems (see Section 4.8). Recall that second-order cone optimization problems [160], such as (4.27), are efficiently solved by interior-point methods (e.g., using `SeDuMi` [256]).

When the intra-cell channels are perfectly known (i.e., $\mathbf{B}_{j_k k} = \mathbf{0}_{N_{j_k}} \forall k$), we can utilize the sufficiency of single-stream beamforming to reduce the computational complexity in Theorem 4.3. The following convex problem formulation was obtained in [17].

**Corollary 4.6.** If $\mathbf{B}_{j_k k} = \mathbf{0}_{N_{j_k}}\ \forall k$, the problem in (4.14) can be reformulated into the convex feasibility problem

$$\begin{aligned} &\text{find } \mathbf{S}_{j_k k} \succeq \mathbf{0}_{N_{j_k}} \quad \forall k \\ &\text{subject to } \sum_{k=1}^{K_r} \mathbf{v}_{j_k k}^H \mathbf{Q}_{l j_k k} \mathbf{v}_{j_k k} \leq q_l \quad \forall l \\ &\qquad \begin{bmatrix} \vartheta_{mk} - \lambda_{mk} & \widehat{\mathbf{h}}_{mk}^H \mathbf{V}_m & \mathbf{0} \\ \mathbf{V}_m^H \widehat{\mathbf{h}}_{mk} & \vartheta_{mk} \mathbf{I}_{|\mathcal{D}_m|} & \mathbf{V}_m^H \mathbf{B}_{mk} \\ \mathbf{0} & \mathbf{B}_{mk}^H \mathbf{V}_m & \lambda_{mk} \mathbf{I}_{N_m} \end{bmatrix} \succeq \mathbf{0} \quad m \neq j_k \end{aligned} \tag{4.28}$$

$$\sqrt{1 + \frac{1}{\gamma_k}} \mathbf{h}_{j_k k}^H \mathbf{C}_{j_k k} \mathbf{v}_{j_k k} \geq \sqrt{\|\mathbf{h}_{j_k k}^H \mathbf{C}_{j_k k} \mathbf{V}_{j_k}\|_2^2 + \sum_{m \neq j_k} \vartheta_{mk}^2 + \sigma_k^2}.$$

where $\mathbf{V}_j = [\mathbf{v}_{j\mathcal{D}_j(1)} \ldots \mathbf{v}_{j\mathcal{D}_j(|\mathcal{D}_j|)}]$ contains the beamforming vectors of users served by $\text{BS}_j$.

*Proof.* The proof is similar to that of Theorem 4.3 but utilizes an extension of the S-procedure that handles beamforming vectors; see [66, Proposition 2]. The intra-cell constraint can be formulated as a second-order cone and further details are available in [17]. □

We conclude this section by illustrating that Theorem 4.3 and Corollary 4.6 provide a way to solve any robust resource allocation problem under coordinated beamforming. We consider $K_t = 2$ base stations with $N_j = 4$ antennas and two users per cell (i.e., $K_r = 4$). We generate the exact intra-cell channel as $\mathbf{h}_{jk} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{N_j})$ for $k \in \mathcal{D}_j$ and the uncertain inter-cell channels as $\widehat{\mathbf{h}}_{ji} \sim \mathcal{CN}(\mathbf{0}, \frac{1}{2}\mathbf{I}_{N_j})$ for $i \notin \mathcal{D}_j$. Spherical uncertainty sets are assumed with $\mathbf{B}_{ji} = \sqrt{\xi}\mathbf{I}_{N_j}$, where $\sqrt{\xi}$ is the radius. The optimal worst-case sum information rate is obtained using the BRB algorithm and is shown in Figure 4.3. We also show the performance of SLNR-MAX beamforming with heuristic power allocation based on Theorem 3.16.[3] The simulation shows that the heuristic strategy is

[3] The worst-case performance of a heuristic transmit strategy is computed by solving one optimization problem per user: Maximize $\gamma_k$ subject to (4.16) for $j = 1, \ldots, K_t$. This is a convex problem when the transmit strategy is fixed and only $\lambda_{jk} \geq 0\ \forall j$ and $\vartheta_{jk} \geq 0\ \forall j \neq j_k$ are considered to be optimization variables.

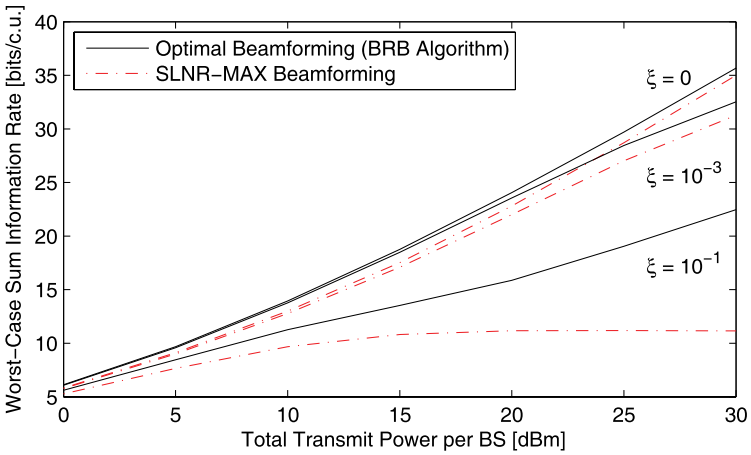



Fig. 4.3 Average worst-case sum information rate as a function of the total transmit power (per base station). The optimal beamforming is computed using the BRB algorithm. The heuristic transmit strategy is robust to small inter-cell CSI errors, but becomes highly suboptimal as the uncertainty grows.

relatively robust to small errors (e.g., $\xi = 10^{-3}$) but becomes highly suboptimal as the CSI uncertainty increases.

## 4.2 Distributed Resource Allocation

Several seemingly nonconvex resource allocation problems were reformulated as convex problems in Section 2.2, thus showing that these can be solved in polynomial time using numerical algorithms such as interior-point methods [37]. The discussions stopped when the problems were identified as convex, but it is important to also design the optimization functionality of the system: where are the different pieces of information available and where should the numerical computations be performed? Ideally, the optimal resource allocation is computed at a central control station with aggregate CSI knowledge from the whole system, but this is practically infeasible in terms of computational complexity, backhaul signaling, delays, and scalability [21]. The decomposability of a resource allocation problem is therefore an important feasibility factor, in addition to the convexity [194]. Removing the reliance on a central control station also provides resilience against various hardware failures that can occur in the network.

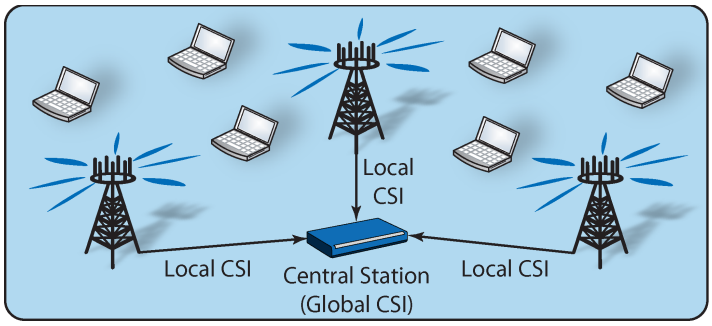
This section outlines two different approaches to decentralized resource allocation, where the computational load is distributed over the system and the exchange of CSI and control signals is limited. The first approach solves convex optimization problems in a distributed fashion using only local CSI, but requires iterative exchange of control variables. The second approach is truly distributed in the sense that each base station selects its beamforming vectors in a noniterative manner without exchanging any information with the other base stations. In both cases, the local CSI at $\mathrm{BS}_j$ consists of the channel vectors $\mathbf{h}_{jk}$ from the own base station to all users $k \in \mathcal{C}_j$. Observe that these channel vectors can be estimated locally at $\mathrm{BS}_j$ by utilizing channel reciprocity in TDD systems. The distinction between the two distributed approaches and global resource allocation is illustrated in Figure 4.4.

There are certainly other distributed resource allocation approaches; for example, [61, 62, 312] where the base stations make distributed decisions based on different estimates of the global CSI. Capacity results under different backhaul models are surveyed in [81]. Furthermore, [207] presents two iterative distributed algorithms where the subproblem for $\mathrm{MS}_k$ is only based on perfect knowledge of $\mathbf{C}_k\mathbf{h}_k$.

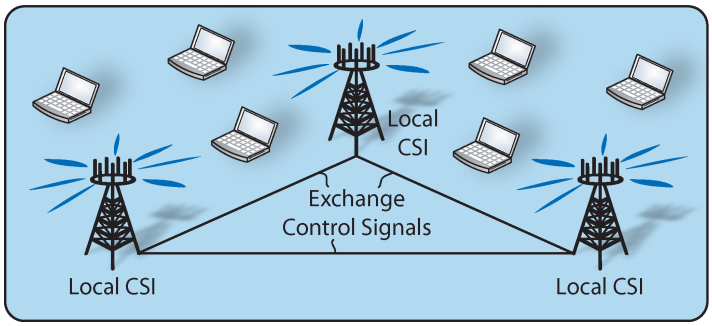
### 4.2.1 Distributed Implementation of Convex Optimization

The very essence of resource allocation problems is the coupling between the users, in terms of inter-user interference and power constraints. A tutorial on decomposition methods that relax the coupling is provided in [194]. In our area, these methods can decompose the original centralized optimization problem into a sequence of distributed subproblems only requiring local CSI and not any user involvement. The beamforming for each user is optimized separately and sequentially. If multiple base stations serve a given user, then they appoint a *master base station* (MBS). The MBS gathers the relevant CSI from other base stations and takes care of all computations related to the given user in the optimization.

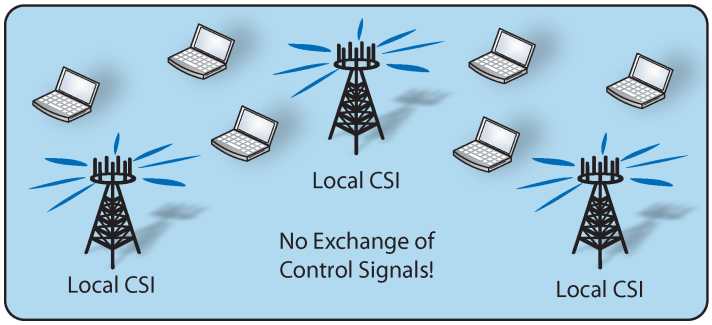
To exemplify the decomposability, we will show how the resource allocation with QoS requirements and the quasi-convex curve-search




(a) Centralized Optimization and Resource Allocation




(b) Distributed Allocation with Control Signaling (Subsection 4.2.1)




(c) Truly Distributed Allocation (Subsection 4.2.2)

Fig. 4.4 Different implementations of resource allocation in multi-cell systems: (a) Global CSI is gathered at a central station that allocates resources; (b) Base stations perform distributed resource allocation by iteratively exchanging control variables (but not CSI); and (c) Base stations perform distributed resource allocation without exchanging any information (but the problem formulation and local CSI is available).

procedure in Section 2.2 can be solved to global optimality in a distributed manner. The former problem is considered in [204, 257, 265] under total power minimization, while the latter is considered in [14, 239, 257].

As a first step, the feasibility problem with QoS requirements in (2.29) (with $\mathrm{SINR}_k \geq \gamma_k \forall k$) is rewritten as

$$\begin{aligned} &\text{find } \mathbf{v}_k, \Theta_{ik}, \widetilde{\Theta}_{ik}, q_{lk} \; \forall k,i,l, \quad i \neq k \\ &\text{subject to } \frac{|\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_k \mathbf{v}_k|^2}{\sigma_k^2 + \sum\limits_{i \neq k} \widetilde{\Theta}_{ik}^2} \geq \gamma_k \quad \forall k, \\ &\qquad |\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_i \mathbf{v}_i|^2 \leq \Theta_{ik}^2, \quad \Theta_{ik} \leq \widetilde{\Theta}_{ik} \quad \forall k,i, \; i \neq k, \\ &\qquad \mathbf{v}_k^H \mathbf{Q}_{lk} \mathbf{v}_k \leq q_{lk}, \quad \sum_{i=1}^{K_r} q_{li} \leq q_l \quad \forall l,k \end{aligned} \tag{4.29}$$

by adding nonnegative auxiliary variables $\Theta_{ik}, \widetilde{\Theta}_{ik}, q_{lk}$. The squared variable $\Theta_{ik}^2$ is the actual interference generated at $\mathrm{MS}_k$ by signals intended for $\mathrm{MS}_i$, while $\widetilde{\Theta}_{ik}^2$ is its believed value in the beamforming optimization for $\mathrm{MS}_k$. The reason for defining these variables as the square roots of the interference is to enable the SINR constraints to be expressed as second-order cones $\Re(\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_k \mathbf{v}_k) \geq \sqrt{\gamma_k} \sqrt{\sigma_k^2 + \sum_{i \neq k} \widetilde{\Theta}_{ik}^2}$. Similarly, the power constraints are separated into per-user constraints using the variables $q_{lk}$.

Looking at (4.29), we observe that the transmission to different users is only coupled by the so-called consistency constraints $\Theta_{ik} \leq \widetilde{\Theta}_{ik}$ and $\sum_{i=1}^{K_r} q_{li} \leq q_l$. If the coupling variables $\Theta_{ik}, \widetilde{\Theta}_{ik}, q_{lk}$ are fixed, the beamforming optimization for the different users would decouple. The classic decomposition methods basically pretend that these variables are constants and update them iteratively [194, 204, 265].

We will take a dual decomposition approach where the coupling is relaxed by forming a partial Lagrangian for the consistency constraints. If the Lagrange multipliers are denoted $y_{ik}$ and $z_l$ for the interference and power consistency constraints, respectively, (4.29) can

be decomposed into $K_r$ subproblems where the problem for $\mathrm{MS}_k$ is

$$\begin{aligned} \underset{\mathbf{v}_k,\{\Theta_{ki}\}_{\forall i},\{\widetilde{\Theta}_{ik}\}_{\forall i},\{q_{lk}\}_{\forall l}}{\text{minimize}} \quad & \sum_{i\neq k}\left(y_{ki}\Theta_{ki} - y_{ik}\widetilde{\Theta}_{ik}\right) + \sum_{l=1}^{L} z_l q_{lk} \\ \text{subject to} \quad & \frac{|\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_k \mathbf{v}_k|^2}{\sigma_k^2 + \sum_{i\neq k}\widetilde{\Theta}_{ik}^2} \geq \gamma_k \quad \forall k, \\ & \mathbf{v}_k^H \mathbf{Q}_{lk} \mathbf{v}_k \leq q_{lk}, \quad q_{lk} \leq q_l \quad \forall l, \\ & |\mathbf{h}_i^H \mathbf{C}_i \mathbf{D}_k \mathbf{v}_k|^2 \leq \Theta_{ki}^2 \quad i \neq k \end{aligned} \tag{4.30}$$

and a master dual problem

$$\underset{\{y_{ik}\}_{\forall k,i},\{z_l\}_{\forall l}}{\text{maximize}} \sum_{k=1}^{K_r}\sum_{i\neq k} y_{ik}\left(\Theta_{ik}^* - \widetilde{\Theta}_{ik}^*\right) + \sum_{l=1}^{L} z_l \left(\sum_{k=1}^{K_r} q_{lk}^* - q_l\right), \tag{4.31}$$

where $\Theta_{ik}^*, \widetilde{\Theta}_{ik}^*, q_{lk}^*, \mathbf{v}_k^*$ are the subproblem solutions.

This decomposition enables an iterative procedure where the subproblems in (4.30) are solved for constant values on the Lagrange multipliers $y_{ki}, y_{ik}\ \forall i \neq k$ and $z_l\ \forall l$. This is called *dual* decomposition because the Lagrange multipliers are viewed as constants in the subproblems, and not the coupling variables themselves. The Lagrange multipliers can be viewed as prices for causing interference and for consuming transmit power, and these prices are iteratively adjusted by the master problem (4.31) until convergence. The update procedure requires backhaul signaling, but we will see that it can be implemented by distributed message passing between the involved transmitters. In other words, the heavy CSI signaling required to solve the resource allocation problem centrally is replaced by iterative interference and power control signaling. This confirms the observation in Section 3.2 that coordinated decision making is the limiting factor in multi-cell resource allocation, and not the localness of the CSI at the base stations.

The subproblems in (4.30) resembles the interference-constrained beamforming in Subsection 2.2.1 (with interference limits $\Gamma_{ik} = \Theta_{ik}^2$), with the difference that also the power constraints are decoupled. The problem (4.30) is convex and can be solved to global optimality using standard techniques. The master problem has a more complicated

structure and is typically solved by subgradient methods [194]. This implies that the Lagrange multipliers in iteration $n+1$ are achieved as

$$y_{ik}^{(n+1)} = \left[ y_{ik}^{(n)} - \xi(\widetilde{\Theta}_{ik}^{(n)} - \Theta_{ik}^{(n)}) \right]_+, \tag{4.32}$$

$$z_{l}^{(n+1)} = \left[ z_{l}^{(n)} - \xi\big(q_l - \sum_{k=1}^{K_r} q_{lk}^{(n)}\big) \right]_+, \tag{4.33}$$

where $\xi > 0$ is the step size. The update (4.32) requires exchange of $\widetilde{\Theta}_{ik}^{(n)}$ and $\Theta_{ik}^{(n)}$ (i.e., the Lagrange multipliers computed in the $n$th iteration) between the master base stations of $\mathrm{MS}_k$ and $\mathrm{MS}_i$.[4] The update (4.33) requires exchange of $q_{lk}^{(n)}$ between base stations that share the $l$th power constraint. The backhaul signaling load is quantified in [265], where several variations are discussed.

If the original problem (4.29) is feasible and the step size diminishes with $n$, iterating between the master problem and subproblems will eventually provide the globally optimal solution [194].[5] The problem is solved if all consistency constraints are satisfied as $\Theta_{ik}^{(n)} \leq \widetilde{\Theta}_{ik}^{(n)} + \varepsilon$ and $\sum_{k=1}^{K_r} q_{lk}^{(n)} \leq q_l + \varepsilon$, for some predefined accuracy $\varepsilon > 0$. The distributed algorithm is summarized in Algorithm 4. The stopping criterion can limit the number of iterations or detect if the original problem seems infeasible, but some central entity might need to enforce it.

Algorithm 4 includes some optional steps where the beamforming vectors are rescaled to satisfy all power constraints with equality and the corresponding user performance is evaluated. This improves the convergence by making the current beamforming vectors feasible, but at the expense of exchanging all power allocation coefficients $q_{lk}^{(n)}$.

Since Algorithm 4 provides a distributed solution to resource allocation with fixed QoS requirements, it can also be used as a subproblem in algorithms that successively increase the QoS requirements for the purpose of obtaining a Pareto optimal point. We will exemplify how the curve-search procedure in (2.40) of Subsection 2.2.3 can be solved

[4] If a base station is responsible for multiple users, their subproblems can be solved jointly and there is no need to introduce any coupling variables between these users.

[5] This convergence to the optimal solution only holds for convex problems. If an algorithm that solves a nonconvex problem is decomposed, we might not converge to the global optimum (and not even converge to something at all).

**Algorithm 4**: Distributed Optimization with QoS Requirements

**Result**: Distributed algorithm for solving (4.29).
**Input**: QoS requirements $g_k(\mathrm{SINR}_k) \geq r_k^*$ for each user $k$;
**Input**: Step-size $\xi > 0$ (fixed or adaptive);
**Input**: Stopping criterion and accuracy $\varepsilon > 0$;
**Input**: Initialization of $y_{ik}^{(1)}, z_l^{(1)}$ (e.g., equal to zero);
Set $n = 0$ and $\gamma_k = g_k^{-1}(r_k^*)$ $\forall k$;
**while** *stopping criterion is not satisfied* **do**
  Set $n = n + 1$;
  Solve subproblem (4.30) (using $y_{ik}^{(n)}, z_l^{(n)}$) at the master base station of each user $k$. Save current $\Theta_{ik}^{(n)}, \widetilde{\Theta}_{ik}^{(n)}, q_{lk}^{(n)}, \mathbf{v}_k^{(n)}$;
  Exchange relevant solution variables $\Theta_{ik}^{(n)}, \widetilde{\Theta}_{ik}^{(n)}, q_{lk}^{(n)}$ between base stations coupled by consistency constraints;
  (Optional:) Compute $\varsigma = \max_{\{l: q_l > 0\}} \sum_{k=1}^{K_r} \frac{q_{lk}^{(n)}}{q_l}$;
  (Optional:) Set $\mathbf{v}_k^{(n)} = \frac{1}{\sqrt{\varsigma}} \mathbf{v}_k^{(n)}$, $\Theta_{ik}^{(n)} = \frac{1}{\sqrt{\varsigma}} \Theta_{ik}^{(n)}$, $\widetilde{\Theta}_{ik}^{(n)} = \frac{1}{\sqrt{\varsigma}} \widetilde{\Theta}_{ik}^{(n)}$, $q_{lk}^{(n)} = \frac{1}{\sqrt{\varsigma}} q_{lk}^{(n)}$;
  Compute $y_{ik}^{(n+1)}, z_l^{(n+1)}$ at relevant places, using (4.32)-(4.33);
  Check if consistency constraints are satisfied (to accuracy $\varepsilon$);
  **if** $\varsigma \leq 1$ *and all consistency constraints are satisfied* **then**
    Problem (4.29) has been solved and the algorithm ends;
  (Optional:) Compute $\mathrm{SINR}_k^{(n)} = \frac{|\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_k \mathbf{v}_k^{(n)}|^2}{\sigma_k^2 + \sum_{i \neq k} (\Theta_{ik}^{(n)})^2}$;
  **if** *(Optional:)* $g_k(\mathrm{SINR}_k^{(n)}) \geq \gamma_k$ *for all* $k$ **then**
    Problem (4.29) has been solved and the algorithm ends;

**Output**: Optimal beamforming vectors $\mathbf{v}_k^{(n)}$;

distributively using Algorithm 4. Recall that this problem finds the best feasible point on a strictly increasing curve $\mathbf{r}(\tau)$. The centralized approach in Algorithm 1 solves the curve-search by bisection, but the dual decomposition approach is relatively slow at declaring that an operating point is infeasible. The distributed Algorithm 5 therefore starts at the first point, $\mathbf{r}(\tau^{\mathrm{lower}})$, and moves step-by-step along the

**Algorithm 5**: Distributed Curve-Search Procedure

**Result**: Distributed solution to optimization problem (2.40).
**Input**: Lower bound $\tau^{\text{lower}}$ on $\tau$ that guarantees feasibility;
**Input**: Step size $\tau^{(\text{step})} > 0$ of curve search;
**Input**: Step size $\xi > 0$ of subproblems;
**Input**: Subproblem accuracy $\varepsilon > 0$ and stopping criterion;
Set $y_{ik}^{(1)} = 0$, $z_l^{(1)} = 0$ $\forall k, i, l$ $i \neq k$;
Set $\tau^{(0)} = \tau^{\text{lower}}$ and $m = 0$;
**while** *stopping criterion is not satisfied* **do**
  Set $m = m + 1$;
  Set $\tau^{(m)} = \tau^{(m-1)} + \tau^{(\text{step})}$ and $r_k^* = r_k(\tau^{(m)})$ $\forall k$;
  Run Algorithm 4 using $\{r_k^*, y_{ik}^{(m)}, z_l^{(m)}, \xi, \varepsilon\}$;
  **if** *Algorithm 4 solves the problem* **then**
    Store variables $\mathbf{v}_k^{(m)}, y_{ik}^{(m+1)}, z_l^{(m+1)}$ in solution;
    (Optional:) Compute $\text{SINR}_k$ as in Step 12 of Algorithm 4;
    (Optional:) Find minimal $\tilde{\tau}$ with $r_k(\tilde{\tau}) \geq g_k(\text{SINR}_k)$ $\forall k$;
    (Optional:) Set $\tau^{(m)} = \tilde{\tau}$;
  **else**
    Decrement $m = m - 1$ and stop the algorithm;
**Output**: Operating point $\mathbf{r}(\tau^{(m)})$ and beamforming $\mathbf{v}_k^{(m)}$;

curve using a step size of $\tau^{(\text{step})} > 0$ (it can be either fixed or adaptive). Thereby, the algorithm approaches the Pareto boundary by moving inside of the performance region. Some stopping criterion is required (e.g., on the number of iterations) and there is an optional part that can improve the convergence, at the expense of some extra signaling. There will typically be some central entity that oversees the algorithm and informs the base stations when to update $\tau^{(m)}$ and start a new iteration.

The convergence of Algorithm 5 is illustrated in Figure 4.5 for the same two-user global joint transmission scenario and channel realizations as in Figure 4.2(a) of Section 3.3. We try to obtain the max-min fairness point using $\tau^{(\text{step})} = 0.2$ and $\xi = \frac{0.25}{\sqrt{n}}$, where $n$ is the step number in the subproblems. Only 2 iterations are required to achieve

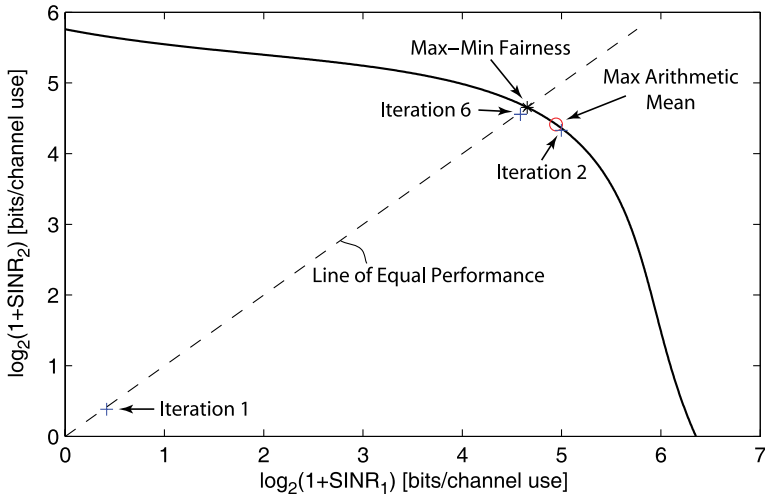



Fig. 4.5 Illustration of the convergence of the distributed resource allocation in Algorithm 5. Max-min fairness is the system utility function and 98% of the optimal performance is achieved after 6 iterations.

an operating point close to the Pareto boundary, but not exactly on the line where the users achieve equal performance. This behavior is due to distributed resource allocation that has not yet converged. 98% of the optimal max-min fairness utility is achieved after 6 iterations, showing the efficiency of the algorithms in this simple scenario.

To summarize, the dual decomposition approaches in Algorithms 4 and 5 should be seen as proofs-of-concept: convex and quasi-convex resource allocation problems can be implemented in a distributed fashion by exchanging control variables rather than CSI. The algorithms in this subsection are not intended for practical implementation, but illustrates a decomposition concept that can also include robustness to CSI uncertainty [257] and time-correlated fading that changes the channels in between iterations [265]. The convergence can be improved using the alternating direction method of multipliers; see [36] for a survey and [239] for applications to robust multi-cell resource allocation.

Uplink–downlink duality provides an alternative decomposition approach where we update and exchange the parameters in Theorem 3.5 in an iterative manner. For weighted max-min optimization with a total power constraint, there are computation-

ally efficient fixed-point algorithms [42, 208, 226, 228, 296] that are also amenable to distributed implementation [42, 59, 208]. These algorithms are less suitable under general multi-cell power constraints, although such constraints can be handled exactly for single-antenna transmitters [43], by iterative subgradient methods for multi-antenna transmitters [59, 308], or by suboptimal approximation of the power constraints [43, 108]. Furthermore, the beamforming parametrization with interference-temperature constraints in Theorem 3.2 enables a simple decentralized algorithm for moving an operating point toward the Pareto boundary [325]. The final operating point greatly depends on the starting point and on parameter values that roughly describe the user priorities; therefore, the approach in [325] is suitable for refining the heuristic truly distributed strategies described in the next subsection.

---

**Remark 4.2 (Distributed Nonconvex Optimization).** If the resource allocation problem is nonconvex, both centralized and distributed solution algorithms are practically infeasible (although they are theoretically implementable by combining the dual decomposition approach in this subsection with the PA or BRB algorithms in Section 2). The natural approach is to search for a locally optimal point instead of a globally optimal point in the performance region. This remark will exemplify some recent algorithms for multi-cell systems, and we refer to [104] for a more thorough survey on centralized and distributed resource allocation algorithms that find stationary points.

Nonconvex problems can be decomposed using interference-prices [225] (similar to the dual decomposition approach above) and the prices are iteratively updated to converge to a local optimum. The distributed algorithm in [280] searches for beamforming vectors that satisfy the KKT conditions. A convex conservative approximation of the weighted sum information rate is obtained in [257], which enables distributed optimization of a lower bound on the system utility. The multi-cell resource allocation is decomposed into many single-cell problems in [291], thus enabling iterative use of algorithms developed for low-complexity single-cell sum information rate optimization. Uplink–downlink duality and the corresponding

beamforming parametrization in Theorem 3.5 are considered in [201] and an algorithm is proposed to iteratively adapt the parameters to the system utility function.

There are many papers claiming that a large fraction of the optimal system utility can be achieved by suboptimal low-complexity optimization algorithms [18, 257, 291], but it is hard to verify for large and complicated multi-cell systems.

### 4.2.2 Truly Distributed Resource Allocation

The previous subsection showed that convex resource allocation problems can be solved to global optimality in a distributed fashion. Several iterations and rounds of control signaling are generally necessary to achieve the solution, which might not be desirable or feasible in practice. By sacrificing the optimality assurance, noniterative resource allocation algorithms that only utilize local CSI can be obtained; for example, by imitating the structure of optimal beamforming [18, 21, 23, 88, 135, 142, 180, 311]. This subsection describes a simple heuristic approach that can be seen as a pragmatic approach to multi-cell coordination but also as a reasonable starting-point for iterative algorithms. We call it a *truly* distributed approach since neither CSI nor other coordination variables (such as $\Theta_{ik}, \widetilde{\Theta}_{ik}, q_{lk}$ in the previous subsection) are exchanged between the base stations.

To bring some insights on the consequences of truly distributed resource allocation, we consider an arbitrary scalarized resource allocation problem

$$\begin{aligned} \underset{\{\mathbf{v}_{jk}\}_{j=1,k=1}^{K_t,K_r}}{\text{maximize}} \quad & f\left(g_1(\mathrm{SINR}_1),\ldots,g_{K_r}(\mathrm{SINR}_{K_r})\right) \\ \text{subject to} \quad & \mathrm{SINR}_k = \frac{\left|\sum_{j=1}^{K_t} \mathbf{h}_{jk}^H \mathbf{C}_{jk} \mathbf{D}_{jk} \mathbf{v}_{jk}\right|^2}{\sigma_k^2 + \sum_{i \neq k} \left|\sum_{j=1}^{K_t} \mathbf{h}_{jk}^H \mathbf{C}_{jk} \mathbf{D}_{ji} \mathbf{v}_{ji}\right|^2} \quad \forall k, \\ & \sum_{k=1}^{K_r} \begin{bmatrix} \mathbf{v}_{1k}^H & \ldots & \mathbf{v}_{K_t k}^H \end{bmatrix} \mathbf{Q}_{lk} \begin{bmatrix} \mathbf{v}_{1k} \\ \vdots \\ \mathbf{v}_{K_t k} \end{bmatrix} \leq q_l \quad \forall l, \end{aligned} \tag{4.34}$$

where the beamforming vectors are decomposed as $\mathbf{v}_k = [\mathbf{v}_{1k}^T \dots \mathbf{v}_{K_t k}^T]^T$ and $\mathbf{v}_{jk} \in \mathbb{C}^{N_j}$ is the contribution at $\text{MS}_k$ from $\text{BS}_j$. An important question is how to maximize $\text{SINR}_k$ in (4.34) in a distributed manner using only local CSI. Starting with the numerator, coherent signal reception can be achieved, for instance, by synchronizing the joint transmissions such that $\mathbf{h}_{jk}^H \mathbf{C}_{jk} \mathbf{D}_{jk} \mathbf{v}_{jk}$ is positive real-valued (or zero) for each $\text{BS}_j$. This requires phase-synchronization between the cooperating base stations (e.g., using GPS-locked reference clocks [124] or common reference signals [177]), but no further control signaling. Achieving coherent interference cancelation (i.e., $|\sum_{j=1}^{K_t} \mathbf{h}_{jk}^H \mathbf{C}_{jk} \mathbf{D}_{ji} \mathbf{v}_{ji}|^2$ is small without enforcing that every term in the sum is small) is more difficult under local CSI, if not impossible in noniterative multi-cell systems [18, 91].[6] Without coherent interference cancelation, there are few reasons for joint transmission; it is more power efficient and reliable to serve each user only by the base station with the strongest channel, although somewhat more unbalanced interference patterns might arise if the user distribution is highly heterogeneous.[7] Truly distributed joint transmission is certainly possible (e.g., by minimizing each term in $|\sum_{j=1}^{K_t} \mathbf{h}_{jk}^H \mathbf{C}_{jk} \mathbf{D}_{ji} \mathbf{v}_{ji}|^2$ individually) and was pioneered in [23], but recent work indicates that the performance gain over coordinated beamforming is small [18]. It will not justify the increased backhaul signaling required deliver the same data signals to multiple base stations. To summarize, joint transmission requires iterative resource allocation with some kind of information exchange between the base stations, while only coordinated beamforming (where $|\sum_{j=1}^{K_t} \mathbf{h}_{jk}^H \mathbf{C}_{jk} \mathbf{D}_{ji} \mathbf{v}_{ji}|^2$ only has one nonzero term) is reasonable in truly distributed systems.

Under coordinated beamforming with per-transmitter power constraints, the beamforming parametrization in Theorem 3.5 takes the following form.

[6] Coherent interference cancelation means that $\text{BS}_j$ selects $\mathbf{v}_{ji}$ to make $\mathbf{h}_{jk}^H \mathbf{C}_{jk} \mathbf{D}_{ji} \mathbf{v}_{ji} \approx -\sum_{m \neq j} \mathbf{h}_{mk}^H \mathbf{C}_{mk} \mathbf{D}_{mi} \mathbf{v}_{mi}$ for all $k \neq i$. The phase and magnitude of the aggregate interference from the other base stations participating in the joint transmission are required, in addition to phase-synchronization. This is more involved than just aligning the useful signals at an intended user. A similar problem arise in interference alignment, where distributed implementations require iterations to find suitable interference subspaces [91].

[7] This problem can be resolved in the dynamic clustering by increasing the range of weakly loaded cells and decreasing the range of heavily loaded cells; see Section 4.7.

**Corollary 4.7.** Suppose the sets $\mathcal{D}_j$ are disjoint between the base stations and that $\mathrm{BS}_j$ has a per-transmitter power constraint of $q_j$. Each Pareto optimal point is achieved by beamforming vectors $\mathbf{v}_{jk} = \sqrt{p_{jk}}\bar{\mathbf{v}}_{jk}$ for $k \in \mathcal{D}_j$ where

$$\bar{\mathbf{v}}_{jk} = \frac{\left(\frac{\mu_j}{q_j}\mathbf{I}_{N_j} + \sum\limits_{i \in \mathcal{C}_j \setminus \{k\}} \frac{\lambda_i}{\sigma_i^2}\mathbf{h}_{ji}\mathbf{h}_{ji}^H\right)^{\dagger}\mathbf{h}_{jk}}{\left\|\left(\frac{\mu_j}{q_j}\mathbf{I}_{N_j} + \sum\limits_{i \in \mathcal{C}_j \setminus \{k\}} \frac{\lambda_i}{\sigma_i^2}\mathbf{h}_{ji}\mathbf{h}_{ji}^H\right)^{\dagger}\mathbf{h}_{jk}\right\|_2} \tag{4.35}$$

for some positive parameters $\{\mu_j\}_{j=1}^{K_r}$ and $\{\lambda_i\}_{i=1}^{K_r}$ that satisfy $\sum_{j=1}^{K_t}\mu_j = \sum_{i=1}^{K_r}\lambda_i = 1$. Furthermore, $p_{jk} \geq 0$ denotes the power allocation and is identically zero whenever $k \notin \mathcal{D}_j$.

This corollary parameterizes the optimal beamforming direction. Recall from Section 3.4 that the heuristic MRT, ZFBF, and SLNR-MAX strategies are related to this optimal structure. The parameter selection should however be adapted to the problem at hand, meaning that the structure of the system utility function $f(\cdot)$ and user performance functions $g_k(\cdot)$ should be taken into account. If these functions are completely symmetric among the users, then this should be reflected in a symmetry in the variables $\lambda_i$ since these describe user priorities (see Corollary 3.7). Additionally, any asymmetry in the performance measures should lead to a corresponding asymmetry in $\{\lambda_i\}_{i=1}^{K_r}$.

Assume that all user performance functions are the same (e.g., information rates or MSEs) and any of the weighted system utility functions in Example 1.11 is used. The weighting factors $w_k \geq 0$ are then the best priority indicators available in the problem formulation. It makes sense to select

$$\lambda_k^{(\text{heuristic})} = \frac{w_k}{\sum_{i=1}^{K_r} w_i} \quad \forall k. \tag{4.36}$$

Furthermore, the parameter $\mu_j$ describes the relative importance of enforcing the power constraint at $\mathrm{BS}_j$. All base stations have total power constraints, thus only the number of users served by $\mathrm{BS}_j$ determines the relative importance of using much transmit power in this

cell. It makes sense to select

$$\mu_j^{(\text{heuristic})} = \frac{|\mathcal{D}_j|}{\sum_{m=1}^{K_t} |\mathcal{D}_m|} \quad \forall j. \tag{4.37}$$

This heuristic parameter selection equals SLNR-MAX when all weighting factors are equal (i.e., $w_k = \frac{1}{K_r}$) but is generally different due to the user priority adaptation. It also resembles the DVSINR beamforming in [18], but the parameters are slightly different. To distinguish this method from previous work, we denote our distributed beamforming scheme as *weighted SLNR-MAX beamforming*.

Observe that (4.36) and (4.37) are both characterized by the problem formulation and can be computed independently at each base station. The beamforming direction (4.35) only depends on local CSI (i.e., $\mathbf{h}_{ji}$ for $i \in \mathcal{C}_j$), thus the transmitting base station can compute $\bar{\mathbf{v}}_{jk}^{(\text{heuristic})}$ by itself using Corollary 4.7, (4.36), and (4.37).

The power allocation can in principle be computed as in Theorem 3.5, but the allocation depends on a system of equations that generally cannot be solved without exchanging information between the base stations. In addition, our parameter selection only utilizes the weighting factors and number of users per cell, and not the exact structure of the system utility and user performance functions. An alternative approach is to solve a heuristic power allocation problem

$$\begin{aligned} &\underset{p_{jk} \geq 0 \; \forall k \in \mathcal{D}_j}{\text{maximize}} \; f\left(g_1(\text{SINR}_1), \ldots, g_{K_r}(\text{SINR}_{K_r})\right) \\ &\text{subject to } \text{SINR}_k = \begin{cases} \frac{p_{jk} |\mathbf{h}_{jk}^H \bar{\mathbf{v}}_{jk}^{(\text{heuristic})}|^2}{\sigma_k^2}, & k \in \mathcal{D}_j, \\ 0, & k \notin \mathcal{D}_j, \end{cases} \\ &\qquad \sum_{k \in \mathcal{D}_j} p_{jk} \leq q_j \end{aligned} \tag{4.38}$$

at $\text{BS}_j$, where all inter-user interference has been ignored. This approximation is intuitive in the high-SNR regime, since $\bar{\mathbf{v}}_{jk}^{(\text{heuristic})}$ will be similar to ZFBF. It also makes sense in the low-SNR regime, because then noise term dominates the inter-user interference. The use at intermediate SNR can be motivated numerically [18]. The power allocation problem in (4.38) can often be solved in closed form; see Theorem 3.16 for the waterfilling solution obtained for the weighted arithmetic mean.

This subsection is concluded with a measurement-based comparison of centralized and distributed resource allocation strategies [18, 111].

---

**Example 4.2 (Measurement-Based Evaluation).** We will compare the performance of different resource allocation strategies, ranging from the optimal centralized strategy to the truly distributed strategy described in this subsection. To capture practical channel fading, spatial correlation, and path loss effects, we utilize narrowband channel measurements conducted in Stockholm, Sweden, using two base stations with four-element uniform linear arrays with $0.56\lambda$ antenna spacing and one user device. The system bandwidth was 9.6 kHz at a carrier frequency in the 1800 MHz band. The user had a uniform circular array with four directional antennas, but we average the signal over its antennas to create a single virtual omni-directional antenna. Further measurement details and maps are available in [18, 111].

In this example, we utilize the channel measurements to create a two-cell scenario where 8 users are randomly located along the measured routes.[8] The system utility function is the weighted sum information rate with

$$w_k = \frac{c_w}{\mathbb{E}\left\{\log_2\left(1 + \frac{K_t}{K_r\sigma_k^2}\max_j(q_j\|\mathbf{h}_{jk}\|_2^2)\right)\right\}} \quad \forall k. \tag{4.39}$$

These weights balance aggregate utility and user fairness (and $c_w$ is scaled to make $\sum_{k=1}^{K_r} w_k = 1$). We assume a transmit power of 15 dBm (per base station), 5 dBi antenna gain, $-131$ dBm noise power, and the measured path losses (from the strongest base station) ranges between $-37$ dB and $-85$ dB. We evaluate five resource allocation strategies:

(1) Optimal Resource Allocation: Calculated using the BRB algorithm in Section 2.3.
(2) Optimal Resource Allocation with Incoherent Interference Reception: Similar to the optimal strategy, but with the additional assumption that base stations cannot cancel out each

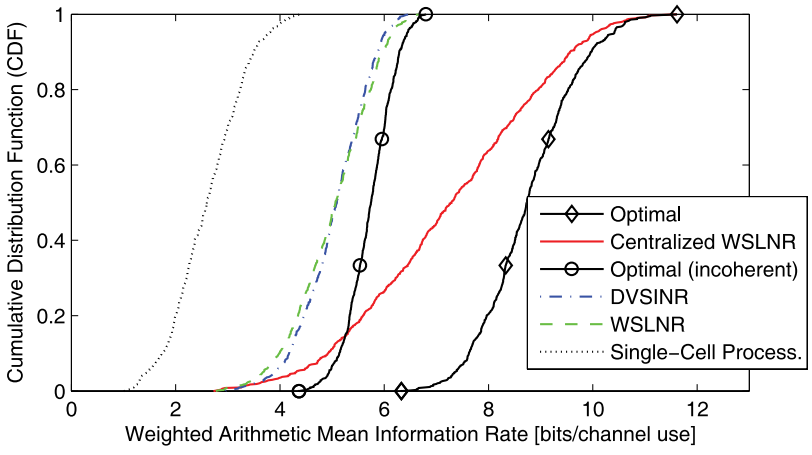
[8] To create balance, two users are placed to have their strongest channel gains $\|\mathbf{h}_{jk}\|_2^2$ from $\mathrm{BS}_1$ while two users have their strongest channel gains from $\mathrm{BS}_2$. The other users are uniformly distributed along the measured routes; see [18].

other's interference through joint transmission. This is an upper bound on truly distributed strategies and maximizes the weighted sum utility with $\mathrm{SINR}_k$ replaced by

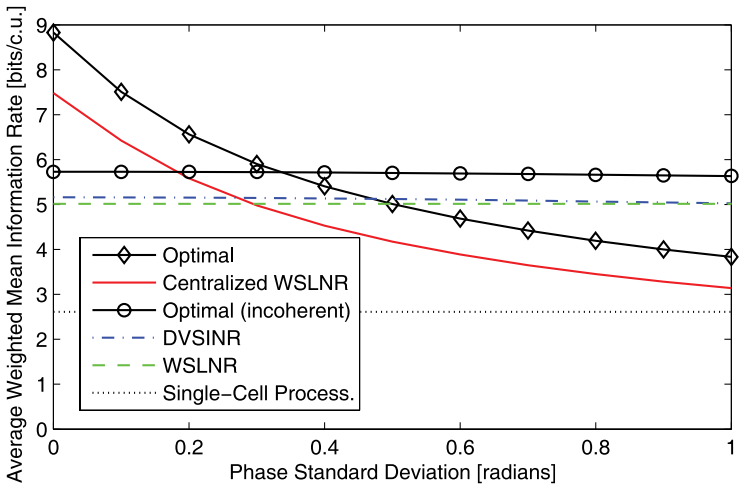
$$\mathrm{SINR}_k^{(\text{incoherent})} = \frac{\left|\sum_{j=1}^{K_t} \sqrt{p_{jk}} \mathbf{h}_{jk}^H \mathbf{C}_{jk} \mathbf{D}_{jk} \bar{\mathbf{v}}_{jk}\right|^2}{\sigma_k^2 + \sum_{i \neq k} \sum_{j=1}^{K_t} \left|\sqrt{p_{ji}} \mathbf{h}_{jk}^H \mathbf{C}_{jk} \mathbf{D}_{ji} \bar{\mathbf{v}}_{ji}\right|^2} \quad \forall k.$$

(3) Centralized WSLNR: Beamforming Parametrization in Theorem 3.5 (using Corollary 3.6) with the heuristic parameter values given in (4.36) and (4.37).
(4) Distributed WSLNR: Truly distributed coordinated beamforming strategy described in this subsection, with user selection based on the distributed ProSched scheme in [18].
(5) JT-DVSINR: Truly distributed joint transmission strategy proposed in [23], with user selection based on the distributed ProSched scheme in [18].
(6) Single-cell processing: Base stations acts as if there is only one cell and out-of-cell interference is included in $\sigma_k^2$.

The cumulative distribution functions (CDFs) of the weighted arithmetic mean information rates (over user locations and channel measurements) are shown in Figure 4.6(a). Centralized WSLNR is relatively close to the optimal solution (on average), but improvements can be made at the lower end. This heuristic noniterative strategy can thus be viewed as a good starting-point for centralized resource allocation. Furthermore, we observe that the truly distributed WSLNR and JT-DVSINR strategies are close together, thus confirming that joint transmission is only useful if there is some kind of control signaling between the base stations. There is a large gap between the centralized and truly distributed strategies because coherent interference cancelation enables transmission to 8 users, while only 4 users are efficiently served in the case of incoherent interference reception. Observe that rudimentary coordinated beamforming brings large improvements over single-cell processing.




(a) CDF of Weighted Arithmetic Mean Information Rate




(b) Weighted Arithmetic Mean Information Rate Under Imperfect Synchronization

Fig. 4.6 Results from the measurement-based evaluation of a two-cell system with 8 users. The performance is evaluated over different random user locations and showed for different resource allocation strategies. (a) CDF of the weighted sum information rate; and (b) Average weighted sum information rate as a function the phase standard deviation $\sigma_\phi$, where the actual channels are modeled as $\mathbf{h}_{jk}^{(\text{effective})} = \mathbf{h}_{jk} e^{\imath \phi_{jk}}$ with $\phi_{jk} \sim \mathcal{N}(0, \sigma_\phi^2)$.

Figure 4.6(b) shows the average weighted arithmetic mean information rates when we introduce synchronization errors between the base stations; it is generally difficult to achieve perfect

phase-synchronization between antennas distributed over a wide area (e.g., due to clock drifts, carrier frequency offsets, and insufficient cyclic prefixes [18, 262, 318, 322]). We assume a very simple but illustrative model where the antennas at each base station are perfectly synchronized, but there is a phase mismatch between the base stations. The effective channels are $\mathbf{h}_{jk}^{(\text{effective})} = \mathbf{h}_{jk} e^{\imath \phi_{jk}}$, where $\imath$ denotes the imaginary number and $\phi_{jk} \sim \mathcal{N}(0, \sigma_\phi^2)$ are random phase deviations. Note that $\sigma_\phi = 0$ means perfect synchronization. The optimal resource allocation and centralized WSLNR are very sensitive to synchronization errors as they rely on coherent interference cancelation where the interfering signals from different base stations should cancel out perfectly. The other schemes are more-or-less unaffected by synchronization errors, since they are not utilizing coherent interference cancelation. The optimal strategy under incoherent interference reception provides a useful performance bound, and the truly distributed schemes are remarkably close to it. We conclude that tight synchronization is required to benefit from joint transmission, while the coordinated beamforming provides a large and relatively robust gain over single-cell processing.

## 4.3 Transceiver Impairments

The beamforming optimization in multi-antenna systems has traditionally been separated from the design of transceiver hardware; that is, the hardware has been assumed to give rise to perfectly linear input–output models such as (1.1) and (1.9). While these models might also include multiplicative and additive distortions that are *independent* of the data signals, physical hardware implementations of radio frequency (RF) transceivers also suffer from impairments that are signal-dependent; for example, due to nonlinear power amplifiers, phase noise, and IQ-imbalance [103, 224].[9] These impairments have a

[9] These might be the most severe impairments in OFDM systems, but there is also carrier-frequency offsets, sampling-rate offsets, quantization noise, etc. [103].

minor impact on point-to-point systems with low-order modulations that can be operated at low SNR (e.g., quadrature phase-shift keying (QPSK) [76]). This perhaps explains why transceiver impairments have received much less attention from the resource allocation community than other nonidealities such as CSI uncertainty and limited backhaul capacity. However, the degradations can be particularly severe in modern multi-cell systems using OFDM (which requires amplifiers with high dynamic range; see Remark 4.4), high-order modulations (which require high SINRs), low-cost equipment (which are relatively non-ideal), and transmit-side interference mitigation (which needs accurate CSI and channel models) [72, 94, 255].

Many of these impairments can be mitigated by proper modeling of the associated distortions, followed by calibration and compensation algorithms [60, 224, 260]. This is related to the *dirty RF paradigm* where the analog components are designed based on some suitable criterion (e.g., high energy-efficiency or small chip area), while nonidealities are compensated by digital signal processing techniques [5, 72]. These techniques cannot remove the distortions completely, but the residual distortions are well-modeled as additive Gaussian noise with a variance that increases with the power of the transmitted signal. The Gaussianity is explained by the aggregate residual of many impairments, whereof some are Gaussian distributed and some behave as Gaussian when summed up [60, 103, 254, 255].

This section will show that the performance of multi-cell systems can be improved and better predicted if the existence of transceiver impairments is taken into consideration in the resource allocation. The analysis builds upon the generalized impairment model in [16, 24, 25, 224, 254, 319], which considers the combined influence of all impairments rather than separately modeling the behavior of each hardware component. This model has been utilized to study the performance of point-to-point systems [25, 76, 254, 255], nonlinear single-cell transmission [93], multi-cell ZFBF [319], and optimal coordinated beamforming [24]. We will show that this model enables generalizations of most concepts in Sections 1 and 2 with retained computational feasibility.

### 4.3.1 Generalized Impairment Model

We consider a generalization of the multi-cell system model in (1.9) where the received signal at $\text{MS}_k$ is

$$y_k = \mathbf{h}_k^H \mathbf{C}_k \left( \sum_{i=1}^{K_r} \mathbf{D}_i \mathbf{v}_i s_i + \boldsymbol{\xi}^{(t)} \right) + n_k + \xi_k^{(r)}, \tag{4.40}$$

where $\mathbf{v}_i s_i$ is the transmitted signal to $\text{MS}_i$ under single-stream beamforming. The new terms $\boldsymbol{\xi}^{(t)} \in \mathbb{C}^N$ and $\xi_k^{(r)} \in \mathbb{C}$ $\forall k$ are the additive *transmitter-distortion* and *receiver-distortion*, respectively. These distortions are modeled as zero-mean complex Gaussian and are statistically independent of the data signals, but the covariance matrices depend on the power of the transmitted and received signals, respectively.

The transmitter-distortion $\boldsymbol{\xi}^{(t)}$ describes the mismatch between the aggregate data signal $\sum_{i=1}^{K_r} \mathbf{D}_i \mathbf{v}_i s_i$ designed by the resource allocation and what is actually transmitted by the RF hardware; see Figure 4.7. The structure of the distortion depends on many things; for example, the quality of the hardware, which compensation algorithms are applied, the number of subcarriers, and whether adjacent transmit antennas share components or essentially are decoupled. We assume that the elements of $\boldsymbol{\xi}^{(t)}$ are uncorrelated, which can be confirmed by measurements on decoupled antenna branches [224, 254, 319].[10] In

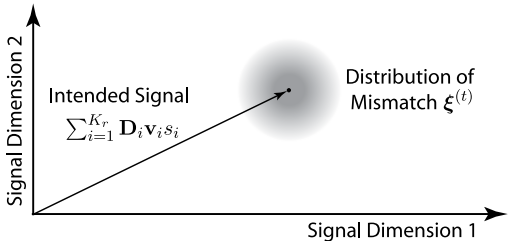



Fig. 4.7 Schematic illustration of the additive mismatch $\boldsymbol{\xi}^{(t)}$ between the aggregate data signal $\sum_{i=1}^{K_r} \mathbf{D}_i \mathbf{v}_i s_i$ and the signal actually created and emitted by the RF hardware. The mismatch is due to transceiver impairments in the transmitter.

[10] The transmitter-distortion is not uncorrelated in general, as proved in [184], but the correlation is expected to be rather weak. Nevertheless, the results in this subsection can

other words, $\boldsymbol{\xi}^{(t)} \sim \mathcal{CN}(\mathbf{0}, \boldsymbol{\Xi})$, where $\boldsymbol{\Xi} \in \mathbb{C}^{N \times N}$ is a diagonal covariance matrix.

The signal power allocated to the $n$th transmit antenna can be computed as $\|\mathbf{T}_n \mathbf{V}_{\text{tot}}\|_F^2$, where $\mathbf{T}_n \in \mathbb{C}^{N \times N}$ is zero except at the $n$th diagonal element and $\mathbf{V}_{\text{tot}} = [\mathbf{D}_1 \mathbf{v}_1 \dots \mathbf{D}_{K_r} \mathbf{v}_{K_r}]$ includes all the signals. The distortion at this antenna increases with $\|\mathbf{T}_n \mathbf{V}_{\text{tot}}\|_F^2$, thus

$$\boldsymbol{\Xi} = \begin{bmatrix} \left(\eta_1(\|\mathbf{T}_1 \mathbf{V}_{\text{tot}}\|_F)\right)^2 & & \\ & \ddots & \\ & & \left(\eta_N(\|\mathbf{T}_N \mathbf{V}_{\text{tot}}\|_F)\right)^2 \end{bmatrix}, \tag{4.41}$$

where $\eta_n(\cdot)$ is a continuous and monotonically increasing function.[11] Observe that this distortion function maps the transmit magnitude to the distortion magnitude (both in unit $\sqrt{\text{mW}}$), and not the powers. This definition simplifies analysis and clarifies the connection to *error vector magnitude* (EVM), which is a common quality measure for RF transceivers [103, 224, 254]. The EVM is the ratio between the average distortion magnitude and the average transmit magnitude, defined as

$$\text{EVM}_n^{(t)} = \sqrt{\frac{\mathbb{E}\left\{\left|[\boldsymbol{\xi}^{(t)}]_n\right|^2\right\}}{\mathbb{E}\left\{\left|\left[\sum_{k=1}^{K_r} \mathbf{D}_k \mathbf{v}_k s_k\right]_n\right|^2\right\}}} = \frac{\eta_n\left(\|\mathbf{T}_n \mathbf{V}_{\text{tot}}\|_F\right)}{\|\mathbf{T}_n \mathbf{V}_{\text{tot}}\|_F}, \tag{4.42}$$

and is often reported as a percentage. Consequently, we can expect $\eta_n(\cdot)$ to behave as $\eta_n(x) = \text{EVM}_n^{(t)} x$ and increase at least linearly with the transmit magnitude $x = \|\mathbf{T}_n \mathbf{V}_{\text{tot}}\|_F$. It can also increase faster than linear if $\text{EVM}_n^{(t)}$ becomes worse/larger when $x$ is large (e.g., due to nonlinearities in the power amplifiers). The EVM requirements in 3GPP Long Term Evolution (LTE) are 8%–17.5% at the transmitter, depending on the anticipated spectral efficiency [103, Section 14.3.4].

---

probably be extended also to correlated distortions, but yet there does not exist a general impairment model that is feasible for mathematical analysis.

[11] The transmitter-distortion in multi-carrier systems (see Section 4.5) generally depend on the power allocated over all the $K_c$ subcarriers. The direct impact of the transmit strategy diminishes with increasing $K_c$ and the distortion power mainly depend on the average power used at each base station. In other words, $\boldsymbol{\Xi}$ may converge to a constant matrix as $K_c$ grows large.

Since the actual transmitted signal under transceiver impairments is $\sum_{i=1}^{K_r} \mathbf{D}_i \mathbf{v}_i s_i + \boldsymbol{\xi}^{(t)}$ (i.e., includes the transmitter-distortion), it is the average power of this signal that should be limited by the power constraints. Due to the statistical independence of each term, the power constraints in (1.4) can be generalized as[12]

$$\sum_{k=1}^{K_r} \mathbf{v}_k^H \mathbf{Q}_{lk} \mathbf{v}_k + \mathrm{tr}(\mathbf{Q}_{l\xi} \boldsymbol{\Xi}) \leq q_l \quad l = 1, \ldots, L. \tag{4.43}$$

The second term models the additional power consumed by the impairments. The weighting matrix $\mathbf{Q}_{l\xi} \in \mathbb{C}^{N \times N}$ is Hermitian positive semi-definite and should typically have the same structure as $\mathbf{Q}_{lk}$, but note that we might have $\mathrm{tr}(\mathbf{Q}_{l\xi}) < \mathrm{tr}(\mathbf{Q}_{lk})$ since not all types of impairments increase the power consumption.[13] As the transmitter-distortions are much weaker than the useful signals, the second term of (4.43) often have a negligible impact on the system [16]. An alternative model is to keep the original power constraints (i.e., set $\mathbf{Q}_{l\xi} = \mathbf{0}_N$) and simply reduce each limit $q_l$ to account for the distortions.

Furthermore, the receiver-distortion $\xi_k^{(r)}$ of $\mathrm{MS}_k$ models the mismatch between ideal and practical reception. This term is modeled as

$$\xi_k^{(r)} \sim \mathcal{CN}(0, \sigma_{k,\xi}^2) \quad \text{where } \sigma_{k,\xi} = \nu_k \left( \| \mathbf{h}_k^H \mathbf{C}_k \mathbf{V}_{\mathrm{tot}} \|_F \right) \tag{4.44}$$

and $\nu_k(\cdot)$ is a continuous and monotonically increasing function. This distortion function describes the receiver impairment characteristics and maps the average received signal magnitude to the corresponding distortion magnitude (both in unit $\sqrt{\mathrm{mW}}$). The main error sources are phase noise and IQ-imbalance, thus we can expect $\nu_k(\cdot)$ to behave as $\nu_k(x) = \mathrm{EVM}_k^{(r)} x$, where $\mathrm{EVM}_k^{(r)}$ is a constant EVM-term.

[12] A portion of the useful signal is basically transformed into distortion in practice. From a modeling perspective, this is essentially the same thing as viewing the remaining signal as the design parameter and the distortion as an additive noise.

[13] Nonlinearities in power amplifiers result in a saturation that might reduce the power consumption of these components.

### 4.3.2 Resource Allocation with Transceiver Impairments

The generalized system model in (4.40) results in a generalized SINR expression for $\text{MS}_k$,

$$\text{SINR}_k = \frac{|\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_k \mathbf{v}_k|^2}{\sigma_k^2 + \sigma_{k,\xi}^2 + \sum_{i \neq k} |\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_i \mathbf{v}_i|^2 + \mathbf{h}_k^H \mathbf{C}_k \boldsymbol{\Xi} \mathbf{C}_k^H \mathbf{h}_k}, \tag{4.45}$$

and the generalized power constraints (4.43), but otherwise the multi-objective and single-objective resource allocation problems in (1.35) and (1.40), respectively, are the same. For example, the feasibility problem in (2.29) (with the QoS constraints $g_k(\text{SINR}_k) \geq r_k^*$) can be generalized as

$$\begin{aligned} &\text{find } \mathbf{v}_1 \ldots, \mathbf{v}_{K_r} && (4.46)\\ &\text{subject to } |\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_k \mathbf{v}_k|^2 \geq g_k^{-1}(r_k^*) \Big( \sigma_k^2 + \sigma_{k,\xi}^2 \\ &\qquad + \sum_{i \neq k} |\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_i \mathbf{v}_i|^2 + \mathbf{h}_k^H \mathbf{C}_k \boldsymbol{\Xi} \mathbf{C}_k^H \mathbf{h}_k \Big) \quad \forall k, \\ &\qquad \sum_{k=1}^{K_r} \mathbf{v}_k^H \mathbf{Q}_{lk} \mathbf{v}_k + \text{tr}(\mathbf{Q}_{l\xi} \boldsymbol{\Xi}) \leq q_l \quad \forall l, \end{aligned}$$

where $\{r_k^*\}_{k=1}^{K_r}$ are fixed. Theorem 2.6 showed that this is a convex problem under ideal hardware, and the following corollary proves the same thing for (4.46).

---

**Corollary 4.8** The feasibility problem in (4.46) can be reformulated as

$$\text{find } \mathbf{v}_k, t_n, \rho_k \quad \forall k, n \tag{4.47}$$

$$\text{subject to } \sum_{k=1}^{K_r} \mathbf{v}_k^H \mathbf{Q}_{lk} \mathbf{v}_k + \sum_{n=1}^{N} \text{tr}(\mathbf{Q}_{l\xi} \mathbf{T}_n) t_n^2 \leq q_l \quad \forall l, \tag{4.48}$$

$$\sqrt{\sigma_k^2 + \rho_k + \sum_{i=1}^{K_r} |\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_i \mathbf{v}_i|^2 + \sum_{n=1}^{N} t_n^2 \mathbf{h}_k^H \mathbf{C}_k \mathbf{T}_n \mathbf{C}_k^H \mathbf{h}_k} \leq \sqrt{\frac{1 + g_k^{-1}(r_k^*)}{g_k^{-1}(r_k^*)}} \Re(\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_k \mathbf{v}_k) \quad \forall k, \tag{4.49}$$

$$\Im(\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_k \mathbf{v}_k) = 0 \quad \forall k, \tag{4.50}$$

$$\eta_n(\|\mathbf{T}_n \mathbf{V}_{\text{tot}}\|_F) \leq t_n \quad \forall n, \tag{4.51}$$

$$\nu_k\left(\|\mathbf{h}_k^H \mathbf{C}_k \mathbf{V}_{\text{tot}}\|_F\right) \leq \rho_k \quad \forall k \tag{4.52}$$

and is jointly convex in the beamforming vectors and the auxiliary optimization variables $\{t_n\}_{n=1}^N$, $\{\rho_k\}_{k=1}^{K_r}$ provided that $\eta_n(\cdot)$ and $\nu_k(\cdot)$ are monotonically increasing convex functions.

---

*Proof.* The auxiliary variable $t_n$ is defined to describe the transmitter-distortion magnitude at the $n$th antenna: $t_n = \eta_n(\|\mathbf{T}_n \mathbf{V}_{\text{tot}}\|_F)$. This enables rewriting

$$\begin{aligned} \mathbf{h}_k^H \mathbf{C}_k \boldsymbol{\Xi} \mathbf{C}_k^H \mathbf{h}_k &= \sum_{n=1}^N t_n^2 \mathbf{h}_k^H \mathbf{C}_k \mathbf{T}_n \mathbf{C}_k^H \mathbf{h}_k \\ \text{tr}(\mathbf{Q}_{l\xi} \boldsymbol{\Xi}) &= \sum_{n=1}^N t_n^2 \text{tr}(\mathbf{Q}_{l\xi} \mathbf{T}_n), \end{aligned} \tag{4.53}$$

which are terms that appear in the SINR and power constraints, respectively. These constraints are convex in $t_n$ and by minimizing over $t_n$, we can have inequality in (4.51) and be sure that equality holds at the optimal solution. Next, we introduce the auxiliary variable $\rho_k$ as in (4.52) to represent the receiver-distortion magnitude at the $k$th user. Equality will always hold if we minimize over $\rho_k$, thus we can replace $\nu_k(\cdot)$ with $\rho_k$ in the SINR expression. The remaining terms in the SINR expressions can be rewritten as convex second-order cones, just as in Theorem 2.6. Finally, the convexity of (4.51) and (4.52) follows if $\eta_n(\cdot)$ and $\nu_k(\cdot)$ are increasing convex functions, as the arguments are convex functions of the beamforming vectors. □

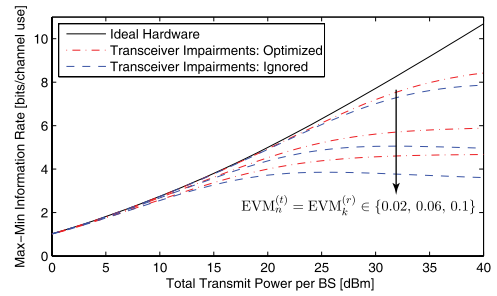
This corollary shows that resource allocation problems with QoS requirements are convex under transceiver impairments. As (4.46) is a subproblem of both the different types of quasi-convex weighted max-min fairness problems in Subsection 2.2.3 and the PA and BRB algorithms for arbitrary monotonic problems in Section 2.3, we thus have established an approach to solve any single-objective resource allocation problem under transceiver impairments.

**Example 4.3 (Max-Min Fairness under Impairments).** We consider a coordinated beamforming scenario with two base stations with $N_j = 4$ antennas and two users per cell. The average single-user SNR $\frac{\mathbb{E}\{q_j \|\mathbf{h}_{jk}\|_2^2\}}{\sigma_k^2}$ is $q_j N_j$ if user $k \in \mathcal{D}_j$ and $q_j \frac{N_j}{3}$ if $k \notin \mathcal{D}_j$, thus each user is closer to its serving base station. We consider the performance under ideal hardware and for different levels of transceiver impairments, which are either handled by optimizing the beamforming vectors as described in this subsection or by ignoring the impairments and optimize as shown in Section 2.2.[14]

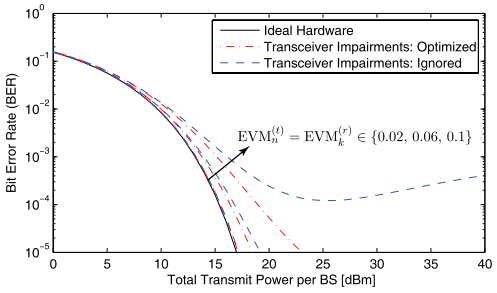
Max-min fairness optimization with different user performance functions is shown in Figure 4.8. First, Figure 4.8(a) considers the information rate. Impairments only yield a minor degradation at low SNR, but the difference to the ideal case is huge at high SNR. This is explained by the bounded asymptotic performance under transceiver impairments, while the ideal case behaves as $1 \cdot \log_2(P) + \mathcal{O}(1)$ and is said to achieve a *multiplexing gain* of one. Although the multiplexing gain is zero in practice, it is shown in [24] that SDMA can still provide several-fold higher performance than TDMA. The figure also shows a clear gain of optimizing the beamforming vectors with impairments in mind.

Figure 4.8(b)–(d) show the BER with three different modulations: QPSK (4-QAM), 16-QAM, and 64-QAM. The more constellation points, the higher SNR is required to achieve a certain BER. We observe that QPSK can handle much larger impairments than 16-QAM and 64-QAM. The impairment-ignoring approach behaves strangely at high SNR (the BER is degraded), simply because it optimize another criterion than we are considering. To summarize, this example shows that impairment modeling and optimization become increasingly important when moving toward higher spectral efficiencies (i.e., larger modulations).

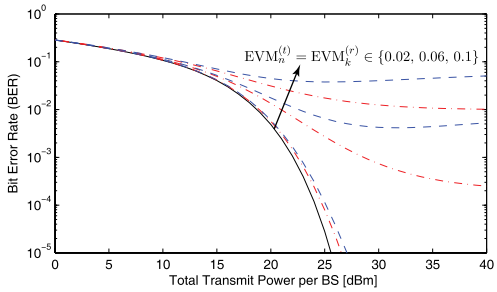
[14] The impairment-optimized approach includes the impairments in the power constraints by using $\mathbf{Q}_{l\xi} = \mathbf{Q}_{lk} \forall l$ in (4.43), while the impairment-ignoring approach only considers the signal power in the power constraints. The latter approach might consume more power than is available, but this has a negligible impact on the results since the signal power greatly dominates the distortion power at practical EVMs.



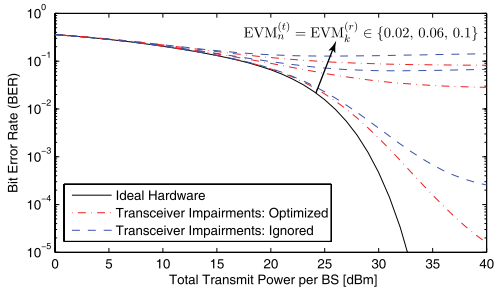

(a) Information Rate




(b) BER with QPSK (4-QAM)




(c) BER with 16-QAM




(d) BER with 64-QAM

Fig. 4.8 Max-min fairness with four different user performance functions in a coordinated beamforming scenario. The performance with different levels of transceiver impairments is compared; the EVM at the transmitters and receivers is 0%, 2%, 6%, or 10%.

We can also obtain a necessary and sufficient characterization of the Pareto boundary by combining Corollary 4.8 with Theorem 3.9. We illustrate this by considering the same global joint transmission scenario as in Subsection 3.3.1, but with transceiver impairments with $\eta_n(x) = \mathrm{EVM}_n^{(t)} x$ and $\nu_k(x) = \mathrm{EVM}_k^{(r)} x$, where $\mathrm{EVM}_n^{(t)} = \mathrm{EVM}_k^{(r)} \in \{0, 0.05, 0.1, 0.15\}$. The performance regions for the information rate is shown in Figure 4.9 for two random channel realizations. We see that impairments reduce the size of the regions, while the shape is mostly unchanged. Furthermore, the user with the strongest channel is also the most sensitive to impairments.

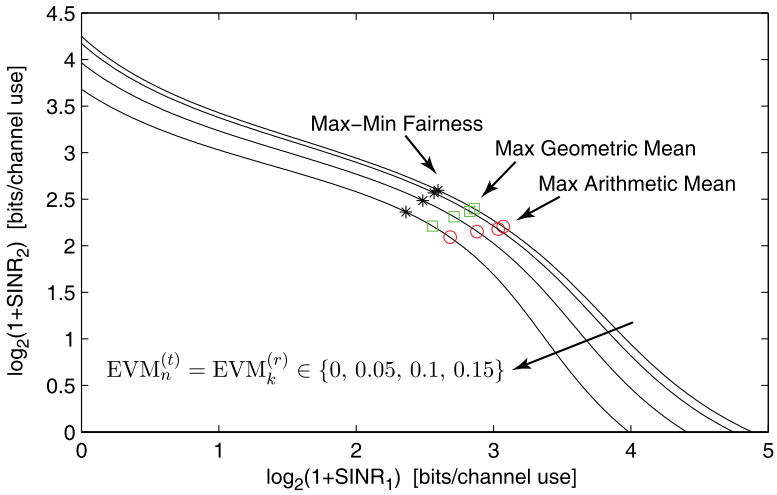
---

**Remark 4.3 (Convexity of Distortion Functions).** Corollary 4.8 requires that the distortion functions $\eta_n(\cdot), \nu_k(\cdot)$ are convex, which is a rather mild requirements and satisfied by any polynomial function with positive coefficients. The interpretation is that the distortion power should increase equally fast or faster than the signal power. For example, $\eta_n(x) = \mathrm{EVM}_n^{(t)} x$ and $\nu_k(x) = \mathrm{EVM}_k^{(r)} x$ are linear (and thus convex) functions when the EVM-terms are constant.

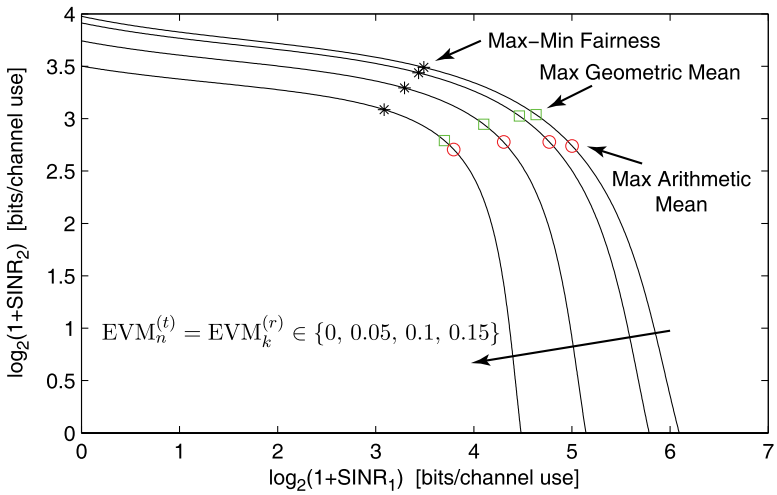
Another example is given in [24], where the transmitter-distortion of a practical LTE power amplifier is modeled as $\eta_n(x) = \mathrm{EVM}_n^{(t)} x(1 + (\frac{x}{\omega})^4)$ and the second term models a fifth-order nonlinearity with a cutoff magnitude of $\omega$ $[\sqrt{\mathrm{mW}}]$ (i.e., $\mathrm{EVM}_n^{(t)}$ is the EVM at low transmit power while it has doubled at $\omega^2$ [mW] and continues to increase). If any of $\eta_n(\cdot)$ and $\nu_k(\cdot)$ increase faster than linear, it is not always beneficial to increase the transmit power since the impact of distortions become more severe [24, 255]. In other words, all power constraint might be inactive at the optimal solution under transceiver impairments, while Theorem 1.9 showed that at least one power constraint is active under ideal hardware.

---

We can obtain a compact expression for the optimal beamforming structure in the special case of linear distortion functions [16]. The parametrization in Theorem 3.5, which utilizes uplink–downlink duality, is easily extended by adding a few terms to account for impairments.




(a) Channel Realization 1




(b) Channel Realization 2

Fig. 4.9 Performance regions for two different channel realizations under global joint transmission (see Subsection 3.3.1 for details). The Pareto boundary is generated by the characterization in Theorem 3.9 using Corollary 4.8 with different levels of transceiver impairments; the EVM at the transmitters and receivers is 0%, 5%, 10%, or 15%.

**Corollary 4.9** Every feasible point $\mathbf{g} \in \mathcal{R}$ under transceiver impairments with $\eta_n(x) = a_n x$ and $\nu_k(x) = b_k x$ are achieved by the parametrization in Theorem 3.5 by replacing (3.18) and (3.20) with

$$\begin{aligned}\boldsymbol{\Psi}_k = &\Bigg( \sum_{i=1}^{K_r} \frac{\lambda_i}{\sigma_i^2} \mathbf{D}_k^H \mathbf{C}_i^H \Big( \mathbf{h}_i \mathbf{h}_i^H (1+b_k^2) + \sum_{n=1}^{N} a_n^2 \mathbf{T}_n^H \mathbf{h}_i \mathbf{h}_i^H \mathbf{T}_n \Big) \mathbf{C}_i \mathbf{D}_k \\ &+ \sum_{l=1}^{L} \frac{\mu_l}{q_l} \Big( \mathbf{Q}_{lk} + \sum_{n=1}^{N} a_n^2 \mathbf{T}_n^H \mathbf{Q}_{l\xi} \mathbf{T}_n \big) \Big), \end{aligned} \tag{4.54}$$

$$[\mathbf{M}]_{ik} = \begin{cases} |\mathbf{h}_i^H \mathbf{C}_i \mathbf{D}_i \bar{\mathbf{v}}_i|^2 (1 - b_i^2 \gamma_i) - \gamma_i \sum_{n=1}^{N} a_n^2 |\mathbf{h}_i^H \mathbf{C}_i \mathbf{T}_n \mathbf{D}_i \bar{\mathbf{v}}_i|^2, & i = k, \\ -\gamma_k \big( |\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_i \bar{\mathbf{v}}_i|^2 (1 + b_i^2) + \sum_{n=1}^{N} a_n^2 |\mathbf{h}_k^H \mathbf{C}_k \mathbf{T}_n \mathbf{D}_i \bar{\mathbf{v}}_i|^2 \big), & i \neq k, \end{cases} \tag{4.55}$$

respectively. Furthermore, every Pareto optimal point $\mathbf{g} \in \partial^+ \mathcal{R}$ is achieved in this way for some nonnegative parameters $\{\lambda_k\}_{k=1}^{K_r}$ and $\{\mu_l\}_{l=1}^{L}$ satisfying $\sum_{k=1}^{K_r} \lambda_k = 1$ and $\sum_{l=1}^{L} \mu_l = 1$.

*Proof.* The proof is identical to the proof of Theorem 3.5, except for the different SINR expression and power constraints. See [16] for details. □

This corollary shows that transceiver impairments have only a minor impact on the optimal beamforming structure. The beamforming directions $\bar{\mathbf{v}}_k = \frac{\boldsymbol{\Psi}_k^\dagger \mathbf{D}_k^H \mathbf{h}_k}{\|\boldsymbol{\Psi}_k^\dagger \mathbf{D}_k^H \mathbf{h}_k\|}$ are rotated to further reduce the inter-user interference and compensate for uneven channel gains (the terms $\sum_{n=1}^{N} a_n^2 \mathbf{T}_n^H \mathbf{h}_i \mathbf{h}_i^H \mathbf{T}_n$ in $\boldsymbol{\Psi}_k$ and $\sum_{n=1}^{N} a_n^2 \mathbf{T}_n^H \mathbf{Q}_{l\xi} \mathbf{T}_n$ act as additional per-antenna constraints). The power allocation is modified by reducing the anticipated channel gain of the useful signal and amplifying the interfering signals.

Interestingly, the number of parameters in Corollary 4.9 is the same under transceiver impairments as with ideal hardware in Theorem 3.5. It is also possible to parameterize the optimal beamforming directions

under arbitrary distortion functions by increasing the number of parameters; we refer to [24] for further details.

## 4.4 Multi-Cast Transmission

Previous sections have considered the scenario when each transmitted data signal is only intended for a single unique user. This section describes the extension to a scenario in which one transmitter equipped with $N$ antennas sends the same data signal to a set of $K_r$ users. Since the transmission performance (e.g., information rate) depends on the weakest link in the user set, the transmitter optimizes its transmission to achieve max-min fairness at the users [80]. The multi-cast beamforming problem to achieve max-min fairness is proven to be nonconvex and NP-hard for $K_r \geq N$ in [245], which stands in contrast to the quasi-convexity proved in Subsection 2.2.3 without multi-cast. For single-antenna transmitters, optimization of multi-cast transmission can in general be solved by the BRB algorithm; see [293].

There are two problem formulations typical for multi-cast beamforming optimization. We consider the same setting as in previous sections but concentrate on $\mathrm{BS}_j$ and denote the set of its multi-cast receivers by $\mathcal{K}_j \subseteq \mathcal{C}_j$ (see [127] for the extension to multiple multi-cast groups). Then, the maximization of the minimum achievable SNR leads to problem statement

$$\underset{\mathbf{v}:\|\mathbf{v}\|\leq q}{\text{maximize}} \min_{k\in\mathcal{K}_j} \frac{|\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_k \mathbf{v}|^2}{\sigma_k^2}. \tag{4.56}$$

Alternatively, the problem can be posed as minimizing the transmit power under a fixed SNR requirement $\gamma$ at all users in $\mathcal{K}_j$, which leads to the problem

$$\underset{\mathbf{v}}{\text{minimize}} \|\mathbf{v}\|^2 \quad \text{subject to} \quad \frac{|\mathbf{h}_k^H \mathbf{C}_k \mathbf{D}_k \mathbf{v}|^2}{\sigma_k^2} \geq \gamma \quad \forall k \in \mathcal{K}_j. \tag{4.57}$$

Note that it is not possible to reformulate (4.56) and (4.57) as convex second-order cone problems (as was done in Subsection 2.2.3 without multi-cast) since the same beamforming vector is used for multiple users [245]. The multi-stream beamforming counterpart to (4.56) will

however be a convex problem, but the solution is far from always rank-one and thus only approximations are viable in practice [127].

In [266], the multi-cast max-min fairness problem (4.56) is studied for $K_r = 2$ users and the set of beamforming vectors which includes the optimal solution is characterized. Using the channel gain region, we obtain the following generalized characterization for an arbitrary number of users.

---

**Corollary 4.10** For a fixed total transmit power $q$, the multi-cast beamforming vector which solves (4.56) is given by

$$\mathbf{v}_k^*(\boldsymbol{\lambda}) = q\,\boldsymbol{v}_{\max}\left(\sum_{k\in\mathcal{K}_j}\lambda_k\mathbf{D}_k^H\mathbf{C}_k^H\mathbf{h}_k\mathbf{h}_k^H\mathbf{C}_k\mathbf{D}_k\right) \tag{4.58}$$

for some set of $|\mathcal{K}_j|$ parameters that satisfies $\lambda_k \geq 0$ and $\sum_{k\in\mathcal{K}_j}\lambda_k = 1$. The operator $\boldsymbol{v}_{\max}$ gives the dominating unit-norm eigenvector.

---

*Proof.* The proof follows from the characterization of the channel gain region in direction $[+1,\ldots,+1]$ and by a contradiction assuming that the operation point is not on the Pareto boundary in this direction. □

The result in Corollary 4.10 is illustrated in Figure 4.10. The max-min SNR point that solves (4.56) is indicated. Note that this point is always on the upper boundary of the channel gain region in direction $[+1\ldots+1]$. The characterization of the solution to the power minimization problem in (4.57) is more difficult, because the solution is not guaranteed to satisfy all SNR constraints with equality. This is shown in Figure 4.10(c) where (4.57) can be feasible although the point $[\gamma\ \gamma]^T$ is outside the channel gain region; the optimal operating point is then at the boundary of the channel gain region but not necessarily in direction $[+1,\ldots,+1]$.

Finally, note that the number of parameters required to describe the optimal beamforming vector in (4.58) increases with number of multi-cast receivers.

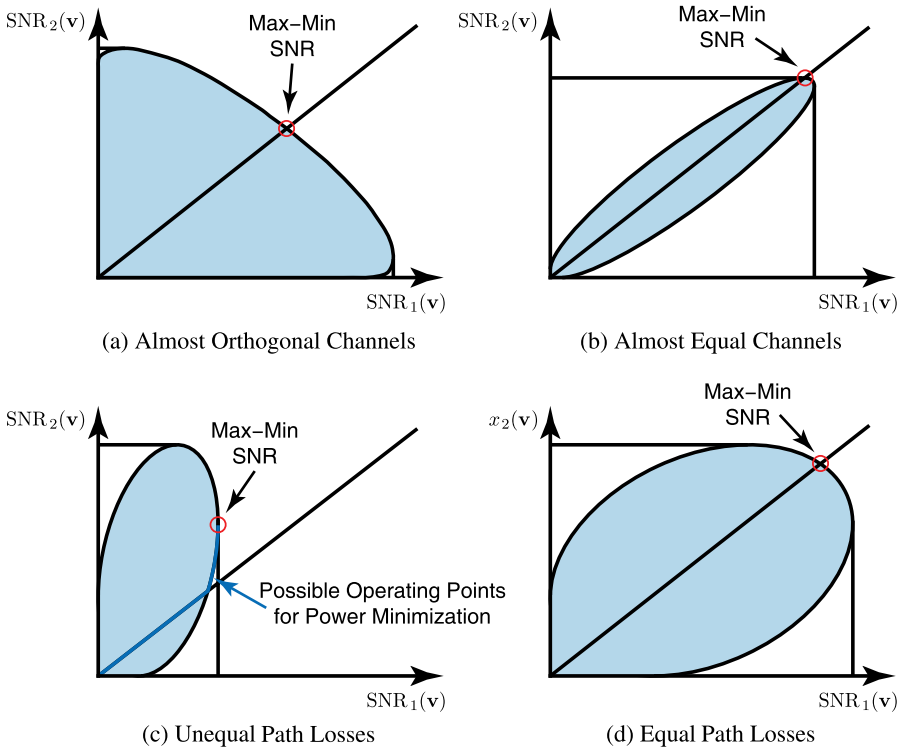



Fig. 4.10 Illustration of the channel gain region $\Omega$ for different channel realizations. The straight lines connecting the upper boundary with the horizontal and vertical axes describe the achievable SNR region. The multi-cast max-min SNR value is given by the intersection of the line in direction $[+1\ +1]^T$ with the SNR region, which not necessarily coincides with the intersection with boundary of $\Omega$.

## 4.5 Multi-Carrier Systems

The single-carrier system model in (1.9) is readily extendable to a multi-carrier system with $K_c$ subcarriers. The received symbol-sampled complex-baseband signal at $\text{MS}_k$ on the $c$th subcarrier is then

$$y_{kc} = \mathbf{h}_{kc}^H \mathbf{C}_k \sum_{i=1}^{K_r} \mathbf{D}_i \mathbf{s}_{ic} + n_{kc}, \tag{4.59}$$

where $\mathbf{h}_{kc} \in \mathbb{C}^N$ is the channel vector, $\mathbf{s}_{ic} \in \mathbb{C}^{N\times 1}$ is the signal intended for $\text{MS}_i$, and $n_{kc} \sim \mathcal{CN}(0, \sigma_{kc}^2)$ is the noise term. Observe that (4.59) is achieved from (1.9) by simply adding a subcarrier-index $c$ at every

term, except $\mathbf{C}_k, \mathbf{D}_k$ which for simplicity are assumed to be the same over the subcarriers. Assuming that all signal and noise variables are independent, the SINR at $\mathrm{MS}_k$ on the $c$th subcarrier is

$$\mathrm{SINR}_{kc}(\mathbf{S}_{1c}, \ldots, \mathbf{S}_{K_r c}) = \frac{\mathbf{h}_{kc}^H \mathbf{C}_k \mathbf{D}_k \mathbf{S}_{kc} \mathbf{D}_k^H \mathbf{C}_k^H \mathbf{h}_{kc}}{\sigma_{kc}^2 + \mathbf{h}_{kc}^H \mathbf{C}_k (\sum_{i \neq k} \mathbf{D}_i \mathbf{S}_{ic} \mathbf{D}_i^H) \mathbf{C}_k^H \mathbf{h}_{kc}} \tag{4.60}$$

with $\mathbf{S}_{kc} = \mathbb{E}\{\mathbf{s}_{kc}\mathbf{s}_{kc}^H\}$. We also extend the power constraints as

$$\sum_{k=1}^{K_r} \sum_{c=1}^{K_c} \mathrm{tr}(\mathbf{Q}_{lkc} \mathbf{S}_{kc}) \leq q_l \qquad l = 1, \ldots, L, \tag{4.61}$$

where $\mathbf{Q}_{lkc} \succeq \mathbf{0}_N$ might model subcarrier-specific characteristics. Multi-carrier power constraints are further discussed in Remark 4.4.

The multi-objective resource allocation problem under multi-carrier transmission can be formulated as

$$\begin{aligned} \underset{\mathbf{S}_{kc} \succeq \mathbf{0}_N \, \forall k,c}{\text{maximize}} \;\; & \{g_1, \ldots, g_{K_r}\} \\ \text{subject to} \;\; & g_k = g_k\left(\{\mathrm{SINR}_{kc}(\mathbf{S}_{1c}, \ldots, \mathbf{S}_{K_r c})\}_{c=1}^{K_c}\right) \;\; \forall k, \\ & \sum_{k=1}^{K_r} \sum_{c=1}^{K_c} \mathrm{tr}(\mathbf{Q}_{lkc} \mathbf{S}_{kc}) \leq q_l \quad \forall l \end{aligned} \tag{4.62}$$

and the sufficiency of single-stream beamforming (on each subcarrier) is easily proved using Lemma 1.6. Compared with the single-carrier MOP in (1.19), the multi-carrier problem in (4.62) can be viewed as joint optimization of $K_c$ superimposed single-carrier systems. The subcarriers are coupled by the user performance functions $g_k(\mathrm{SINR}_{k1}, \ldots, \mathrm{SINR}_{kK_c})$ and the power constraints which generally are shared over the subcarriers. In other words, (4.62) has roughly a factor $K_c$ more optimization variables than (1.19), where $K_c$ can be on the order of several hundred in 3GPP LTE [330]. The scalarized resource allocation problems that were shown to be convex in Section 2.2 might still be convex in the multi-carrier setting (depending on the structure of $g_k(\cdot, \ldots, \cdot)$), but the polynomial computational complexity might not be practically feasible when there are thousands of optimization variables.

The overwhelming multi-carrier complexity can be handled by dividing the subcarriers into subsets of manageable size and solve

these separately (cf. *physical resource blocks* in 3GPP LTE [330]). The coupling in user performance and power constraints can then be resolved by having separable user performance functions (e.g., $g_k = \sum_{c=1}^{K_c} g_{kc}(\mathrm{SINR}_{kc})$ [18]) and fixed power division (since the optimal power allocation can be almost flat over the subcarriers [212]). Alternatively, the optimization problem can be decomposed similarly to the dual decomposition approach in Subsection 4.2.1, giving an iterative optimization procedure with subproblems of manageable complexity [231].

The state-of-the-art beamforming parametrizations in Section 3.2 can be extended with retained computational simplicity. For example, the uplink–downlink duality based parametrization in Theorem 3.5 is generalized to multi-carriers systems in [18].

---

**Corollary 4.11** Every feasible point $\mathbf{g} \in \mathcal{R}$ is achieved by beamforming vectors $\mathbf{v}_{kc} = \sqrt{p_{kc}}\bar{\mathbf{v}}_{kc}$ for all $k, c$, where

$$\bar{\mathbf{v}}_{kc} = \frac{\boldsymbol{\Psi}_{kc}^{\dagger}\mathbf{D}_k^H\mathbf{h}_{kc}}{\|\boldsymbol{\Psi}_{kc}^{\dagger}\mathbf{D}_k^H\mathbf{h}_{kc}\|}, \tag{4.63}$$

$$\begin{bmatrix} p_{1c} & \cdots & p_{K_rc} \end{bmatrix} = \begin{bmatrix} \gamma_{1c}\sigma_{1c}^2 & \cdots & \gamma_{K_rc}\sigma_{K_rc}^2 \end{bmatrix} \mathbf{M}_c^{\dagger}, \tag{4.64}$$

$$\boldsymbol{\Psi}_{kc} = \left( \sum_{l=1}^{L} \frac{\mu_l}{q_l}\mathbf{Q}_{lkc} + \sum_{i=1}^{K_r} \frac{\lambda_{ic}}{\sigma_{ic}^2}\mathbf{D}_k^H\mathbf{C}_i^H\mathbf{h}_{ic}\mathbf{h}_{ic}^H\mathbf{C}_i\mathbf{D}_k \right), \tag{4.65}$$

$$\gamma_{kc} = \frac{\lambda_{kc}}{\sigma_{kc}^2}\mathbf{h}_{kc}^H\mathbf{D}_k\big(\boldsymbol{\Psi}_{kc} - \frac{\lambda_{kc}}{\sigma_{kc}^2}\mathbf{D}_k^H\mathbf{C}_k^H\mathbf{h}_{kc}\mathbf{h}_{kc}^H\mathbf{C}_k\mathbf{D}_k\big)^{\dagger}\mathbf{D}_k^H\mathbf{h}_k, \tag{4.66}$$

$$[\mathbf{M}_c]_{ik} = \begin{cases} |\mathbf{h}_{ic}^H\mathbf{C}_i\mathbf{D}_i\bar{\mathbf{v}}_{ic}|^2, & i = k, \\ -\gamma_{kc}|\mathbf{h}_{kc}^H\mathbf{C}_k\mathbf{D}_i\bar{\mathbf{v}}_{ic}|^2, & i \neq k, \end{cases} \tag{4.67}$$

for some non-negative parameters $\{\lambda_{kc}\}_{k=1,c=1}^{K_r,K_c}$ and $\{\mu_l\}_{l=1}^{L}$.

---

This corollary provides a beamforming parametrization with only $K_rK_c + L$ parameters, and the approach in Corollary 3.6 reduces it to $K_rK_c + L - 2$ parameters between zero and one. Heuristic selection of

these parameters might yield a reasonable starting point for further multi-carrier performance optimization [18]. The similarity between Corollary 4.11 and the single-carrier parametrization in Theorem 3.5 shows that the introduction of multiple subcarriers has only a small impact on the optimal solution structure.

---

**Remark 4.4 (Multi-Carrier Power Constraints).** The physical power constraints in OFDM-based multi-carrier systems are not easily formulated as (4.61); the connection between the transmit power allocated in the complex baseband and the resulting RF waveform is complicated and nonconvex. In fact, the waveform is typically Gaussian-like (as it is the superposition of $K_c$ random signals where $K_c$ is large) and can exhibit a high peak-to-average power ratio (PAPR), which is undesirable as it requires hardware components with high dynamic range. The PAPR can be reduced by bounding the per-antenna transmit power in the complex baseband from both above and below [11, 195]. Such constraints can be formulated as convex if we optimize over signal correlation matrices $\mathbf{S}_{kc}$ of arbitrary rank. However, we are not aware of any tractable problem formulation that enables complete control over the PAPR.

For a given OFDM signal, the PAPR can be improved by distorting the signal before transmission. The simplest approach might be to remove the largest amplitude spikes by clipping techniques, but there are more powerful techniques that utilize convex optimization to basically minimize the PAPR under constraints on the tolerable error in the modulated signal; we refer to [2, 154, 188] for further details on this subject. The combination of beamforming optimization and distorting of the signal for PAPR reduction seems to be an open problem.

---

## 4.6 Multi-Antenna Users

The analysis in this tutorial is based on having a single effective antenna at each user, which according to Section 1.2 means that $\mathrm{MS}_k$ is equipped with either a single antenna or $M_k > 1$ antennas that are combined into a single effective antenna prior to resource allocation (e.g., using receive combining or antenna selection). This assumption

simplifies analysis, but has also practical advantages such as requiring less hardware on the user devices and acquiring less CSI per user.

This section explores the possibility of also including the use of multiple receive antennas (per user) in the resource allocation optimization, which certainly should increase the performance region. In order to explain the fundamental difference between multi-cell MIMO systems and the multi-cell MISO setup in the rest of this tutorial, we first focus on the two-user MIMO interference channel in which $\mathrm{BS}_j$ transmits a single data stream (single-stream beamforming) to $\mathrm{MS}_j$ for $j = 1, 2$ [47].

Suppose $\mathrm{BS}_j$ is equipped with $N_j \geq 2$ transmit antennas, while $\mathrm{MS}_k$ has $M_k \geq 2$ receive antennas. Only one data stream is transmitted to each user and the received signal at $\mathrm{MS}_1$ is modeled as[15]

$$y_1 = \boldsymbol{\zeta}_1^H \big( \mathbf{H}_{11} \mathbf{v}_1 s_1 + \mathbf{H}_{21} \mathbf{v}_2 s_2 + \mathbf{n}_1 \big), \tag{4.68}$$

where $s_j \in \mathbb{C}$ is the data signal with $\mathbb{E}\{|s_j|^2\} = 1$ transmitted by $\mathrm{BS}_j$ employing the beamforming vector $\mathbf{v}_j \in \mathbb{C}^{N_j}$ for $j = 1, 2$. Furthermore, $\boldsymbol{\zeta}_k \in \mathbb{C}^{M_k}$ is the receive combining vector employed at $\mathrm{MS}_k$. The term $\mathbf{n}_k \in \mathbb{C}^{M_k}$ is the additive white Gaussian noise vector with zero-mean and covariance matrix $\sigma_k^2 \mathbf{I}_{M_k}$. The channel matrix between $\mathrm{BS}_j$ and $\mathrm{MS}_k$ is denoted $\mathbf{H}_{jk} \in \mathbb{C}^{M_k \times N_j}$. For simplicity, we assume per-transmitter power constraints $\|\mathbf{v}_j\|^2 \leq 1$.

Assuming that the users perform single-user decoding and treat interference as additional additive white Gaussian noise, the performance of $\mathrm{MS}_k$ can be modeled as

$$g_k(\mathbf{v}_1, \mathbf{v}_2, \boldsymbol{\zeta}_k) = g_k\big(\mathrm{SINR}_k(\mathbf{v}_1, \mathbf{v}_2, \boldsymbol{\zeta}_k)\big), \tag{4.69}$$

for a strictly monotonically increasing user performance function $g_k(\cdot)$ (see Definition 1.4). The SINR of $\mathrm{MS}_1$ is then given by

$$\begin{aligned} \mathrm{SINR}_1(\mathbf{v}_1, \mathbf{v}_2, \boldsymbol{\zeta}_1) &= \frac{|\boldsymbol{\zeta}_1^H \mathbf{H}_{11} \mathbf{v}_1|^2}{\sigma_1^2 + |\boldsymbol{\zeta}_1^H \mathbf{H}_{21} \mathbf{v}_2|^2} \\ &\leq \mathbf{v}_1^H \mathbf{H}_{11}^H \big( \sigma_1^2 \mathbf{I}_{M_1} + \mathbf{H}_{21} \mathbf{v}_2 \mathbf{v}_2^H \mathbf{H}_{21}^H \big)^{-1} \mathbf{H}_{11} \mathbf{v}_1 \\ &= \mathrm{SINR}_1(\mathbf{v}_1, \mathbf{v}_2), \end{aligned} \tag{4.70}$$

[15] The expressions for the link $\mathrm{BS}_2 \mapsto \mathrm{MS}_2$ are obtained by interchanging indices.

where the maximal SINR is achieved using the linear MMSE filter $\boldsymbol{\zeta}_1^{(\mathrm{MMSE})} = \left(\sigma_1^2 \mathbf{I}_{M_1} + \mathbf{H}_{11}\mathbf{v}_1\mathbf{v}_1^H\mathbf{H}_{11}^H + \mathbf{H}_{21}\mathbf{v}_2\mathbf{v}_2^H\mathbf{H}_{21}^H\right)^{-1}\mathbf{H}_{11}\mathbf{v}_1$. This is the optimal receive combining vector under linear receive processing; see Section 3.4. By replacing $\boldsymbol{\zeta}_k$ with the expression for the optimal $\boldsymbol{\zeta}_k^{(\mathrm{MMSE})}$, we can thus write the SINRs as $\mathrm{SINR}_k(\mathbf{v}_1, \mathbf{v}_2)$ instead.

The expression for $\mathrm{SINR}_1(\mathbf{v}_1, \mathbf{v}_2)$ in (4.70) has a difficult mathematical structure that makes it hard to analyze the SINR directly. The following result from [48, Proposition 1] shows that the difficulty lies in the coupling between the beamforming vectors $\mathbf{v}_1$ and $\mathbf{v}_2$.

---

**Lemma 4.12** For the two-user single-stream MIMO interference channel, the SINR in (4.70) of each user under MMSE receive filtering can be expressed as

$$\mathrm{SINR}_1(\mathbf{v}_1, \mathbf{v}_2) = \sin^2(\theta_1)\,\frac{\|\mathbf{H}_{11}\mathbf{v}_1\|_2^2}{\sigma_1^2} + \cos^2(\theta_1)\,\frac{\|\mathbf{H}_{11}\mathbf{v}_1\|_2^2}{\sigma_1^2 + \|\mathbf{H}_{21}\mathbf{v}_2\|_2^2}, \tag{4.71}$$

where $\cos(\theta_1) = \frac{|\mathbf{v}_1^H\mathbf{H}_{11}^H\mathbf{H}_{21}\mathbf{v}_2|}{\|\mathbf{H}_{11}\mathbf{v}_1\|_2\|\mathbf{H}_{21}\mathbf{v}_2\|_2}$ and $\theta_1 \in [0, \frac{\pi}{2}]$.

---

This lemma shows that $\mathrm{SINR}_1(\mathbf{v}_1, \mathbf{v}_2)$ can be viewed as a linear combination of $\frac{\|\mathbf{H}_{11}\mathbf{v}_1\|_2^2}{\sigma_1^2}$ and $\frac{\|\mathbf{H}_{11}\mathbf{v}_1\|_2^2}{\sigma_1^2+\|\mathbf{H}_{21}\mathbf{v}_2\|_2^2}$ with the weights $\sin^2(\theta_1)$ and $\cos^2(\theta_1)$. That is, the SINR depends not only on the values of the desired signal power $\|\mathbf{H}_{11}\mathbf{v}_1\|^2$ and the interference power $\|\mathbf{H}_{21}\mathbf{v}_2\|^2$, but also on the Hermitian angle $\theta_1$ between their effective channel directions. Observe that $\cos(\theta_1) = 1$ and $\sin(\theta_1) = 0$ in the MISO case, while the existence of a flexible $\theta_1$ creates an additional coupling in the SINRs in the MIMO case. Therefore, it is significantly more difficult to analyze and optimize the performance of a MIMO interference channel. In fact, even the resource allocation problem with fixed QoS requirements is provably NP-hard in the MIMO case [158, 210]. As this is the computationally simplest problem in the MISO case, we cannot expect to solve *any* multi-cell single-stream MIMO resource allocation problem to global optimally in practice.

Consequently, practical algorithms need to search for stationary points. An alternating optimization approach is developed in [48] to

find stationary points close to arbitrary Pareto optimal points for the two-user single-stream MIMO interference channel. The turning points (i.e., points where the vertical and horizontal weak Pareto boundary changes to the strict Pareto boundary) are characterized in closed form.

Under an arbitrary number of users, [158] finds stationary points of the max-min fairness optimization by alternating between optimizing the transmit strategy with fixed receive combining (which equals the MISO scenario solved in Subsection 2.2.3) and updating the receive combining vector $\boldsymbol{\zeta}_k$ as the MMSE filter for the current beamforming vectors. A linear transmission algorithm for weighted sum information rate maximization for the MIMO interfering broadcast channel is proposed in [243], where the optimization problem is transformed to an equivalent sum-MSE minimization problem. An alternating optimization algorithm with three steps is proposed: (1) update the weight matrices; (2) update the MMSE receive matrices; and (3) update the transmit covariance matrices. The iterative algorithm is guaranteed to converge to a stationary point. Therefore, a series of operating points can be achieved corresponding to the maximum weighted sum information rates with different weights. However, this approach cannot achieve all the Pareto optimal points when the performance region is nonconvex [324]. Furthermore, it is not clear how to obtain the corresponding weights in order to achieve given rate tuples. An alternative approach for the same optimization problem was recently proposed in [150], based on on iterative multi-cell waterfilling.

A few works have considered the performance region of the MIMO interference channel in general multi-stream multi-cell multi-user scenarios. For example, jointly optimized MMSE and zero-forcing MIMO transceiver algorithms for the two-user MIMO interference channel (called interference aware-coordinated beamforming (IA-CBF)) are proposed in [49]. However, the MMSE IA-CBF can only achieve a lower bound on the sum information rate, and the zero-forcing IA-CBF only finds operating points achievable by zero-forcing strategies.

---

**Remark 4.5 (Optimality of Single-Stream Beamforming).** The sufficiency of single-stream beamforming for single-antenna users was proved in Theorem 1.8 under perfect CSI, and it also seem to hold

true under CSI uncertainty [51, 239, 251]. The result is rather intuitive, because single-antenna users can only make a single observation. Multi-antenna users will however make multiple signal observations and can thus receive and efficiently decode multi-stream beamforming transmissions; in fact, the capacity-achieving TDMA scheme is multi-stream beamforming where the signal correlation matrix is adapted to the right singular vectors of the channel matrix [269]. Apart from decoding multi-stream beamforming signals, the receive antennas at a given user can also be utilized for interference-aware receive combining that essentially creates an effective MISO channel with relatively good properties (i.e., balance between strong channel gain and good co-user separability). This raises a fundamental question: Should the existence of multiple antennas at each user be utilized for multi-stream beamforming or is it better to still perform single-stream beamforming and exploit receive combining instead? To put it differently, suppose the system should convey $N$ data streams in parallel. Should we divide these among just a few users that are served with multiple streams or should we select $N$ different users and perform single-stream beamforming?

The line of work in [15, 20, 28, 268] shows that it is advisable to perform single-stream beamforming also in the multi-user MIMO case, especially when the resources for channel estimation and feedback are limited. The basic explanation is that receive combining provides resilience toward spatial correlation and nonorthogonality between co-user channels, which are two major limiting factors in SDMA. This recent observation motivates further study on single-stream beamforming transmission to multi-antenna users. It is also very positive from a hardware perspective, because reception of single-stream beamforming is less demanding than reception of multi-stream beamforming.

## 4.7 Design of Dynamic Cooperation Clusters

This section will discuss the design of dynamic cooperation clusters (DCCs), where the word *dynamic* refers to adaptation to time-variant characteristics such as channel properties, user mobility, activity levels, user load distribution, and base station failure. Recall from

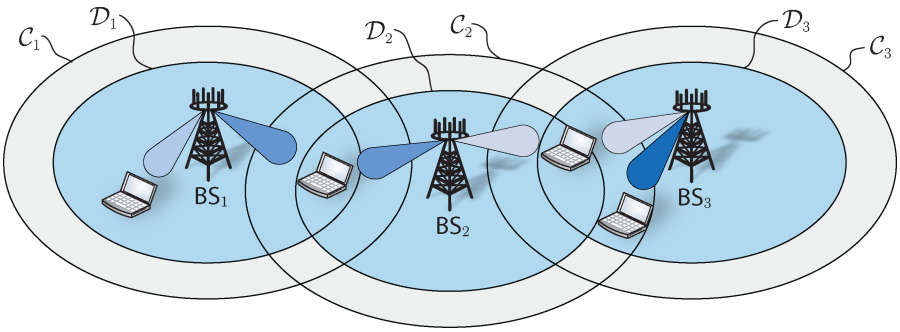



Fig. 4.11 Illustration of the overlapping nature of dynamic cooperation clusters. $\text{BS}_j$ forms different cooperation constellations when serving different users. $\mathcal{D}_j$ are the users that it serves and $\mathcal{C}_j$ are the users considered in the spatial interference coordination.

Definition 1.3 that this tutorial considers a DCC model where $\text{BS}_j$ is serving a set of users $\mathcal{D}_j$, while taking interference caused at users in the set $\mathcal{C}_j$ into consideration. This structure is illustrated in Figure 4.11. The users served by a base station are naturally a subset of those users that it tries to avoid interference at, thus $\mathcal{D}_j \subseteq \mathcal{C}_j \, \forall j$. As illustrated by Examples 1.1–1.5, this type of sets can describe a variety of different multi-cell scenarios. It ranges from interference channels (where each base station serves a single unique user) to global joint transmission (where all base stations transmit jointly to all users). A key property of the DCCs is that each base station is allowed to cooperate with all of its neighboring base stations and form different cooperation constellations when serving different users — this stands in contrast to the earliest work on static and dynamic clustering where the base stations are divided into disjoint groups (see Figure 1.4) [106, 174, 199, 323]. The overlapping (nondisjoint) nature of DCCs is illustrated in Figure 4.11, where $\text{BS}_2$ cooperates simultaneously with different base stations when serving different users.

While this tutorial provides a thorough framework for resource allocation for given DCCs, the practical design of DCCs is a relatively new and unexplored research topic. This section describes some fundamental factors that limit the cardinality and shape of $\mathcal{C}_j, \mathcal{D}_j$ in practical applications and outlines some recent dynamic clustering approaches.

The sets $\{\mathcal{D}_j\}_{j=1}^{K_r}$ of users served by each base station (and potentially by multiple ones) are selected under the following conditions:

(1) Each active user should have a *master base station* (MBS) that guarantees its data services. This makes sure that no one falls through the cracks and also creates a natural hierarchy between the MBS and other base stations that might participate in a joint transmission. The distributed resource allocation in Subsection 4.2.1 requires the existence of an MBS that computes the beamforming vector for the user. Each user can either suggest or be appointed an MBS. In any case, the choice should be based on CSI and the base station with the strongest channel conditions is the natural choice. It might however be beneficial to select another base station when the strongest one has a heavy user load.
(2) The backhaul infrastructure should support the joint transmission to a user, in terms of enabling fast exchange of control signals for resource allocation and phase synchronization [86]. Furthermore, joint transmission increases the delay spread and thereby limits the number of base station that can perform coherent interference cancelation [322]; recall from Subsection 4.2.2 that joint transmission is only useful when coherent interference cancelation can be achieved.
(3) Joint transmission requires the same data signal to be delivered (and equally encoded for transmission) to all of the serving base stations, which can significantly increase the backhaul signaling [175, 313]. Therefore, the limited backhaul capacity suggests that joint transmission only is used when the increase in throughput outweighs the increased demands on the backhaul infrastructure.[16]

[16] The backhaul capability depends greatly on the infrastructure; fiber-optic cables might have almost infinite capacity for practical purposes, while conventional copper cables and wireless links are much more capacity limited. Cellular networks based on existing/conventional infrastructure are however expected to have heterogeneous backhaul networks, where new high-capacity fiber-optic links coexist with older links having modest capacities.

(4) Proximity is not measured geographically but by the average channel gain, taking possible differences in transmit power between base stations into account. We let users with one strong dominating channel from one of the base stations be called *cell center users*, while users with relatively similar channel gains from multiple base stations are called *cell edge users*. These different user scenarios call for a variable number of base stations per user [92]; only the base stations that might cause relatively strong interference to a user should consider participating in joint transmission to this user. Cell edge users are therefore prone for joint transmission from several neighboring base stations, while cell center users might as well only be served by their MBSs.

(5) The base stations and many objects in the propagation environment are static. The geographical area can therefore be divided into *location bins* where the statistics of the channel propagation is almost static [109]. It therefore makes sense to apply the same clustering on all users that are located in the same bin. The size and structure of the location bins can be very different at different places, but can be determined in advance through system calibration. Although the channel statistics capture many important large-scale fading effects, the clustering should also depend on user mobility and macroscopic conditions such as congestion.

The sets $\{\mathcal{C}_j\}_{j=1}^{K_r}$ of users that are considered in the beamforming at each base station are selected under the following conditions:

(6) The channels between the base station and all the users that it includes in its interference coordination need to be estimated, for example, by using training signaling [22]. The resources available for channel estimation are fundamentally limited by the coherence time of the channels [45, 122], since the estimates should both be acquired and utilized for resource allocation and transmission during this time period. The number of orthogonal training signals is therefore limited and need to be simultaneously reused at multiple base

stations in FDD systems and multiple users in TDD systems. The distance between entities using the same training signals essentially bounds the area where a base station can obtain reliable CSI.

(7) All neighboring base stations in TDD can estimate their own individual channel components by listening to the *same* uplink training signal from a user. On the contrary, separate estimation and feedback of the channels from each base station antenna is required in FDD systems, which increases the feedback load linearly with the number of base stations that request channel estimates — and the user needs to know which these base stations are. Backhaul signaling of CSI might also be needed in FDD, depending on whether the feedback is decoded at the MBS (and then sent over the backhaul) or directly at the corresponding base stations. Generally speaking, TDD systems can enable larger coordination sets than FDD systems, and also the use of an arbitrary number of antennas per base station (i.e., $N$ has no impact on the estimation resources [122, 220]). On the other hand, FDD systems have a potential advantage in the fact that the CSI are fed back; neighboring base stations can then listen to all the CSI feedback from a user and thereby achieve CSI also for other cells. Although the obtained CSI might be slightly different (due to variations in the feedback channel conditions), this information can improve the convergence of distributed resource allocation schemes since base stations can predict the decisions of their neighbors [200, 312].

The selection of $\mathcal{C}_j, \mathcal{D}_j$ should certainly be based on some kind of CSI, but the question is how much information is necessary to make good decisions. Intuitively, the cluster dynamics are predominated by large-scale channel properties (e.g., distant-dependent attenuation and shadowing) and should thus be modeled as a function of the current channel statistics (e.g., $\mathbb{E}\{\mathbf{h}_{jk}\mathbf{h}_{jk}^H\}$) measured over some suitable time-window. This intuition is confirmed in [92, 198], where cluster-adaptation based on the instantaneous channel vectors only shows a

marginal gain over statistical clustering. It is also practically desirable to change the clusters over a larger time-scale than the resource allocation decisions are made. Each base station can then easily be aware of which users it serves and which neighboring base stations that serve the same users. Furthermore, it enables CSI acquisition for the right set of users.

---

**Example 4.4 (Simple Clustering).** A simple clustering algorithm would be to include $\text{MS}_k$ in $\mathcal{C}_j$ if the average channel gain $\mathbb{E}\{\|\mathbf{h}_{jk}\|_2^2\}$ (over some suitable time-window) is above a certain threshold value [18]. The fulfillment of this condition is rechecked at the same time-scale as the estimate of $\mathbb{E}\{\|\mathbf{h}_{jk}\|_2^2\}$ is updated. $\text{MS}_k$ appoints base station

$$m = \underset{\{j:\, k \in \mathcal{C}_j\}}{\arg\max}\, \mathbb{E}\{\|\mathbf{h}_{jk}\|_2^2\} \tag{4.72}$$

as its MBS, which is the one with the strongest channel. The user then computes the ratio between the average channel gain of the MBS and the gain of the second strongest base station,

$$\frac{\mathbb{E}\{\|\mathbf{h}_{mk}\|_2^2\}}{\max_{\{j:k\in\mathcal{C}_j\}\setminus\{m\}} \mathbb{E}\{\|\mathbf{h}_{jk}\|_2^2\}}. \tag{4.73}$$

Knowing that base stations that perform joint transmission should have relatively similar channel gains to the user, this ratio is compared with a threshold that determines if it seems likely that the system will benefit from joint transmission [92]. This procedure can be repeated to also include a third (and fourth, and so on) base station in the joint transmission. Observe that this heuristic algorithm is distributed in the sense that $\text{BS}_j$ decides on $\mathcal{C}_j$ and $\text{MS}_k$ decides which base stations it want to be served by.

---

The clustering algorithm can be made more analytic than in Example 4.4, but the combinatorial nature makes it easy to formulate optimization problems that are too difficult to be solved in practice.

A generic problem formulation (similar to [128]) is

$$\begin{aligned} &\underset{\mathcal{C}_j, \mathcal{D}_j \, \forall j}{\text{maximize}} \ \varphi(\{\mathcal{D}_j\}_{j=1}^{K_t}, \{\mathcal{C}_j\}_{j=1}^{K_t}) \\ &\text{subject to } \mathcal{D}_j \subseteq \mathcal{C}_j \quad \forall j, \\ &\qquad\qquad \text{upper bounds on } |\mathcal{C}_j| \text{ and } |\mathcal{D}_j| \quad \forall j, \\ &\qquad\qquad \text{constraints on which base stations that} \\ &\qquad\qquad \text{BS}_j \text{ may perform joint transmission with} \quad \forall j. \end{aligned} \tag{4.74}$$

The utility function $\varphi(\cdot,\cdot)$ describes the preference of a certain clustering. The inherent difficulty in (4.74) is threefold: (1) The number of possible clusterings increase rapidly with the size of the system; (2) The utility function $\varphi(\cdot,\cdot)$ should be explicitly defined and selected to indicate the utility of the final resource allocation (without having to solve an optimization problem to evaluate it); and (3) The problem requires global CSI (statistical or instantaneous) but should be formulated to enable distributed implementation.

A greedy algorithm for solving (4.74) under zero-forcing constraints is proposed in [128]. Formulations of (4.74) as a linear combinatorial problem are given in [173, 290]. Alternatively, the problem can be formulated as a graph where users and base stations are nodes [86]. There will be an edge between a user and the base stations that might serve it, and also edges between base stations that might cooperate. In this interpretation, clustering corresponds to selecting a subset of all edges.

This section has presented some guidelines and algorithms for dynamic clustering that appeared during the last few years [18, 86, 92, 128, 173, 198, 198, 290]. We hope to see many more results on this subject in the near future, both in terms of distributed low-complexity clustering algorithms and evaluation of such algorithms in large multi-cell systems with practical properties.

### 4.7.1 Distributed Multi-Cell Scheduling

This tutorial basically describes a scenario where all $K_r$ users are expected to be served at once; all system utility functions in Example 1.11 (except the arithmetic mean) require that all users with strictly positive weighted factors, $w_k > 0$, are allocated nonzero performance.

*Scheduling* (or *user selection*) is however an essential part of many multi-cell systems, since the number of potential users can be much larger than the number of data streams that can be transmitted with manageable inter-user interference. With a proper scheduling algorithm, the system can exploit multi-user diversity by riding on the peaks of the channel fading and maintain a certain relative user fairness over time [141, 284].

This subsection provides a short introduction to recent approaches for coordinated multi-cell multi-antenna scheduling; we refer to [197] for a recent tutorial on scheduling in single-antenna systems. As indicated above, scheduling can be represented as setting $w_k = 0$ for users that should be inactive in the current resource allocation. Roughly speaking, multi-cell multi-antenna scheduling is based on two factors: (1) User-specific application requirements (e.g., constraints on delay and average throughput); and (2) Spatial separability among users. The first part is very application-dependent but can perhaps be described by a weight $\tilde{w}_k$ that represents the urgency and a QoS request $\gamma_k = g_k(\mathrm{SINR}_k)$ that represents an acceptable performance level for the user. The spatial separability describes the benefit of selecting users with either near-orthogonal channel vectors or weak channel gains from each other's transmitters, since this will automatically limit the inter-user interference without the need for intricate beamforming selection. Conversely, it is probably beneficial to allocate orthogonal time/frequency slots to users with very similar spatial signatures[17] because beamforming and power control cannot separate them adequately [52].

Scheduling is conceptually similar to clustering, but is updated at a much smaller time-scale to achieve continuous fairness among all users and adapt to changes in the separability due to small-scale channel fading. The combinatorial nature of scheduling makes it practically infeasible to consider all possibilities; there are $2^{K_r}$ different ways to select a subset of $K_r$ users. Fortunately, greedy algorithms that iteratively select the user that provides the largest improvement in system utility

[17] Recall that the performance region is convex when considering two well-separated users, while two users with similar spatial signatures give a concave regions. Roughly speaking, scheduling represents the removal of users such that the remaining performance region becomes increasingly convex and keeps much of its volume.

seem to provide close-to-optimal performance in single-cell applications [64, 75, 240]. Multi-cell scheduling is more involved since the spatial separability of $\mathrm{MS}_k$ should be considered for every base station with $k \in \mathcal{C}_j$, but the asymptotic results in [83] show that simple distributed scheduling algorithms that ignore inter-cell interference can achieve the optimal high-SNR scaling in information rates. The decomposability is easily motivated if the average inter-cell interference power is independent of the user location [134], which turns out to be a relatively good approximation when having omni-directional single-antenna transmitters. These are encouraging results as the scalability of multi-cell systems requires some kind of distributed scheduling. A semi-distributed approach is suggested in [323], where users are jointly scheduled within each disjoint cooperation cluster, but there is no cooperation between clusters.

As with dynamic clustering, distributed multi-cell scheduling is a relatively new research area. One recent trend is the exploitation of time-correlation, which means that users that are currently scheduled are more probable than others to be good candidates in the next scheduling round. The motivation is that a user with high application requirements and/or good spatial separability will partially keep these characteristics over multiple scheduling decisions. If we assume that each base station knows the outcome of the last scheduling decision, $\mathrm{BS}_j$ can update its user selection to improve on the previous result. These updates can either be simultaneous (i.e., all base station changes their decisions in each iteration [18]) or sequential (i.e., one base station updates at a time [100, 248] or only spatially well-separated base stations make parallel updates [105]). This concept is illustrated in Figure 4.12, where $\mathrm{BS}_1$ knows which users are currently scheduled by the adjacent base stations and tries to make a spatially compatible scheduling decision. A related idea is to fix the transmit strategy and check if the system utility can be improved by changing the intended user for each beamforming vectors [307]. Such scheduling updates can be performed locally (if users report the SINR of their preferable beam, similar to what is done in random beamforming [238]) and the system can iterate between updating scheduling and updating beamforming vectors for the currently active users. We refer to the

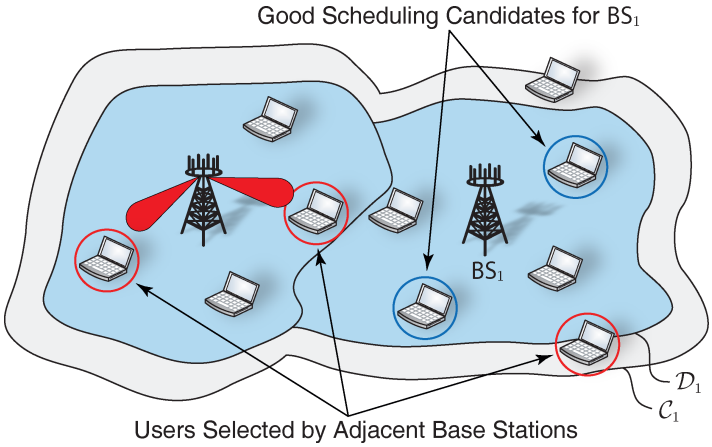



Fig. 4.12 Illustration of iterative multi-cell scheduling that utilizes time-correlation. $BS_1$ selects users that are compatible with the current scheduling decisions in the adjacent cells.

recent works of [18, 100, 105, 248, 307] for further details and note that we are not aware of any evaluation of distributed scheduling algorithms under practical conditions.

## 4.8 Cognitive Radio Systems

In a cognitive radio scenario, the systems are capable of detecting their environment and reconfiguring their operations accordingly. These capabilities are feasible due to measuring and feedback mechanisms in the network [102]. Consider a network composed of licensed *primary users*. The offered radio resources might not be utilized completely by these systems such that more users can be supported in the network. Additional users, having cognitive radio capabilities, can be also supported in the network. These users are called *secondary users*,[18] and they can use the resources licensed to the primary users under the condition of not imposing quality-of-service (QoS) degradations to these systems.

[18] Note that primary and secondary refers to the role in the network and the corresponding priorities whereas *cognitive* and *noncognitive* (sometimes called legacy) refers to the capabilities of the links.

In the following, we introduce the most relevant models of coexistence of cognitive radio systems and legitimate systems: interweave, underlay, and overlay cognitive radio [90].

An *interweave* cognitive radio is an intelligent wireless communication system that periodically monitors the radio spectrum, intelligently detects occupancy in the different parts of the spectrum, and then opportunistically communicates over spectrum holes with minimal interference to the active primary users. Cognitive users transmit simultaneously with a noncognitive user only in the event of a false spectral hole detection. The transmit power of the secondary system is limited by the range of its spectral hole sensing. In normal operation, this type of cognitive radio system does not lead to the multi-cell system model (with dynamic cooperation clusters) introduced in Section 1.2 between the primary and cognitive system. However, for the cognitive links, our multi-cell framework can be applied.

The *underlay* paradigm mandates that concurrent noncognitive and cognitive transmissions may occur only if the interference generated by the cognitive devices at the noncognitive receivers is below some acceptable threshold. The interference constraint for the noncognitive users may be met by using multiple antennas to guide the cognitive signals away from the noncognitive receivers, or by using a wide bandwidth over which the cognitive signal can be spreaded below the noise floor, then despreaded at the cognitive receiver. In both cases, the interference created by the cognitive transmitter to the primary user as well as the interference from the primary transmitter at the cognitive receiver can be described by the multi-cell framework of this tutorial. The special characteristics are the interference temperature constraints (ITC) and the assumptions on the cooperation between legacy and cognitive system. This scenario was shown in Example 1.4.

Finally, cognitive radio systems that have cooperation with the primary system as key feature are typically denoted as *overlay* cognitive radio systems. In general, spectrum overlay refers to the situation where the primary system changes its transmission strategy to involve the secondary system and to set up cooperation. Cooperation between the primary and secondary system can be established, for example, on the transmitter side or the receiver side. Again, this leads to the multi-cell

system model of this tutorial with specific requirements on the ITC at the primary receiver and on the adaptivity and cooperation.

In this subsection, we focus on the underlay cognitive radio system and begin with null-shaping constraints [120] followed by general ITC [170]. However, we note that the results from this tutorial could be also applied in overlay cognitive radio systems with cooperation between primary and cognitive transmitter [169].

The ITCs are distinguished in soft- and peak-power-shaping constraints [229]. These constraints refer to the maximum average power and average peak power tolerated at the primary receivers, respectively. In our case, the two types of constraints are equivalent since we consider only single-stream beamforming (motivated by Theorem 1.8). Reference [327] considers the setting of a single secondary transmitter sharing the same spectral band with multiple primary users. The authors provide optimal transmit strategies under ITCs for the secondary transmitter. Moreover, convex optimization techniques for solving cognitive radio problems are studied in [326].

We will focus and elaborate on the scenario described in Example 1.4. The $K_{\text{primary}}$ soft-shaping constraints $\mathbf{Q}_{kl} = \mathbf{D}_k \mathbf{C}_l \mathbf{h}_l \mathbf{h}_l^H \mathbf{C}_l \mathbf{D}_k$ are assumed to be null-shaping constraints (i.e., $\mathbf{v}_k^H \mathbf{Q}_{kl} \mathbf{v}_k = 0 \, \forall l \in \mathcal{K}_{\text{primary}}$). We collect all null-shaping constraints for transmission to $\text{MS}_k$ in a matrix

$$\mathbf{Z}_k = \left[ \mathbf{Q}_{k1} \dots \mathbf{Q}_{kK_{\text{primary}}} \right]. \tag{4.75}$$

In order to satisfy the null-shaping constraints, we can define a new effective channel of $\text{MS}_k$ as

$$\tilde{\mathbf{h}}_k = \mathbf{\Pi}_{\mathbf{Z}_k}^{\perp} \mathbf{h}_k \tag{4.76}$$

by projecting the original channel vector $\mathbf{h}_k$ onto the null-space of the null-shaping matrix $\mathbf{Z}_k$. Based on the effective channels $\tilde{\mathbf{h}}_k$ in (4.76), the complete framework developed in this tutorial can be applied. The achievable performance region shrinks compared with the case without null-shaping constraints. This is visualized for two secondary users in Figure 4.13, using $N = 3$ transmit antennas and an SNR of 10 dB.

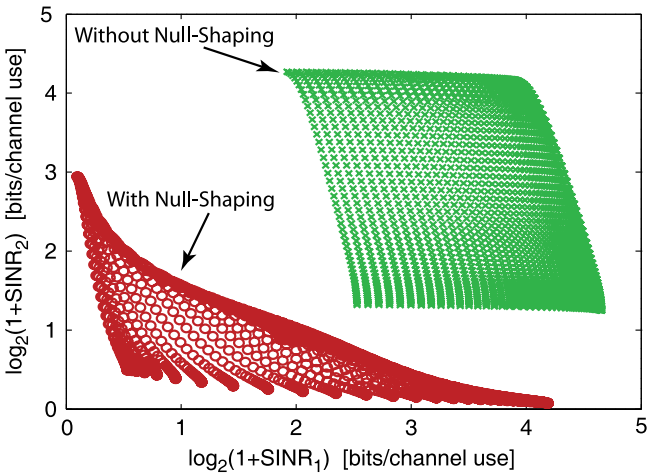



Fig. 4.13 Illustration of the rate region for two secondary users and the reduction imposed by null-shaping constraints. Sample points of the region without null-shaping constraints are marked with crosses, while sample points of the region with null-shaping constraints are marked with circles.

Another interesting question concerns the performance of multiple noncooperating cognitive transmitter–receiver pairs under null-shaping constraints. It can be shown that by properly selecting the null-shaping constraints, it is possible to achieve all points on the Pareto boundary of the corresponding performance region. We have the following result from [180, Corollary 1].

---

**Corollary 4.13** Assume that the number of antennas at each secondary transmitter $j$ is larger than the total number of secondary users (i.e., $N_j \geq K_r \; \forall j$). Construct the null-shaping constraint matrix as

$$W\mathbf{Z}_k(\boldsymbol{\lambda}_k) = \left[\mathbf{z}_1^k(\boldsymbol{\lambda}_k) \ldots \mathbf{z}_{K_{\text{primary}}}^k(\boldsymbol{\lambda}_k)\; \mathbf{z}_{N_j}^k(\boldsymbol{\lambda}_k)\right], \tag{4.77}$$

where

$$\begin{aligned} \mathbf{z}_i^k(\boldsymbol{\lambda}_k) &= \boldsymbol{v}_i \left( \sum_{i=1}^{K_r} \lambda_{ki} e_{ki} \mathbf{D}_k^H \mathbf{C}_i^H \mathbf{h}_i \mathbf{h}_i^H \mathbf{C}_i \mathbf{D}_k \right), \\ e_{ki} &= \begin{cases} +1, & k = i, \\ -1, & k \neq i, \end{cases} \end{aligned} \tag{4.78}$$

with $\boldsymbol{v}_i$ as the unit-norm eigenvector corresponding to the $i$th strongest eigenvalue and $\boldsymbol{\lambda}_k = [\lambda_{k1} \ldots \lambda_{kK_r}]^T \in \mathbb{R}_+^{K_r}$ (with $\sum_{i=1}^{K_r} \lambda_{ki} = 1$).

All points on the Pareto boundary of the performance region can be achieved with (noncooperative) beamforming directions

$$\bar{\mathbf{v}}_k(\boldsymbol{\lambda}_k) = \frac{\boldsymbol{\Pi}^{\perp}_{\mathbf{Z}_k(\boldsymbol{\lambda}_k)} \mathbf{D}_k^H \mathbf{C}_k^H \mathbf{h}_k}{\|\boldsymbol{\Pi}^{\perp}_{\mathbf{Z}_k(\boldsymbol{\lambda}_k)} \mathbf{D}_k^H \mathbf{C}_k^H \mathbf{h}_k\|}. \tag{4.79}$$

Finally, we relax the null-shaping constraints to general ITC and consider one primary transmitter–receiver pair and one secondary transmitter–receiver pair. The ITC is given by $\mathbf{v}_2^H \mathbf{Q}_{21} \mathbf{v}_2 \leq q_1$, where the limit $q_1 \geq 0$ can be selected in different ways. For example, it can be related to the loading factor of the primary system; that is,

$$R_1(\texttt{load}) = \texttt{load} \cdot \log_2 \left( 1 + \frac{|\mathbf{h}_1^H \mathbf{C}_1 \mathbf{D}_1 \mathbf{v}_1|^2}{\sigma_1^2} \right), \tag{4.80}$$

where `load` is the loading factor between zero and one (one means 100% load). Suppose an information rate of $R_1(\widetilde{\texttt{load}}) > 0$ is required to support the QoS of the primary system, then the resulting ITC limit is

$$q_1 = |\mathbf{h}_1^H \mathbf{C}_1 \mathbf{D}_1 \mathbf{v}_1|^2 \left(2^{R_1(\widetilde{\texttt{load}})} - 1\right)^{-1} - \sigma_1^2. \tag{4.81}$$

The optimization problem for the cognitive transmitter is to maximize the performance of the cognitive user while satisfying the ITC and the power constraint

$$\underset{\mathbf{v}_2 : \|\mathbf{v}_2\| \leq 1}{\text{maximize}} \; g_2(\text{SINR}_2(\mathbf{v}_2)) \quad \text{subject to } \mathbf{v}_2^H \mathbf{Q}_{21} \mathbf{v}_2 \leq q_1. \tag{4.82}$$

Using the characterization in Example 3.1, the solution to (4.82) can be given in closed form, as shown in [170, Proposition 1].

**Theorem 4.14** The optimization problem (4.82) is solved by

$$\mathbf{v}_2(\lambda^*) = \sqrt{\lambda^*} \frac{\boldsymbol{\Pi}_{\mathbf{D}_2^H \mathbf{C}_1^H \mathbf{h}_1} \mathbf{D}_2^H \mathbf{C}_2^H \mathbf{h}_2}{\|\boldsymbol{\Pi}_{\mathbf{D}_2^H \mathbf{C}_1^H \mathbf{h}_1} \mathbf{D}_2^H \mathbf{C}_2^H \mathbf{h}_2\|} + \sqrt{1-\lambda^*} \frac{\boldsymbol{\Pi}^{\perp}_{\mathbf{D}_2^H \mathbf{C}_1^H \mathbf{h}_1} \mathbf{D}_2^H \mathbf{C}_2^H \mathbf{h}_2}{\|\boldsymbol{\Pi}^{\perp}_{\mathbf{D}_2^H \mathbf{C}_1^H \mathbf{h}_1} \mathbf{D}_2^H \mathbf{C}_2^H \mathbf{h}_2\|} \tag{4.83}$$

with

$$\lambda^* = \begin{cases} \lambda_{\text{MRT}}, & \lambda_{\text{MRT}} \leq \frac{q_1}{\|\mathbf{D}_2^H \mathbf{C}_1^H \mathbf{h}_1\|^2}, \\ \frac{q_1}{\|\mathbf{D}_2^H \mathbf{C}_1^H \mathbf{h}_1\|^2}, & \text{otherwise}, \end{cases} \tag{4.84}$$

and $\lambda_{\text{MRT}} = \dfrac{\left\|\boldsymbol{\Pi}_{\mathbf{D}_2^H \mathbf{C}_1^H \mathbf{h}_1} \mathbf{D}_2^H \mathbf{C}_2^H \mathbf{h}_2\right\|^2}{\left\|\boldsymbol{\Pi}_{\mathbf{D}_2^H \mathbf{C}_1^H \mathbf{h}_1} \mathbf{D}_2^H \mathbf{C}_2^H \mathbf{h}_2\right\|^2 + \left\|\boldsymbol{\Pi}^{\perp}_{\mathbf{D}_2^H \mathbf{C}_1^H \mathbf{h}_1} \mathbf{D}_2^H \mathbf{C}_2^H \mathbf{h}_2\right\|^2}$.

Note that with $\texttt{load} = 1$, the solution in Theorem 4.14 reduces to the case with null-shaping constraints.

## 4.9 Physical Layer Security

The data processing, transmission, and encryption in modern communication systems are carried out separately. The typical purpose of the physical layer is to guarantee error-free transmission, whereas encryption is performed at a higher layer in the protocol stack. State-of-the-art encryption algorithms rely on mathematical operations assumed to be hard to compute, however, the classical approach to security becomes increasingly difficult to justify, in particular if we consider that: (a) the underlying intractability assumptions may be wrong; (b) efficient attacks could be developed; (c) the advent of quantum computers is likely to compromise this type of encryption; and (d) fast and reliable communications over *ad hoc* wireless networks require light and effective security architectures.[19]

Information-theoretic results provide an alternative approach by exploiting the randomness of physical communication channels. By proper physical layer design, the network can actually guarantee that the sent messages cannot be decoded by a third party, maliciously eavesdropping on the wireless medium. Shannon pioneered to study the notion of perfect secrecy in his seminal paper [237]. Later, the theoretical basis for an information-theoretic approach was laid by Wyner [298] and Csiszár and Körner [58], who proved that channel codes exist which guarantee both reliability and a prescribed degree of data confidentiality. A good overview of the topic of secrecy on the physical layer

[19] *Light* means that no infrastructure access is required to exchange and manage key pairs.

(including single- and multi-user systems as well as single- and multi-antenna systems) is given in [27, 121, 147]. Furthermore, an overview on current research problems and applications in the field of physical layer secrecy is provided by [156].

In systems with multiple transmit and receive antennas, the spatial degrees-of-freedom provide optimization possibilities for secure transmission as well as clever eavesdropping. In [87], artificial noise is created at the transmitter and relays to ensure secrecy. The secrecy capacity region of MIMO broadcast channels is characterized in [6, 155]. A closed-form expression for the secrecy capacity of the single-user MISO channel is derived in [232]. The corresponding transmit optimization for achieving the secrecy capacity on single-user MIMO channels is more difficult, but a numerical algorithm based on global optimization is proposed in [152].

In this subsection, we exemplify a scenario with four entities (or two links in the framework described in Section 1); see Figure 4.14 [116]. The channel between Alice and Bob is the intended communication link. Another single-antenna node called Eve is trying to overhear the

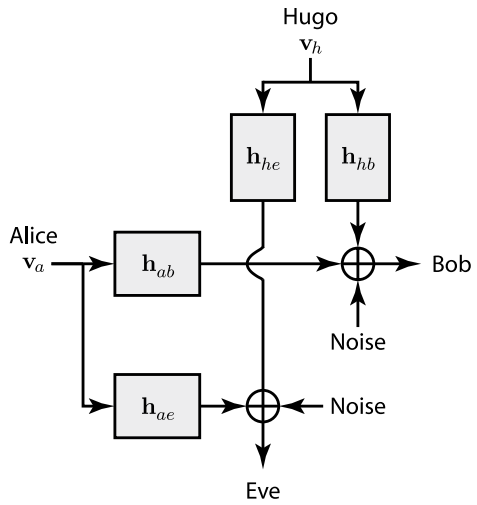



Fig. 4.14 Illustration of a simple eavesdropping scenario with four entities [116]. Alice wants to communicate privately with Bob. Eve is trying to overhear, while Hugo supports the private communication by intentionally creating interference at Eve (while avoiding interference at Bob).

communication between Alice and Bob. Finally, a helper called Hugo equipped with multiple transmit antennas supports the private communication. We basically have two transmissions creating interference to each other: from Alice to Bob and from Hugo to Eve. Following the notation from previous sections, the link from Alice to Bob is described by the vector channel $\mathbf{h}_{ab} = \mathbf{D}_1^H \mathbf{C}_1^H \mathbf{h}_1$, Alice to Eve $\mathbf{h}_{ae} = \mathbf{D}_1^H \mathbf{C}_2^H \mathbf{h}_2$, whereas the link from Hugo to Bob is described by $\mathbf{h}_{hb} = \mathbf{D}_2^H \mathbf{C}_1^H \mathbf{h}_1$ and Hugo to Eve is $\mathbf{h}_{he} = \mathbf{D}_2^H \mathbf{C}_2^H \mathbf{h}_2$. The private communication uses the beamforming vector $\mathbf{v}_a$ and the helper creates interference using $\mathbf{v}_h$. For notational simplicity, the noise variances are normalized toward the channel vectors.

The achievable *secrecy rate* for reliable and secure data transmission between Alice and Bob is given by

$$R_S(\mathbf{v}_a, \mathbf{v}_h) = \left[ \log_2 \left( 1 + \frac{|\mathbf{h}_{ab}^H \mathbf{v}_a|^2}{1 + |\mathbf{h}_{hb}^H \mathbf{v}_h|^2} \right) - \log_2 \left( 1 + \frac{|\mathbf{h}_{ae}^H \mathbf{v}_a|^2}{1 + |\mathbf{h}_{he}^H \mathbf{v}_h|^2} \right) \right]_+ . \tag{4.85}$$

The optimization problem for maximizing the secrecy rate is given by

$$\max_{0 \leq \|\mathbf{v}_a\|^2 \leq q_a} \max_{0 \leq \|\mathbf{v}_h\|^2 \leq q_h} R_S(\mathbf{v}_a, \mathbf{v}_h). \tag{4.86}$$

The outer optimization problem in (4.86) for fixed beamforming vector $\mathbf{v}_h$ is solved similar to [232] and [6, Section V].

---

**Lemma 4.15** The beamforming vector $\mathbf{v}_a'$ that solves (4.86) for fixed $\mathbf{v}_h$ is given by

$$\mathbf{v}_a'(\mathbf{v}_h) = q_a \boldsymbol{\psi}, \tag{4.87}$$

where $\boldsymbol{\psi}$ is the generalized eigenvector associated with the maximum generalized eigenvalue of the pencil $(\mathbf{I} + \frac{1}{1+z_1} \mathbf{h}_{ab} \mathbf{h}_{ab}^H, \mathbf{I} + \frac{1}{1+z_2} \mathbf{h}_{ae} \mathbf{h}_{ae}^H)$ with $z_1 = |\mathbf{h}_{hb}^H \mathbf{v}_h|^2$ and $z_2 = |\mathbf{h}_{he}^H \mathbf{v}_h|^2$.

---

The inner optimization problem in (4.86) for a fixed beamforming vector $\mathbf{v}_a$ cannot be solved in closed form because the terms in the denominator cannot be transformed into a simple quotient. However, the beamforming parametrization in Example 3.1 can be applied to describe the optimal beamforming vector at the helper Hugo.

**Theorem 4.16** For fixed $\mathbf{v}_a$, the beamforming vector $\mathbf{v}'_h$ that solves (4.86) is given by

$$\mathbf{v}'_h(\lambda) = \frac{\lambda \mathbf{\Pi}_{\mathbf{h}_{hb}} \mathbf{h}^*_{he} + (1-\lambda)\mathbf{\Pi}^{\perp}_{\mathbf{h}_{hb}} \mathbf{h}^*_{he}}{\|\lambda \mathbf{\Pi}_{\mathbf{h}_{hb}} \mathbf{h}^*_{he} + (1-\lambda)\mathbf{\Pi}^{\perp}_{\mathbf{h}_{hb}} \mathbf{h}^*_{he}\|} \tag{4.88}$$

for some $\lambda \in [0,1]$.

Assume next that there are $K$ helpers and we denote the channels from helper $k \in \{1,\ldots,K\}$ to Bob as $\mathbf{h}_{kb}$ and to Eve as $\mathbf{h}_{ke}$, respectively. In order to find the optimal transmit strategies at the helpers and at Alice, an iterative approach is described in Algorithm 6. In the algorithm, we define $\mathbf{v}_{-k}(\lambda'_{-k}) = [\mathbf{v}_1(\lambda'_1) \ldots \mathbf{v}_{k-1}(\lambda'_{k-1})\, \mathbf{v}_{k+1}(\lambda'_{k+1}) \ldots \mathbf{v}_K(\lambda_K)]$. Algorithm 6 converges to the global optimum because both steps in the while loop yield unique solutions, the objective function is maximized in each step and there is an upper bound to the objective function given by the peaceful system without any eavesdropper

$$R_S(\mathbf{v}_a, \mathbf{v}_h, \mathbf{v}_1, \ldots, \mathbf{v}_K) \leq \log_2\left(1 + q_a \|\mathbf{h}_{ab}\|^2\right). \tag{4.89}$$

For illustration, we consider the case in which all channel vectors are independent and identically distributed according to a zero-mean complex Gaussian distribution with identity covariance matrices. Figure 4.15 shows the average secrecy rate with and without a helper and using different beamforming strategies:

(1) Upper bound (4.89): Peaceful information rate without Eve.
(2) Optimal beamforming using Algorithm 6.
(3) Alice performs optimal beamforming without a helper.
(4) Alice performs MRT and the helper uses ZFBF.
(5) Alice performs MRT without having a helper.

We can make several observations from Figure 4.15. First, the gap between the naive system where Alice performs MRT while being eavesdropped compared to the peaceful system is significant and increases with the SNR (i.e., the best and worst curves in Figure 4.15). Second, the optimal transmit strategy in (4.87) without any helper performs

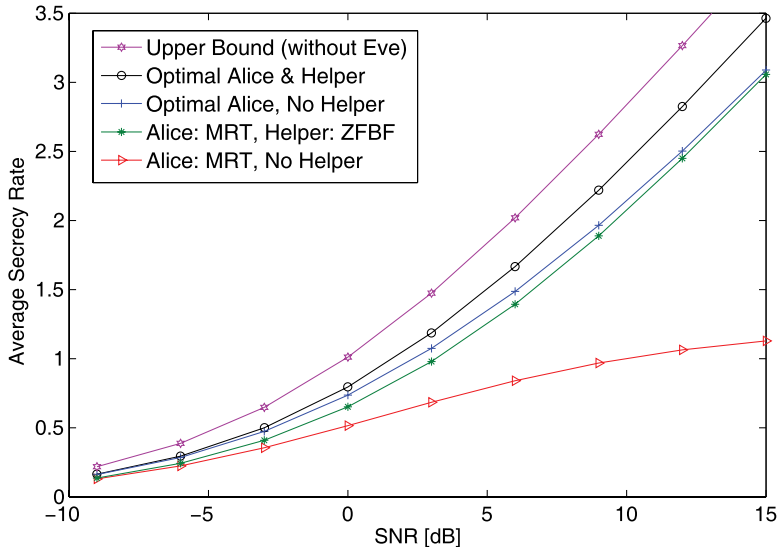



Fig. 4.15 Average secrecy rate with and without a helper, and using different transmit strategies and the upper bound from (4.89). Alice and Hugo have two transmit antennas each.

**Algorithm 6**: Secrecy Rate Maximization with $K$ Helpers

**Result**: Find optimal beamforming vectors $\mathbf{v}_a$ at Alice and optimal helper beamforming $\mathbf{v}_1, \ldots, \mathbf{v}_K$.

**Input**: Channel vectors $\mathbf{h}_{ab}, \mathbf{h}_{ae}, \mathbf{h}_{kb}, \mathbf{h}_{ke}$ for $k = 1, \ldots, K$;

Set $\mathbf{v}'_a = \frac{\mathbf{h}_{ab}}{\|\mathbf{h}_{ab}\|}$ and $\mathbf{v}_1 = \ldots = \mathbf{v}_K = \mathbf{0}$;

**while** *required accuracy not reached* **do**

  **for** $k = 1 : K$ **do**

    $\lambda'_k = \arg \max_{0 \le \lambda \le 1} R_S(\mathbf{v}'_a, \mathbf{v}_k(\lambda), \mathbf{v}_{-k}(\lambda'_{-k}))$;

  $\mathbf{v}'_a(\mathbf{v}_1(\lambda'), \ldots, \mathbf{v}_K(\lambda')) = q_a \boldsymbol{\psi}$ with $\boldsymbol{\psi}$ from (4.87);

**Output**: Optimal beamforming vectors;

reasonably well. If Alice does not adapt to the eavesdropper channel, the helper can almost compensate for it. But the real benefits of having a helper are seen when both Alice and Hugo optimize their transmissions. The average SNR gap between the iterative beamforming solution (Algorithm 6) and the upper bound is about 1.5 dB.

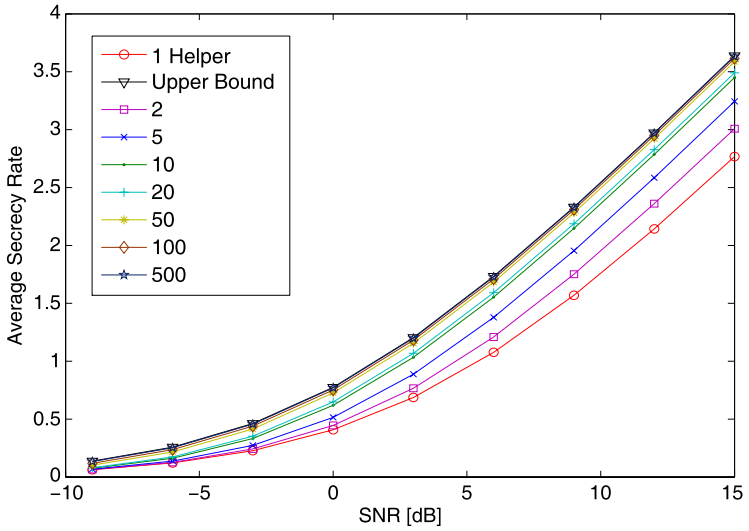



Fig. 4.16 Instantaneous secrecy rate with helpers using optimal beamforming from Algorithm 6. Alice and all helpers are equipped with two transmit antennas.

The last observation leads to the question whether multiple helpers can reduce the gap. We investigate this by having different numbers of helpers with independent and identically distributed channels according to zero-mean complex Gaussian distribution with identity covariance matrices. In order to show the behavior with large (unrealistic) number of helpers, fixed channel vectors $\mathbf{h}_{ab}$ and $\mathbf{h}_{ae}$ are used with fixed channel vectors for varying number of helpers in Figure 4.16. It can be observed that the gap between the upper bound (4.89) and the secrecy rate with helpers reduces with increasing number of helpers. The (unrealistic) case with $K = 500$ helpers achieves a secrecy rate which cannot be distinguished from the upper bound.

Note that we assumed that the transmit strategies are chosen jointly for Alice and Hugo. This requires a central authority to decide on $\lambda$ in (4.88), thus CSI and SNRs need to be available to run Algorithm 6.

A closer look at the optimal beamforming vector $\mathbf{v}_a$ at Alice in (4.87) shows that Alice needs only her own channel vectors $\mathbf{h}_{ab}$ and $\mathbf{h}_{ae}$ and the interference terms $z_1$ and $z_2$ at Bob and Eve, respectively, to compute the generalized eigen decomposition. Bob will voluntarily feedback the SNR $z_1$. In a cellular context, where Eve is an internal

eavesdropper, who behaves well but curiously, the SNR information $z_2$ is also available.

In order to compute the optimal beamforming vector $\mathbf{v}_h$ at Hugo, only information about the own channels $\mathbf{h}_{hb}$ and $\mathbf{h}_{he}$ is required plus the weighting parameter $\lambda$. The parameter selection depends again on the helper model. If Hugo is part of the cellular network, control information such as $\lambda$ can be sent from Alice and the centralized optimization in Algorithm 6 is well motivated. For further discussions on the simple helper scenario considered, please refer to [116].

# Acknowledgments

Emil Björnson thanks the distinguished researchers Björn Ottersten, Mats Bengtsson, David Gesbert, Gan Zheng, and Per Zetterberg for the inspiration and guidance they have provided over the years. This tutorial is the final result of his doctoral journey at the Signal Processing Lab at KTH Royal Institute of Technology. He would also like to express his gratitude to the fellow doctoral students David Hammarwall, Niklas Jaldén, Simon Järmyr, Randa Zakhour, Xueying Hou, and Jinghong Yang for fruitful scientific collaborations.

Eduard Jorswieck thanks his colleagues Erik G. Larsson, Eleftherios Karipidis, Rami Mochaourab, Zuleita Ka Ming Ho, Jan Sykora, Leonardo Badia, Jian Luo, Martin Schubert, David Gesbert, and all members of the SAPHYRE project for interesting and fruitful discussions on the topic of resource sharing in wireless networks.

We are indebted to Rasmus Brandt, Axel Müller, Serveh Shalmashi, and the anonymous reviewers for their careful proofreading.

This work was supported in part by the AMIMOS project under the Advanced Research Grant 228044 from the European Research Council (ERC). It was also supported by the International Postdoc Grant 2012-228 from The Swedish Research Council. Part of this work

was performed in the framework of the European research project SAPHYRE, which was supported in part by the European Union under its FP7 ICT Objective 1.1 — The Network of the Future. This work was also supported in part by the Deutsche Forschungsgemeinschaft (DFG) under Grant Jo 801/4-1.

# Notations and Acronyms

## Mathematical Notations

Upper-case boldface letters are used to denote matrices (e.g., $\mathbf{X}, \mathbf{Y}$), while (column) vectors are denoted with lower-case boldface letters (e.g., $\mathbf{x}, \mathbf{y}$). Scalars are denoted by italic letters (e.g., $X, Y$) and sets by calligraphic letters (e.g., $\mathcal{X}, \mathcal{Y}$). The following mathematical notations are used:

| | |
|---|---|
| $\mathbb{C}^{N\times M}$ | The set of complex-valued $N \times M$ matrices. |
| $\mathbb{R}^{N\times M}$ | The set of real-valued $N \times M$ matrices. |
| $\mathbb{C}^N, \mathbb{R}^N$ | Short forms of $\mathbb{C}^{N\times 1}$ and $\mathbb{R}^{N\times 1}$. |
| $\mathbb{R}_+^N$ | The set of non-negative members of $\mathbb{R}^N$. |
| $\emptyset$ | The empty set. |
| $x \in \mathcal{S}$ | $x$ is a member of the set $\mathcal{S}$. |
| $x \notin \mathcal{S}$ | $x$ is not a member of the set $\mathcal{S}$. |
| $\mathcal{S}_1 \subseteq \mathcal{S}_2$ | $\mathcal{S}_1$ is included in (is a subset of) $\mathcal{S}_2$. |
| $\mathcal{S}_1 \cup \mathcal{S}_2$ | Union set with all members in $\mathcal{S}_1$ and/or $\mathcal{S}_2$. |
| $\mathcal{S}_1 \cap \mathcal{S}_2$ | Intersection set with all members which are in *both* $\mathcal{S}_1$ and $\mathcal{S}_2$. |
| $\mathcal{S}_1 \times \mathcal{S}_2$ | The Cartesian product of sets $\mathcal{S}_1$ and $\mathcal{S}_2$. |

| | |
|---|---|
| $\mathcal{S} \setminus \{x\}$ | The remaining set when member $x$ is removed. |
| $\lvert\mathcal{S}\rvert$ | The cardinality (i.e., number of members) of a set $\mathcal{S}$. |
| $\forall x$ | Means that a statement holds for all $x$ (in the set that $x$ belongs to). |
| $\{x \in \mathcal{S} : P\}$ | The set of all member of $\mathcal{S}$ having a property P. |
| $f : \mathcal{S}_1 \to \mathcal{S}_2$ | Function from $\mathcal{S}_1$ to $\mathcal{S}_2$. |
| $f^{-1}$ | Inverse function of a function $f$. |
| $x_i = [\mathbf{x}]_i$ | Two ways of writing the $i$th element of a vector $\mathbf{x}$. |
| $x_{ij} = [\mathbf{X}]_{ij}$ | Two ways of writing the $i,j$th element of a matrix $\mathbf{X}$. |
| $\mathrm{diag}(\cdot)$ | $\mathrm{diag}(x_1,\ldots,x_N)$ is a diagonal matrix with $x_1,\ldots,x_N$ at the diagonal. $\mathrm{diag}(\mathbf{X}_1,\ldots,\mathbf{X}_N)$ is block-diagonal. |
| $\mathbf{X}^T$ | The transpose of $\mathbf{X}$. |
| $\mathbf{X}^H$ | The conjugate transpose of $\mathbf{X}$. |
| $\mathbf{X}^{-1}$ | The inverse of a square matrix $\mathbf{X}$. |
| $\mathbf{X}^\dagger$ | The Moore–Penrose pseudo-inverse of $\mathbf{X}$. |
| $\boldsymbol{\Pi}_{\mathbf{X}}$ | The orthogonal projection matrix onto the column space of $\mathbf{X}$ (i.e., $\boldsymbol{\Pi}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}^H\mathbf{X})^\dagger\mathbf{X}^H$). |
| $\boldsymbol{\Pi}^\perp_{\mathbf{X}}$ | Projection matrix onto the orthogonal complement of the column space of $\mathbf{X}$ (i.e., $\boldsymbol{\Pi}^\perp_{\mathbf{X}} = \mathbf{I} - \boldsymbol{\Pi}_{\mathbf{X}}$). |
| $[\mathbf{x}]_+$ | Obtained from $\mathbf{x}$ by setting negative entries to zero. |
| $\mathrm{sign}(x)$ | Sign of a real-valued number $x$. |
| $\Re(x)$ | Real part of a scalar $x$. |
| $\Im(x)$ | Imaginary part of a scalar $x$. |
| $\imath$ | The imaginary number. |
| $\lvert x\rvert$ | Absolute value of a scalar $x$. |
| $\angle x$ | Phase of a complex-valued scalar $x$. |
| $\lceil x \rceil$ | The smallest integer not less than the scalar $x \in \mathbb{R}$. |
| $\log_a(x)$ | Logarithm of $x$ using the base $a \in \mathbb{R}_+$. |
| $\mathcal{O}(\cdot)$ | Big O notation: $f(x) = \mathcal{O}(g(x))$ means that it exist $c, x_0 \in \mathbb{R}_+$ such that $\lvert f(x)\rvert \leq c\lvert g(x)\rvert$ for $x > x_0$. |
| $\boldsymbol{v}_{\max}(\mathbf{X})$ | Eigenvector associated with the largest eigenvalue. |
| $\boldsymbol{v}_i(\mathbf{X})$ | Eigenvector associated with the $i$th largest eigenvalue. |
| $\mathrm{tr}(\mathbf{X})$ | Trace of a square matrix $\mathbf{X}$. |
| $\mathrm{rank}(\mathbf{X})$ | Rank of a matrix $\mathbf{X}$ (i.e., nonzero singular values). |
| $\mathrm{span}(\mathbf{X})$ | The column space of a matrix $\mathbf{X}$. |

| | |
|---|---|
| $\nabla f(\mathbf{x})$ | The gradient vector of a scalar function $f$. |
| $\mathcal{N}(\mathbf{x}, \mathbf{R})$ | The multivariate Gaussian distribution with mean $\mathbf{x}$ and covariance matrix $\mathbf{R}$. |
| $\mathcal{CN}(\mathbf{x}, \mathbf{R})$ | The circularly symmetric complex Gaussian counterpart. |
| $x \sim \mathcal{X}(\cdot)$ | The random variable $x$ has distribution $\mathcal{X}(\cdot)$. |
| $\mathbb{E}\{\mathbf{X}\}$ | The mathematical expectation of a stochastic $\mathbf{X}$. |
| $\|\mathbf{x}\|_p$ | The $L_p$-norm $\|\mathbf{x}\|_p = (\sum_i \|x_i\|^p)^{1/p}$ of $\mathbf{x}$. |
| $\|\mathbf{X}\|_F$ | The Frobenius norm $\|\mathbf{X}\|_F = \sqrt{\sum_{i,j} \|x_{ij}\|^2}$ of $\mathbf{X}$. |
| $\mathbf{X} \succ \mathbf{Y}$ | Means $\mathbf{X} - \mathbf{Y}$ is positive definite. |
| $\mathbf{X} \succeq \mathbf{Y}$ | Means $\mathbf{X} - \mathbf{Y}$ is positive semi-definite. |
| $\mathbf{x} > \mathbf{y} (\mathbf{x} \geq \mathbf{y})$ | Means $x_i > y_i$ ($x_i \geq y_i$) for all vector indices $i$. |
| $\mathbf{x} \geq^{\mathbf{e}} \mathbf{y}$ | Means $x_i e_i \geq y_i e_i$ with at least one strict inequality. |
| $\mathbf{I}_N$ | The $N \times N$ identity matrix. |
| $\mathbf{1}_N$ | The $N \times 1$ matrix (i.e., vector) of only ones. |
| $\mathbf{0}_N$ | The $N \times N$ matrix of only zeros. |
| $\mathbf{0}_{N \times M}$ | The $N \times M$ matrix of only zeros. |

## Tutorial Specific Notations

Symbols and functions that are commonly used in the tutorial are summarized as follows:

| | |
|---|---|
| $\text{BS}_j$ | Base station $j$. |
| $\mathcal{C}_j$ | Set of users that $\text{BS}_j$ coordinates interference to. |
| $\mathbf{C}_k$ | Diagonal matrix such that $\mathbf{h}_k^H \mathbf{C}_k$ is the channel that carries nonnegligible interference to user $k$. |
| $\mathbf{C}_{jk}$ | Equal to $\mathbf{I}_{N_j}$ if $\text{BS}_j$ coordinates interference to user $k$. |
| $\mathcal{D}_j$ | Set of users that $\text{BS}_j$ can send data to. |
| $\mathbf{D}_{jk}$ | Equal to $\mathbf{I}_{N_j}$ if $\text{BS}_j$ can send data to user $k$. |
| $\mathbf{D}_k$ | Diagonal matrix such that $\mathbf{h}_k^H \mathbf{D}_k$ is the channel for data. |
| $\delta$ | Predefined line-search accuracy. |
| $\varepsilon$ | Predefined solution accuracy for a monotonic problem. |
| $f(\cdot)$ | System utility function. |
| $g_k(\cdot)$ | Performance function of user $k$. |
| $\mathbf{h}_k$ | Channel vector from all base stations to user $k$. |

| | |
|---|---|
| $\mathbf{h}_{jk}$ | Channel component from $\mathrm{BS}_j$ to user $k$. |
| $j_k$ | Index of the master base station of user $k$. |
| $K_r$ | Number of receiving users. |
| $K_t$ | Number of transmitting base stations. |
| $L$ | Number of power constraints in the system. |
| $\mathrm{MS}_k$ | User $k$. |
| $N$ | Total number of transmit antennas in the system. |
| $N_j$ | Number of antennas at the $j$th base station. |
| $\Omega_k$ | Channel gain region corresponding to $\mathbf{S}_k$. |
| $\mathbf{Q}_{lk}$ | Weighting matrix for user $k$ in the $l$th power constraint. |
| $q_l$ | Total limit of the $l$th power constraint. |
| $q_{lk}$ | Per-user limit of the $l$th power constraint. |
| $\mathcal{R}$ | Performance region. |
| $\mathbf{S}_k$ | Signal correlation matrix for user $k$. |
| $\sigma_k^2$ | Noise variance for user $k$. |
| $\mathrm{SINR}_k$ | Signal-to-interference-and-noise ratio of user $k$. |
| $\mathbf{u}$ | Utopia point. |
| $\mathbf{v}_k$ | Beamforming vector for user $k$. |
| $\bar{\mathbf{v}}_k$ | Beamforming direction for user $k$. |
| $y_k$ | Received signal at user $k$. |

## Acronyms

The following acronyms and abbreviations are used in the tutorial:

| | |
|---|---|
| BER | Bit Error Rate |
| BRB | Branch-Reduce-and-Bound |
| c.u. | Channel Use |
| CDF | Cumulative Distribution Function |
| CoMP | Coordinated Multipoint |
| CSI | Channel State Information |
| dBm | Decibel-Milliwatt |
| DCC | Dynamic Cooperation Clusters |
| FDD | Frequency Division Duplex |
| FPO | Fairness-Profile Optimization |
| GPS | Global Positioning System |

| | |
|---|---|
| ITC | Interference Temperature Constraint |
| KKT | Karush–Kuhn–Tucker |
| LTE | 3GPP Long Term Evolution |
| MBS | Master Base Station |
| MIMO | Multiple-Input Multiple-Output |
| MISO | Multiple-Input Single-Output |
| MMSE | Minimum Mean Square Error |
| MOP | Multi-Objective Optimization Problem |
| MRT | Maximum Ratio Transmission |
| MSE | Mean Square Error |
| mW | Milliwatt |
| NP-hard | Non-Deterministic Polynomial-Time hard |
| OFDM | Orthogonal Frequency-Division Multiplexing |
| PA | Polyblock Outer Approximation |
| PAPR | Peak-to-Average Power Ratio |
| PEP | Pairwise Error Probability |
| QAM | Quadrature Amplitude Modulation |
| QoS | Quality-of-Service |
| QPSK | Quadrature Phase-Shift Keying |
| RF | Radio Frequency |
| SDMA | Spatial Division Multiple Access |
| SER | Symbol Error Rate |
| SINR | Signal-to-Interference-and-Noise Ratio |
| SISO | Single-Input Single-Output |
| SLNR | Signal-to-Leakage-and-Noise Ratio |
| SOP | Single-Objective Optimization Problem |
| SNR | Signal-to-Noise Ratio |
| TDD | Time Division Duplex |
| TDMA | Time Division Multiple Access |
| ZFBF | Zero-Forcing Beamforming |

# References


[1] *Further advancements for E-UTRA physical layer aspects (Release 9)*. 3GPP TS 36.814, March 2010.

[2] A. Aggarwal and T. Meng, "Minimizing the peak-to-average power ratio of OFDM signals using convex optimization," *IEEE Transactions on Signal Processing*, vol. 54, no. 8, pp. 3099–3110, 2006.

[3] H. Al-Shatri and T. Weber, "Achieving the maximum sum rate using D.C. programming in cellular networks," *IEEE Transactions on Signal Processing*, vol. 30, no. 3, pp. 1331–1341, 2012.

[4] V. Annapureddy and V. Veeravalli, "Sum capacity of MIMO interference channels in the low interference regime," *IEEE Transactions on Information Theory*, vol. 57, no. 5, pp. 2565–2581, 2011.

[5] M. Ariaudo, I. Fijalkow, J.-L. Gautier, M. Brandon, B. Aziz, and B. Milevsky, "Green radio despite "dirty RF" front-end," *EURASIP Journal on Wireless Communications and Networking*, 2012.

[6] G. Bagherikaram, A. S. Motahari, and A. K. Khandani, "The secrecy capacity region of the Gaussian MIMO broadcast channel," *IEEE Transactions on Information Theory*, Submitted, arXiv:0903.3261, 2009.

[7] V. Balakrishnan, S. Boyd, and S. Balemi, "Robust downlink beamforming based on outage probability specifications," *International Journal on Robust and Nonlinear Control*, vol. 1, no. 4, pp. 295–317, 1991.

[8] H. Balan, R. Rogalin, A. Michaloliakos, K. Psounis, and G. Caire, "AirSync: Enabling distributed multiuser MIMO with full spatial multiplexing," *IEEE/ACM Transactions on Networking*, Submitted, arXiv:1205.6862.

[9] A. Ben-Tal and A. Nemirovski, "Robust convex optimization," *Mathematics Of Operations Research*, vol. 23, no. 4, pp. 769–805, 1998.

[10] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*. SIAM, 2001.

[11] M. Bengtsson and B. Ottersten, "Optimal and suboptimal transmit beamforming," in *Handbook of Antennas in Wireless Communications*, (L. C. Godara, ed.), CRC Press, 2001.

[12] D. Bertsekas, *Nonlinear Programming*. Athena Scientific, 2nd ed., 1999.

[13] M. Biguesh and A. Gershman, "Downlink channel estimation in cellular systems with antenna arrays at base stations using channel probing with feedback," *EURASIP Journal on Applied Signal Processing*, no. 9, no. 9, pp. 1330–1339, 2004.

[14] E. Björnson, "Multiantenna cellular communications: Channel estimation, feedback, and resource allocation," PhD thesis, KTH Royal Institute of Technology, 2011.

[15] E. Björnson, M. Bengtsson, and B. Ottersten, "Receive combining vs. multistream multiplexing in multiuser MIMO systems," in *Proceedings of IEEE Swedish-Communication Technologies Workshop*, pp. 109–114, 2011.

[16] E. Björnson, M. Bengtsson, and B. Ottersten, "Pareto characterization of the multicell MIMO performance region with simple receivers," *IEEE Transactions on Signal Processing*, vol. 60, no. 8, pp. 4464–4469, 2012.

[17] E. Björnson, M. Bengtsson, G. Zheng, and B. Ottersten, "Computational framework for optimal robust beamforming in coordinated multicell systems," in *Proceedings of IEEE Computational Advance in Multi-Sensor Adaptive Processing*, 2011.

[18] E. Björnson, N. Jaldén, M. Bengtsson, and B. Ottersten, "Optimality properties, distributed strategies, and measurement-based evaluation of coordinated multicell OFDMA transmission," *IEEE Transactions on Signal Processing*, vol. 59, no. 12, pp. 6086–6101, 2011.

[19] E. Björnson and E. Jorswieck, "Optimal resource allocation in coordinated multi-cell systems: Matlab code," Technical Report, doi: http://dx.doi.org/10.1561/0100000069_supp.

[20] E. Björnson, M. Kountouris, M. Bengtsson, and B. Ottersten, "Receive combining vs. multi-stream multiplexing in downlink systems with multi-antenna users," *IEEE Transactions on Signal Processing*, Submitted, arXiv:1207.2776.

[21] E. Björnson and B. Ottersten, "On the principles of multicell precoding with centralized and distributed cooperation," in *Proceedings of International Conference on Wireless Communication and Signal Processing*, 2009.

[22] E. Björnson and B. Ottersten, "A framework for training-based estimation in arbitrarily correlated Rician MIMO channels with Rician disturbance," *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1807–1820, 2010.

[23] E. Björnson, R. Zakhour, D. Gesbert, and B. Ottersten, "Cooperative multicell precoding: Rate region characterization and distributed strategies with instantaneous and statistical CSI," *IEEE Transactions on Signal Processing*, vol. 58, no. 8, pp. 4298–4310, 2010.

[24] E. Björnson, P. Zetterberg, and M. Bengtsson, "Optimal coordinated beamforming in the multicell downlink with transceiver impairments," in *Proceedings of IEEE Global Communication Conference*, 2012.

[25] E. Björnson, P. Zetterberg, M. Bengtsson, and B. Ottersten, "Capacity limits and multiplexing gains of MIMO channels with transceiver impairments," *IEEE Communications Letters*, To appear, 2013.

[26] E. Björnson, G. Zheng, M. Bengtsson, and B. Ottersten, "Robust monotonic optimization framework for multicell MISO systems," *IEEE Transactions on Signal Processing*, vol. 60, no. 5, pp. 2508–2523, 2012.

[27] B. Bloch and J. Barros, *Physical Layer Security — From Information Theory to Security Engineering*. Cambridge University Press, 2011.

[28] F. Boccardi and H. Huang, "A near-optimum technique using linear precoding for the MIMO broadcast channel," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2007.

[29] H. Boche, S. Naik, and M. Schubert, "Characterization of convex and concave resource allocation problems in interference coupled wireless systems," *IEEE Transactions on Signal Processing*, vol. 59, no. 5, pp. 2382–2394, 2011.

[30] H. Boche and M. Schubert, "A general duality theory for uplink and downlink beamforming," in *Proceedings of IEEE Vehicular Technology Conference-Fall*, pp. 87–91, 2002.

[31] H. Boche and M. Schubert, "A calculus for log-convex interference functions," *IEEE Transactions on Information Theory*, vol. 54, no. 12, pp. 5469–5490, 2008.

[32] F. Bock and B. Ebstein, "Assignment of transmitter powers by linear programming," *IEEE Transactions on Electromagnetic Compatibility*, vol. 6, no. 2, pp. 36–44, 1964.

[33] M. Boldi, A. Tölli, M. Olsson, E. Hardouin, T. Svensson, F. Boccardi, L. Thiele, and V. Jungnickel, "Coordinated multipoint (CoMP) systems," in *Mobile and Wireless Communications for IMT-Advanced and Beyond*, (A. Osseiran, J. Monserrat, and W. Mohr, eds.), pp. 121–155, Wiley, 2011.

[34] G. Box and N. Draper, *Empirical Model-building and Response Surfaces*. Wiley, 1987.

[35] S. Boyd, E. F. L. El Ghaoui, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*. Society for Industrial and Applied Mathematics (SIAM), 1994.

[36] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2010.

[37] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[38] J. Branke, K. Deb, K. Miettinen, and R. S. (Eds.), *Multiobjective Optimization: Interactive and Evolutionary Approaches*. Springer, 2008.

[39] J. Brehmer, *Utility Maximization in Nonconvex Wireless Systems*. Springer, 2012.

[40] J. Brehmer and W. Utschick, "Optimal interference management in multi-antenna, multi-cell systems," in *Proceedings of International Zurich Seminar on Communications*, pp. 134–137, 2010.

[41] V. Cadambe and S. Jafar, "Interference alignment and degrees of freedom of the $k$-user interference channel," *IEEE Transactions on Information Theory*, vol. 54, no. 8, pp. 3425–3441, 2008.

[42] D. Cai, T. Quek, and C. Tan, "A unified analysis of max-min weighted SINR for MIMO downlink system," *IEEE Transactions on Signal Processing*, vol. 59, no. 8, pp. 5384–5395, 2011.

[43] D. Cai, T. Quek, C. Tan, and S. Low, "Max-min SINR coordinated multipoint downlink transmission — duality and algorithms," *IEEE Transactions on Signal Processing*, vol. 60, no. 10, pp. 5384–5395, 2012.

[44] G. Caire, N. Jindal, M. Kobayashi, and N. Ravindran, "Multiuser MIMO achievable rates with downlink training and channel state feedback," *IEEE Transactions on Information Theory*, vol. 56, no. 6, pp. 2845–2866, 2010.

[45] G. Caire, S. Ramprashad, and H. Papadopoulos, "Rethinking network MIMO: Cost of CSIT, performance analysis, and architecture comparisons," in *Proceedings of Information Theory and Applications Workshop*, 2010.

[46] G. Caire and S. Shamai, "On the achievable throughput of a multiantenna Gaussian broadcast channel," *IEEE Transactions on Information Theory*, vol. 49, no. 7, pp. 1691–1706, 2003.

[47] P. Cao, S. Shi, and E. Jorswieck, "Efficient computation of the Pareto boundary for the two-user MIMO interference channel," in *Proceedings of IEEE Workshop on Signal Processing Advances in Wireless Communications*, 2012.

[48] P. Cao, S. Shi, and E. Jorswieck, "On the Pareto boundary for the two-user single-beam MIMO interference channel," *IEEE Transactions on Signal Processing*, Submitted, arXiv:1202.5474, 2012.

[49] C.-B. Chae, I. Hwang, R. W. Heath, and V. Tarokh, "Interference aware-coordinated beamforming system in a multi-cell environment," *IEEE Transactions on Wireless Communications*, vol. 11, no. 10, pp. 3692–3703, 2012.

[50] B. Chalise, S. Shahbazpanahi, A. Czylwik, and A. Gershman, "Robust downlink beamforming based on outage probability specifications," *IEEE Transactions on Wireless Communications*, vol. 6, no. 10, pp. 3498–3503, 2007.

[51] T.-H. Chang, W.-K. Ma, and C.-Y. Chi, "Worst-case robust multiuser transmit beamforming using semidefinite relaxation: Duality and implications," in *Proceedings of Asilomar Conference on Signals, Systems, and Computers*, 2011.

[52] M. Chiang, P. Hande, T. Lan, and C. Tan, "Power control in wireless cellular networks," *Foundations and Trends in Networking*, vol. 2, no. 4, pp. 355–580, 2008.

[53] M. Chiang, S. Low, R. Calderbank, and J. Doyle, "Layering as optimization decomposition: A mathematical theory of network architectures," *Proceedings of Institute of Electrical and Electronics Engineers*, vol. 95, no. 1, pp. 255–312, 2007.

[54] D. Chizhik, J. Ling, P. Wolniansky, R. Valenzuela, N. Costa, and K. Huber, "Multiple-input-multiple-output measurements and modeling in Manhattan," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 3, pp. 321–331, 2003.

[55] S. S. Company, "General survey of radio frequency bands (30 MHz to 3 GHz): Vienna, Virginia, September 1–5," Technical Report, version 2.0, 2010.
[56] M. Costa, "Writing on dirty-paper," *IEEE Transactions on Information Theory*, vol. 29, no. 3, pp. 439–441, 1983.
[57] T. Cover and J. Thomas, *Elements of Information Theory*. Wiley, 1991.
[58] I. Csiszár and J. Körner, "Broadcast channels with confidential messages," *IEEE Transactions on Information Theory*, vol. 24, no. 3, pp. 339–348, 1978.
[59] H. Dahrouj and W. Yu, "Coordinated beamforming for the multicell multi-antenna wireless system," *IEEE Transactions on Wireless Communications*, vol. 9, no. 5, pp. 1748–1759, 2010.
[60] D. Dardari, V. Tralli, and A. Vaccari, "A theoretical characterization of nonlinear distortion effects in OFDM systems," *IEEE Transactions on Communications*, vol. 48, no. 10, pp. 1755–1764, 2000.
[61] P. de Kerret and D. Gesbert, "The multiplexing gain of the network MIMO channel with distributed CSI," *IEEE Transactions on Information Theory*, To appear, arXiv:1108.3742.
[62] P. de Kerret and D. Gesbert, "Towards optimal CSI allocation in multicell MIMO channels," in *Proceedings of IEEE International Conference on Communications*, 2012.
[63] P. Dighe, R. Mallik, and S. Jamuar, "Analysis of transmit-receive diversity in rayleigh fading," *IEEE Transactions on Communications*, vol. 51, no. 4, pp. 694–703, 2003.
[64] G. Dimić and N. Sidiropoulos, "On downlink beamforming with greedy user selection: Performance analysis and a simple new algorithm," *IEEE Transactions on Signal Processing*, vol. 53, no. 10, pp. 3857–3868, 2005.
[65] M. Dohler, R. Heath, A. Lozano, C. Papadias, and R. Valenzuela, "Is the PHY layer dead?," *IEEE Communications Magazine*, vol. 49, no. 4, pp. 159–165, 2011.
[66] Y. Eldar and N. Merhav, "A competitive minimax approach to robust estimation of random parameters," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1931–1946, 2004.
[67] K. Eriksson, S. Shi, N. Vučić, M. Schubert, and E. Larsson, "Globally optimal resource allocation for achieving maximum weighted sum rate," in *Proceedings of IEEE Global Communications Conference*, 2010.
[68] R. Etkin, A. Parekh, and D. Tse, "Spectrum sharing for unlicensed bands," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 3, pp. 517–528, 2007.
[69] R. Etkin, D. Tse, and H. Wang, "Gaussian interference channel capacity to within one bit," *IEEE Transactions on Information Theory*, vol. 54, no. 12, pp. 5534–5562, 2008.
[70] K. Fan, "Minimax theorems," *Proceedings of National Academic Society*, vol. 39, no. 1, pp. 42–47, 1953.
[71] C. Farsakh and J. Nossek, "Channel allocation and downlink beamforming in an SDMA mobile radio system," in *Proceedings of IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, pp. 687–691, 1995.

[72] G. Fettweis, M. Löhning, D. Petrovic, M. Windisch, P. Zillmann, and W. Rave, "Dirty RF: A new paradigm," in *Proceedings of IEEE International Symposium on Personal Indoor and Mobile Radio Communications*, pp. 2347–2355, 2005.

[73] M. Fitz and J. Seymour, "On the bit error probability of QAM modulation," *International Journal on Wireless Information Networks*, vol. 1, no. 2, pp. 131–139, 1994.

[74] G. J. Foschini and M. J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless Personal Communications*, vol. 6, no. 3, pp. 311–335, 1998.

[75] M. Fuchs, G. D. Galdo, and M. Haardt, "Low-complexity space-time-frequency scheduling for MIMO systems with SDMA," *IEEE Transactions on Vehicular Technology*, vol. 56, no. 5, pp. 2775–2784, 2007.

[76] C. Galiotto, Y. Huang, N. Marchetti, and M. Zorzi, "Performance evaluation of non-ideal RF transmitter in LTE/LTE-advanced systems," in *Proceedings of European Wireless*, pp. 266–270, 2009.

[77] I. Garcia, N. Kusashima, K. Sakaguchi, and K. Araki, "Dynamic cooperation set clustering on base station cooperation cellular networks," in *Proceedings of IEEE International Symposium on Personal Indoor and Mobile Radio Communications*, pp. 2127–2132, 2010.

[78] L. Georgiadis, M. Neely, and L. Tassiulas, "Power Control in Wireless Cellular Networks," *Foundations and Trends in Networking*, vol. 1, no. 1, pp. 1–144, 2006.

[79] D. Gerlach and A. Paulraj, "Base station transmitting antenna arrays for multipath environments," *Signal Processing*, vol. 54, no. 1, pp. 59–73, 1996.

[80] A. Gershman, N. Sidiropoulos, S. Shahbazpanahi, M. Bengtsson, and B. Ottersten, "Convex optimization-based beamforming," *IEEE Signal Processing Magazine*, vol. 27, no. 3, pp. 62–75, 2010.

[81] D. Gesbert, S. Hanly, H. Huang, S. Shamai, O. Simeone, and W. Yu, "Multi-cell MIMO cooperative networks: A new look at interference," *IEEE Journal on Selected Areas in Communications*, vol. 28, no. 9, pp. 1380–1408, 2010.

[82] D. Gesbert, S. Kiani, A. Gjendemsjø, and G. Øien, "Adaptation, coordination, and distributed resource allocation in interference-limited wireless networks," *Proceedings of Institute of Electrical and Electronics Engineers*, vol. 95, no. 12, pp. 2393–2409, 2007.

[83] D. Gesbert and M. Kountouris, "Rate scaling laws in multicell networks under distributed power control and user scheduling," *IEEE Transactions on Information Theory*, vol. 57, no. 1, pp. 234–244, 2011.

[84] D. Gesbert, M. Kountouris, R. Heath, C.-B. Chae, and T. Sälzer, "Shifting the MIMO paradigm," *IEEE Signal Processing Magazine*, vol. 24, no. 5, pp. 36–46, 2007.

[85] G. Ginis and J. Cioffi, "A multi-user precoding scheme achieving crosstalk cancellation with application to DSL systems," in *Proceedings of Asilomer Conference on Signals, Systems, and Computers*, 2000.

[86] A. Giovanidis, J. Krolikowski, and S. Brueck, "A 0-1 program to form minimum cost clusters in the downlink of cooperating base stations," in *Proceedings of IEEE Wireless Communication and Networking Conference*, 2012.

[87] S. Goel and R. Negi, "Guaranteeing secrecy using artificial noise," *IEEE Transactions on Wireless Communications*, vol. 7, no. 6, pp. 2180–2189, 2008.
[88] J. Goldberg and J. Fonollosa, "Downlink beamforming for spatially distributed sources in cellular mobile communications," *Signal Processing*, vol. 65, no. 2, pp. 181–197, 1998.
[89] A. Goldsmith, S. Jafar, N. Jindal, and S. Vishwanath, "Capacity limits of MIMO channels," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 5, pp. 684–702, 2003.
[90] A. Goldsmith, S. Jafar, I. Maric, and S. Srinivasa, "Breaking spectrum gridlock with cognitive radios: An information theoretic perspective," *Proceedings of Institute of Electrical and Electronics Engineers*, vol. 97, no. 5, pp. 894–914, 2009.
[91] K. Gomadam, V. Cadambe, and S. Jafar, "Approaching the capacity of wireless networks through distributed interference alignment," in *Proceedings of IEEE Global Communications Conference*, 2008.
[92] J. Gong, S. Zhou, Z. Niu, L. Geng, and M. Zheng, "Joint scheduling and dynamic clustering in downlink cellular networks," in *Proceedings of IEEE Global Communications Conference*, 2011.
[93] J. González-Coma, P. Castro, and L. Castedo, "Impact of transmit impairments on multiuser MIMO non-linear transceivers," in *Proceedings of ITG Workshop on Smart Antennas (WSA)*, 2011.
[94] B. Göransson, S. Grant, E. Larsson, and Z. Feng, "Effect of transmitter and receiver impairments on the performance of MIMO in HSDPA," in *Proceedings of IEEE Workshop on Signal Processing Advances in Wireless Communications*, 2008.
[95] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming," http://cvxr.com/cvx, April 2011.
[96] M. Guillaud, D. Slock, and R. Knopp, "A practical method for wireless channel reciprocity exploitation through relative calibration," in *Proceedings of International Symposium on Signal Processing and Its Applications*, pp. 403–406, 2005.
[97] C. Guthy, W. Utschick, and G. Dietl, "Low-complexity linear zero-forcing for the MIMO broadcast channel," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 6, pp. 1106–1117, 2009.
[98] Y. Haimes, L. Lasdon, and D. Wismer, "On a bicriterion formulation of the problems of integrated system identification and system optimization," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 1, no. 3, pp. 296–297, 1971.
[99] D. Hammarwall, M. Bengtsson, and B. Ottersten, "Utilizing the spatial information provided by channel norm feedback in SDMA systems," *IEEE Transactions on Signal Processing*, vol. 56, no. 7, pp. 3278–3293, 2008.
[100] S. Han, Q. Zhang, and C. Yang, "Distributed coordinated multi-point downlink transmission with over-the-air communication," in *Proceedings of International Conference on Communications and Networking in China*, 2010.
[101] T. Han and K. Kobayashi, "A new achievable rate region for the interference channel," *IEEE Transactions on Information Theory*, vol. 27, no. 1, pp. 49–60, 1981.

[102] S. Haykin, "Cognitive radio: Brain-empowered wireless communications," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 2, pp. 201–219, 2005.

[103] H. Holma and A. Toskala, *LTE for UMTS: Evolution to LTE-Advanced*. Wiley, 2nd ed., 2011.

[104] M. Hong and Z.-Q. Luo, "E-reference signal processing," ch. *Signal Processing and Optimal Resource Allocation for the Interference Channel*. Elsevier, 2013.

[105] X. Hou, E. Björnson, C. Yang, and M. Bengtsson, "Cell-grouping based distributed beamforming and scheduling for multi-cell cooperative transmission," in *Proceedings of IEEE International Symposium on Personal Indoor and Mobile Radio Communications*, 2011.

[106] H. Huang, M. Trivellato, A. Hottinen, M. Shafi, P. Smith, and R. Valenzuela, "Increasing downlink cellular throughput with limited network MIMO coordination," *IEEE Transactions on Wireless Communications*, vol. 8, no. 6, pp. 2983–2989, 2009.

[107] Y. Huang and D. Palomar, "Rank-constrained separable semidefinite program with applications to optimal beamforming," *IEEE Transactions on Signal Processing*, vol. 58, no. 2, pp. 664–678, 2010.

[108] Y. Huang, G. Zheng, M. Bengtsson, K. Wong, L. Yang, and B. Ottersten, "Distributed multicell beamforming design approaching Pareto boundary with max-min fairness," *IEEE Transactions on Wireless Communications*, vol. 11, no. 8, pp. 2921–2933, 2012.

[109] H. Huh, G. Caire, H. Papadopoulos, and S. Ramprashad, "Achieving "massive MIMO spectral efficiency with a not-so-large number of antennas," *IEEE Transactions on Wireless Communications*, vol. 11, no. 9, pp. 3226–3239, 2012.

[110] T. Hungerford, *Abstract Algebra: An Introduction*. Brooks Cole, 1996.

[111] N. Jaldén, P. Zetterberg, B. Ottersten, and L. Garcia, "Inter- and intrasite correlations of large-scale parameters from macrocellular measurements at 1800 MHz," *EURASIP Journal on Wireless Communications and Networking*, 2007.

[112] Z. Jiang, Y. Ge, and Y. Li, "Max-utility wireless resource management for best-effort traffic," *IEEE Transactions on Wireless Communications*, vol. 4, no. 1, pp. 100–111, 2005.

[113] N. Jindal, "MIMO broadcast channels with finite-rate feedback," *IEEE Transactions on Information Theory*, vol. 52, no. 11, pp. 5045–5060, 2006.

[114] S. Jing, D. Tse, J. Soriaga, J. Hou, J. Smee, and R. Padovani, "Multicell downlink capacity with coordinated processing," *EURASIP Journal on Wireless Communications and Networking*, 2008.

[115] M. Joham, W. Utschick, and J. Nossek, "Linear transmit processing in MIMO communications systems," *IEEE Transactions on Signal Processing*, vol. 53, no. 8, pp. 2700–2712, 2005.

[116] E. Jorswieck, "Secrecy capacity of single- and multi-antenna channels with simple helpers," in *Proceedings of ITG Conference on Source and Channel Coding*, 2010.

[117] E. Jorswieck and E. Larsson, "The MISO interference channel from a game-theoretic perspective: A combination of selfishness and altruism achieves Pareto optimality," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 5364–5367, 2008.

[118] E. Jorswieck and E. Larsson, "Monotonic optimization framework for the two-user MISO interference channel," *IEEE Transactions on Communications*, vol. 58, no. 7, pp. 2159–2168, 2010.

[119] E. Jorswieck, E. Larsson, and D. Danev, "Complete characterization of the Pareto boundary for the MISO interference channel," *IEEE Transactions on Signal Processing*, vol. 56, no. 10, pp. 5292–5296, 2008.

[120] E. Jorswieck and R. Mochaourab, "Beamforming in underlay cognitive radio: Null-shaping constraints and greedy user selection," in *Proceedings of IEEE CrownCom*, pp. 1–5, 2010.

[121] E. Jorswieck, A. Wolf, and S. Gerbracht, "Trends in Telecommunications Technologies," ch. *Secrecy on the Physical Layer in Wireless Networks*. IN-TECH Education and Publishing, 2010.

[122] J. Jose, A. Ashikhmin, T. Marzetta, and S. Vishwanath, "Pilot contamination and precoding in multi-cell TDD systems," *IEEE Transactions on Communications*, vol. 10, no. 8, pp. 2640–2651, 2011.

[123] S. Joshi, P. Weeraddana, M. Codreanu, and M. Latva-aho, "Coordinated beamforming for MISO interference channel: Complexity analysis and efficient algorithms," *IEEE Transactions on Signal Processing*, vol. 60, no. 4, pp. 2090–2095, 2012.

[124] V. Jungnickel, T. Wirth, M. Schellmann, T. Haustein, and W. Zirwas, "Synchronization of cooperative base stations," in *Proceedings of IEEE International Symposium on Wireless Communication Systems*, pp. 329–334, 2008.

[125] M. Karakayali, G. Foschini, and R. Valenzuela, "Network coordination for spectrally efficient communications in cellular systems," *IEEE Wireless Communications Magazine*, vol. 13, no. 4, pp. 56–61, 2006.

[126] E. Karipidis and E. Larsson, "Efficient computation of the Pareto boundary for the MISO interference channel with perfect CSI," in *Proceedings of International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks*, pp. 573–577, 2010.

[127] E. Karipidis, N. Sidiropoulos, and Z.-Q. Luo, "Quality of service and max-min fair transmit beamforming to multiple cochannel multicast groups," *IEEE Transactions on Signal Processing*, vol. 56, no. 3, pp. 1268–1279, 2008.

[128] S. Kaviani and W. Krzymien, "Multicell scheduling in network MIMO," in *Proceedings of IEEE Global Communications Conference*, 2010.

[129] S. Kaviani, O. Simeone, W. Krzymien, and S. Shamai, "Linear precoding and equalization for network MIMO with partial cooperation," *IEEE Transactions on Vehicular Technology*, vol. 61, no. 5, pp. 2083–2095, 2012.

[130] F. Kelly, "Charging and rate control for elastic traffic," *European Transactions on Telecommunication*, vol. 8, pp. 33–37, 1997.

[131] F. Kelly, A. Maulloo, and D. Tan, "Rate control for communication networks: shadow prices, proportional fairness and stability," *Journal on Operational Research Society*, vol. 49, no. 3, pp. 237–252, 1997.

[132] J. Kermoal, L. Schumacher, K. Pedersen, P. Mogensen, and F. Frederiksen, "A stochastic MIMO radio channel model with experimental validation," *IEEE Journal on Selected Areas in Communications*, vol. 20, no. 6, pp. 1211–1226, 2002.

[133] H. Kha, H. Tuan, and H. Nguyen, "Fast global optimal power allocation in wireless networks by local D.C. programming," *IEEE Transactions on Wireless Communications*, vol. 11, no. 2, pp. 510–515, 2012.

[134] S. Kiani and D. Gesbert, "Optimal and distributed scheduling for multicell capacity maximization," *IEEE Transactions on Wireless Communications*, vol. 7, no. 1, pp. 288–297, 2008.

[135] M. Kobayashi, M. Debbah, and J. Belfiore, "Outage efficient strategies in network MIMO with partial CSIT," in *Proceedings of IEEE International Symposium on Information Theory*, pp. 249–253, 2009.

[136] B. Kolman and R. Beck, *Elementary Linear Programming with Applications*. Academic Press, 1995.

[137] S. Koskie and Z. Gajic, "Signal-to-interference-based power control for wireless networks: A survey, 1992–2005," *Dynamics of Continuous, Discrete and Impulsive Systems B*, vol. 13, no. 2, pp. 187–220, 2006.

[138] J. Kotecha and A. Sayeed, "Transmit signal design for optimal estimation of correlated MIMO channels," *IEEE Transactions on Signal Processing*, vol. 52, no. 2, pp. 546–557, 2004.

[139] M. Ku and D. Kim, "Tx-Rx beamforming with multiuser MIMO channels in multiple-cell systems," in *Proceedings of International Conference on Advanced Communication Technology*, pp. 1767–1771, 2008.

[140] E. Larsson, E. Jorswieck, J. Lindblom, and R. Mochaourab, "Game theory and the flat-fading Gaussian interference channel," *IEEE Signal Processing Magazine*, vol. 26, no. 5, pp. 18–27, 2009.

[141] V. Lau, "Proportional fair spacetime scheduling for wireless communications," *IEEE Transactions on Communications*, vol. 53, no. 8, pp. 1353–1360, 2005.

[142] B. Lee, H. W. Je, I. Sohn, O.-S. Shin, and K. B. Lee, "Interference-aware decentralized precoding for multicell MIMO TDD systems," in *Proceedings of IEEE Global Communications Conference*, 2008.

[143] G. Lee, J. Park, Y. Sung, and M. Yukawa, "Coordinated beamforming with relaxed zero forcing," in *Proceedings of International Conference on Wireless Communications and Signal Processing*, 2011.

[144] J. Lee and N. Jindal, "Symmetric capacity of MIMO downlink channels," in *Proceedings of IEEE International Symposium on Information Theory*, pp. 1031–1035, 2006.

[145] J.-W. Lee, R. Mazumdar, and N. Shroff, "Non-convex optimization and rate control for multi-class services in the internet," *IEEE/ACM Transactions on Networking*, vol. 13, no. 4, pp. 827–840, 2005.

[146] J. Li, H. Zhang, X. Xu, X. Tao, T. Svensson, C. Botella, and B. Liu, "A novel frequency reuse scheme for coordinated multi-point transmission," in *Proceedings of IEEE Vehicular Technology Conference-Spring*, 2010.

[147] Y. Liang, H. V. Poor, and S. Shamai (Shitz), "Information Theoretic Security," *Foundations and Trends in Communications and Information Theory*, vol. 5, no. 4–5, pp. 355–580, 2009.

[148] J. Lindblom, E. Karipidis, and E. Larsson, "Selfishness and altruism on the MISO interference channel: The case of partial transmitter CSI," *IEEE Communications Letters*, vol. 13, no. 9, pp. 667–669, 2009.

[149] J. Lindblom, E. Karipidis, and E. Larsson, "Closed-form parameterization of the Pareto boundary for the two-user MISO interference channel," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 3372–3375, 2011.

[150] A. Liu, Y. Liu, H. Xiang, and W. Luo, "Polite water-filling for weighted sum-rate maximization in MIMO B-MAC networks under multiple linear constraints," *IEEE Transactions on Signal Processing*, vol. 60, no. 2, pp. 834–847, 2012.

[151] H. Liu and G. Xu, "Multiuser blind channel estimation and spatial channel pre-equalization," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 1756–1759, 1995.

[152] J. Liu, T. Hou, and H. D. Sherali, "Optimal power allocation for achieving perfect secrecy capacity in MIMO wire-tap channels," in *Proceedings of IEEE International Conference on Communications*, 2009.

[153] L. Liu, R. Zhang, and K. Chua, "Achieving global optimality for weighted sum-rate maximization in the K-user Gaussian interference channel with multiple antennas," *IEEE Transactions on Wireless Communications*, vol. 11, no. 5, pp. 1933–1945, 2012.

[154] Q. Liu, R. Baxley, X. Ma, and G. Zhou, "Error vector magnitude optimization for OFDM systems with a deterministic peak-to-average power ratio constraint," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 3, pp. 418–429, 2009.

[155] R. Liu, T. Liu., H. Poor, and S. Shamai (Shitz), "Multiple-input multiple-output Gaussian broadcast channels with confidential messages," *IEEE Transactions on Information Theory*, vol. 56, no. 9, pp. 4215–4227, 2010.

[156] R. Liu and W. Trappe, *Securing Wireless Communications at the Physical Layer*. Springer Berlin Heidelberg, 2010.

[157] Y.-F. Liu, Y.-H. Dai, and Z.-Q. Luo, "Coordinated beamforming for MISO interference channel: Complexity analysis and efficient algorithms," *IEEE Transactions on Signal Processing*, vol. 59, no. 3, pp. 1142–1157, 2011.

[158] Y.-F. Liu, Y.-H. Dai, and Z.-Q. Luo, "Max-min fairness linear transceiver design for a multi-user MIMO interference channel," in *Proceedings of IEEE International Conference on Communications*, 2011.

[159] T. Lo, "Maximum ratio transmission," *IEEE Transactions on Communications*, vol. 47, no. 10, pp. 1458–1461, 1999.

[160] M. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret, "Applications of second-order cone programming," *Linear Algebra and its Applications*, vol. 284, no. 1–3, pp. 193–228, 1998.

[161] J. Löfberg, "YALMIP: A toolbox for modeling and optimization in MATLAB," in *Proceedings of IEEE International Symposium on Computer Aided Control System Design*, pp. 284–289, 2004.

[162] S. Loyka, "Channel capacity of MIMO architecture using the exponential correlation matrix," *IEEE Communications Letters*, vol. 5, no. 9, pp. 369–371, 2001.

[163] S. Loyka, V. Kostina, and F. Gagnon, "Error rates of the maximum-likelihood detector for arbitrary constellations: Convex/concave behavior and applications," *IEEE Transactions on Information Theory*, vol. 56, no. 4, pp. 1948–1960, 2010.

[164] A. Lozano, R. Heath, and J. Andrews, "Fundamental limits of cooperation," *IEEE Transactions on Information Theory*, Submitted, arXiv:1204.0011.

[165] A. Lozano and N. Jindal, "Transmit diversity vs. spatial multiplexing in modern MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 9, no. 1, pp. 186–197, 2010.

[166] Z.-Q. Luo, W.-K. Ma, A.-C. So, Y. Ye, and S. Zhang, "Semidefinite relaxation of quadratic optimization problems," *IEEE Signal Processing Magazine*, vol. 27, no. 3, pp. 20–34, 2010.

[167] Z.-Q. Luo and W. Yu, "An introduction to convex optimization for communications and signal processing," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 8, pp. 1426–1438, 2006.

[168] Z.-Q. Luo and S. Zhang, "Dynamic spectrum management: Complexity and duality," *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 1, pp. 57–73, 2008.

[169] J. Lv, R. Blasco-Serrano, E. Jorswieck, R. Thobaben, and A. Kliks, "Optimal Beamforming in MISO Cognitive Channels with Degraded Message Sets," in *IEEE Wireless Communications and Networking Conference*, 2012.

[170] J. Lv and E. Jorswieck, "Spatial shaping in cognitive system with coded legacy transmission," in *Proceedings of ITG Workshop on Smart Antennas (WSA)*, 2011.

[171] A. MacKenzie and L. DaSilva, *Game Theory for Wireless Engineers*. Morgan & Claypool, 2006.

[172] M. Maddah-Ali, A. Mobasher, and A. Khandani, "Fairness in multiuser systems with polymatroid capacity region," *IEEE Transactions on Information Theory*, vol. 55, no. 5, pp. 2128–2138, 2009.

[173] P. Marsch, S. Brück, A. Garavaglia, M. Schulist, R. Weber, and A. Dekorsy, "Clustering," in *Coordinated Multi-Point in Mobile Communications: From Theory to Practice*, ch. 7, (P. Marsch and G. Fettweis, eds.), pp. 139–159, Cambridge, 2011.

[174] P. Marsch and G. Fettweis, "On multicell cooperative transmission in backhaul-constrained cellular systems," *Annals of Telecommunication*, vol. 63, pp. 253–269, 2008.

[175] P. Marsch and G. Fettweis, "On downlink network MIMO under a constrained backhaul and imperfect channel knowledge," in *Proceedings of IEEE Global Communications Conference*, 2009.

[176] P. Marsch and G. Fettweis, "Static clustering for cooperative multi-point (CoMP) in mobile communications," in *Proceedings of IEEE International Conference on Communications*, 2011.

[177] H.-P. Mayer and H. Schlesinger, "Antenna synchronization for coherent network MIMO," US Patent, 20120002967, 2010.

[178] M. McKay, A. Grant, and I. Collings, "Performance analysis of MIMO-MRC in double-correlated rayleigh environments," *IEEE Transactions on Communications*, vol. 55, no. 3, pp. 497–507, 2007.

[179] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," *IEEE/ACM Transactions on Networking*, vol. 8, no. 5, pp. 556–567, 2000.

[180] R. Mochaourab and E. Jorswieck, "Optimal beamforming in interference networks with perfect local channel information," *IEEE Transactions on Signal Processing*, vol. 59, no. 3, pp. 1128–1141, 2011.

[181] R. Mochaourab and E. Jorswieck, "Walrasian equilibrium in two-user multiple-input single-output interference channels," in *Proceedings of IEEE International Conference on Communications*, 2011.

[182] R. Mochaourab and E. Jorswieck, "Robust beamforming in interference channels with imperfect transmitter channel information," *Signal Processing*, vol. 92, no. 11, pp. 2509–2518, 2012.

[183] P. Mogensen, W. Na, I. Kovács, F. Frederiksen, A. Pokhariyal, K. Pedersen, T. Kolding, K. Hugl, and M. Kuusela, "LTE capacity compared to the Shannon bound," in *Proceedings of IEEE Vehicular Technology Conference-Spring*, 2007.

[184] N. Moghadam, P. Zetterberg, P. Händel, and H. Hjalmarsson, "Correlation of distortion noise between the branches of MIMO transmit antennas," in *Proceedings of IEEE International Symposium on Personal Indoor and Mobile Radio Communications*, 2012.

[185] M. Mohseni, R. Zhang, and J. Cioffi, "Optimized transmission for fading multiple-access and broadcast channels with multiple antennas," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 8, pp. 1627–1639, 2006.

[186] G. Montalbano, I. Ghauri, and D. Slock, "Multiuser blind channel estimation and spatial channel pre-equalization," in *Proceedings of Asilomer Conference on Signals, Systems, and Computers*, pp. 1337–1341, 1998.

[187] R. Mueller and G. Foschini, "The capacity of linear channels with additive Gaussian noise," *The Bell System Technical Journal*, pp. 81–94, 1970.

[188] C. Nader, P. Händel, and N. Björsell, "Peak-to-average power reduction of OFDM signals by convex optimization: Experimental validation and performance optimization," *IEEE Transactions on Instrumentation and Measurement*, vol. 60, no. 2, pp. 473–479, 2011.

[189] M. Naeem and D. Lee, "On convexity of MQAM's and MPAM's bit error probability functions," *International Journal on Communication Systems*, vol. 22, pp. 1465–1477, 2009.

[190] R. Nettleton and H. Alavi, "Power control for a spread spectrum cellular mobile radio system," in *Proceedings of IEEE Vehicular Technology Conference*, pp. 242–246, 1983.

[191] B. L. Ng, J. Evans, S. Hanly, and D. Aktas, "Distributed downlink beamforming with cooperative base stations," *IEEE Transactions on Information Theory*, vol. 54, no. 12, pp. 5491–5499, 2008.

[192] H. Nikaido, *Convex Structures and Economic Theory.* Academic Press, 1968.

[193] M. Nokleby and A. Swindlehurst, "Bargaining and the MISO interference channel," *EURASIP Journal on Advances in Signal Process*, 2009.

[194] D. Palomar and M. Chiang, "A tutorial on decomposition methods for network utility maximization," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 8, pp. 1439–1451, 2006.

[195] D. Palomar, J. Cioffi, and M. Lagunas, "Joint Tx-Rx beamforming design for multicarrier MIMO channels: A unified framework for convex optimization," *IEEE Transactions on Signal Processing*, vol. 51, no. 9, pp. 2381–2401, 2003.

[196] D. Palomar and Y. Jiang, "MIMO transceiver design via majorization theory," *Foundations and Trends in Communications and Information Theory*, vol. 3, no. 4–5, pp. 331–551, 2006.

[197] A. Pantelidou and A. Ephremides, "Scheduling in wireless networks," *Foundations and Trends in Communications and Information Theory*, vol. 4, no. 4, pp. 421–511, 2010.

[198] A. Papadogiannis and G. Alexandropoulos, "A dynamic clustering approach in wireless networks with multi-cell cooperative processing," in *Proceedings of IEEE International Conference on Fuzzy Systems*, 2010.

[199] A. Papadogiannis, D. Gesbert, and E. Hardouin, "A dynamic clustering approach in wireless networks with multi-cell cooperative processing," in *Proceedings of IEEE International Conference on Communications*, 2008.

[200] A. Papadogiannis, E. Hardouin, and D. Gesbert, "Decentralising multicell cooperative processing: A novel robust framework," *EURASIP Journal on Wireless Communications and Networking*, 2009.

[201] S.-H. Park, H. Parkand, H. Kong, and I. Lee, "New beamforming techniques based on virtual SINR maximization for coordinated multi-cell transmission," *IEEE Transactions on Wireless Communications*, vol. 11, no. 3, pp. 1034–1044, 2012.

[202] S. Parkvall, E. Dahlman, A. Furuskär, Y. Jading, M. Olsson, S. Wänstedt, and K. Zangi, "LTE-advanced — evolving LTE towards IMT-advanced," in *Proceedings of IEEE Vehicular Technology Conference-Fall*, 2008.

[203] C. Peel, B. Hochwald, and A. Swindlehurst, "A vector-perturbation technique for near-capacity multiantenna multiuser communication — part I: Channel inversion and regularization," *IEEE Transactions on Communications*, vol. 53, no. 1, pp. 195–202, 2005.

[204] H. Pennanen, A. Tölli, and M. Latva-aho, "Decentralized coordinated downlink beamforming via primal decomposition," *IEEE Signal Processing Letters*, vol. 18, no. 11, pp. 647–650, 2011.

[205] S. Pillai, T. Suel, and S. Cha, "The Perron-Frobenius theorem: Some of its applications," *IEEE Signal Processing Magazine*, vol. 22, no. 2, pp. 62–75, 2005.

[206] L. Qian, Y. Zhang, and J. Huang, "MAPEL: Achieving global optimality for a non-convex wireless power control problem," *IEEE Transactions on Wireless Communications*, vol. 8, no. 3, pp. 1553–1563, 2009.

[207] J. Qiu, R. Zhang, Z.-Q. Luo, and S. Cui, "Optimal distributed beamforming for MISO interference channels," *IEEE Transactions on Signal Processing*, vol. 59, no. 11, pp. 5638–5643, 2011.

[208] F. Rashid-Farrokhi, K. Liu, and L. Tassiulas, "Transmit beamforming and power control for cellular wireless systems," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 8, pp. 1437–1450, 1998.

[209] F. Rashid-Farrokhi, L. Tassiulas, and K. Liu, "Joint optimal power control and beamforming in wireless networks using antenna arrays," *IEEE Transactions on Communications*, vol. 46, no. 10, pp. 1313–1324, 1998.

[210] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "Linear transceiver design for a MIMO interfering broadcast channel achieving max-min fairness," in *Proceedings of Asilomar Conference on Signals, Systems, and Computers*, pp. 1309–1313, 2011.

[211] T. Ren and R. La, "Downlink beamforming algorithms with inter-cell interference in cellular networks," *IEEE Transactions on Wireless Communications*, vol. 5, no. 10, pp. 2814–2823, 2006.

[212] W. Rhee and J. Cioffi, "Increase in capacity of multiuser OFDM system using dynamic subchannel allocation," in *Proceedings of IEEE Vehicular Technology Conference-Spring*, pp. 1085–1089, 2000.

[213] R. Rockafellar, "Lagrange multipliers and optimality," *SIAM Review*, vol. 35, no. 2, pp. 183–238, 1993.

[214] J. Roh and B. Rao, "Multiple antenna channels with partial channel state information at the transmitter," *IEEE Transactions on Wireless Communications*, vol. 3, no. 2, pp. 677–687, 2004.

[215] M. Rossi, A. Tulino, O. Simeone, and A. Haimovich, "Non-convex utility maximization in Gaussian MISO broadcast and interference channels," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 2960–2963, 2011.

[216] B. Roy and V. Mousseau, "A theoretical framework for analysing the notion of relative importance of criteria," *Journal on Multi-Criteria Decision Analysis*, vol. 5, pp. 145–159, 1996.

[217] R. Roy and B. Ottersten, "Spatial division multiple access wireless communication systems," US Patent, 5515378, 1991.

[218] A. Rubinov, H. Tuy, and H. Mays, "An algorithm for monotonic global optimization problems," *Optimization*, vol. 49, pp. 205–221, 2001.

[219] W. Rudin, *Principles of Mathematical Analysis*. McGraw-Hill, 1976.

[220] F. Rusek, D. Persson, B. Lau, E. Larsson, T. Marzetta, O. Edfors, and F. Tufvesson, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Processing Magazine*, vol. 30, no. 1, pp. 40–60, 2013.

[221] M. Sadek, A. Tarighat, and A. Sayed, "A leakage-based precoding scheme for downlink multi-user MIMO channels," *IEEE Transactions on Wireless Communications*, vol. 6, no. 5, pp. 1711–1721, 2007.

[222] S. Sarkar and L. Tassiulas, "Fair allocation of discrete bandwidth layers in multicast networks," in *Proceedings of IEEE INFOCOM*, 2000.

[223] L. Savage, *The Foundations of Statistics*. Courier Dover Publications, 1972.

[224] T. Schenk, *RF Imperfections in High-Rate Wireless Systems: Impact and Digital Compensation*. Springer, 2008.

[225] D. Schmidt, C. Shi, R. Berry, M. Honig, and W. Utschick, "Distributed resource allocation schemes," *IEEE Signal Processing Magazine*, vol. 26, no. 5, pp. 53–63, 2009.

[226] M. Schubert and H. Boche, "Solution of the multiuser downlink beamforming problem with individual SINR constraints," *IEEE Transactions on Vehicular Technology*, vol. 53, no. 1, pp. 18–28, 2004.

[227] M. Schubert and H. Boche, "QoS-based resource allocation and transceiver optimization," *Foundations and Trends in Communications and Information Theory*, vol. 2, no. 6, pp. 383–529, 2005.

[228] M. Schubert and H. Boche, *Interference Calculus: A General Framework for Interference Management and Network Utility Optimization*. Springer, 2012.

[229] G. Scutari, D. Palomar, J.-S. Pang, and F. Facchinei, "Flexible design of cognitive radio wireless systems," *IEEE Signal Processing Magazine*, vol. 26, no. 5, pp. 107–123, 2009.

[230] G. Scutari, D. P. Palomar, and S. Barbarossa, "Cognitive MIMO radio," *IEEE Signal Processing Magazine*, vol. 25, no. 6, pp. 46–59, 2008.

[231] K. Seong, M. Mohseni, and J. Cioff, "Optimal resource allocation for OFDMA downlink systems," in *Proceedings of IEEE International Symposium on Information Theory*, 2006.

[232] S. Shafiee and S. Ulukus, "Achievable rates in Gaussian MISO channels with secrecy constraints," in *Proceedings of IEEE International Symposium on Information Theory*, 2007.

[233] S. Shamai and B. Zaidel, "Enhancing the cellular downlink capacity via co-processing at the transmitting end," in *Proceedings of IEEE Vehicular Technology Conference-Spring*, pp. 1745–1749, 2001.

[234] X. Shang, B. Chen, G. Kramer, and H. V. Poor, "Noisy-interference sum-rate capacity of parallel Gaussian interference channels," *IEEE Transactions on Information Theory*, vol. 57, no. 1, pp. 210–226, 2011.

[235] X. Shang, B. Chen, and H. V. Poor, "Multiuser MISO interference channels with single-user detection: Optimality of beamforming and the achievable rate region," *IEEE Transactions on Information Theory*, vol. 57, no. 7, pp. 4255–4273, 2011.

[236] C. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, 1948.

[237] C. Shannon, "Communication theory of secrecy systems," *Bell System Technical Journal*, vol. 28, pp. 656–715, 1949.

[238] M. Sharif and B. Hassibi, "On the capacity of MIMO broadcast channels with partial side information," *IEEE Transactions on Information Theory*, vol. 51, no. 2, pp. 506–522, 2005.

[239] C. Shen, T.-H. Chang, K.-Y. Wang, Z. Qiu, and C.-Y. Chi, "Distributed robust multicell coordinated beamforming with imperfect CSI: An ADMM approach," *IEEE Transactions on Signal Processing*, vol. 60, no. 6, pp. 2988–3003, 2012.

[240] Z. Shen, R. Chen, J. Andrews, R. Heath, and B. Evans, "Low complexity user selection algorithms for multiuser MIMO systems with block diagonalization," *IEEE Transactions on Signal Processing*, vol. 54, no. 9, pp. 3658–3663, 2006.

[241] M. B. Shenouda and T. Davidson, "Probabilistically-constrained approaches to the design of the multiple antenna downlink," in *Proceedings of Asilomar Conference on Signals, Systems, and Computers*, pp. 1120–1124, 2008.

[242] M. B. Shenouda and T. Davidson, "Nonlinear and linear broadcasting with QoS requirements: Tractable approaches for bounded channel uncertainties," *IEEE Transactions on Signal Processing*, vol. 57, no. 5, pp. 1936–1947, 2009.

[243] Q. Shi, M. Razaviyayn, Z.-Q. Luo, and C. He, "An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel," *IEEE Transactions on Signal Processing*, vol. 59, no. 9, pp. 4331–4340, 2011.

[244] H. Shirani-Mehr, G. Caire, and M. Neely, "MIMO downlink scheduling with non-perfect channel state knowledge," *IEEE Transactions on Communications*, vol. 58, no. 7, pp. 2055–2066, 2010.

[245] N. Sidiropoulos, T. Davidson, and Z.-Q. Luo, "Transmit beamforming for physical-layer multicasting," *IEEE Transactions on Signal Processing*, vol. 54, no. 6, pp. 2239–2251, 2006.

[246] O. Simeone, N. Levy, A. Sanderovich, O. Somekh, B. Zaidel, H. Poor, and S. Shamai, "Cooperative wireless cellular systems: An information-theoretic view," *Foundations and Trends in Communications and Information Theory*, vol. 8, no. 1–2, pp. 1–177, 2012.

[247] O. Simeone, O. Somekh, H. V. Poor, and S. Shamai, "Downlink multicell processing with limited-backhaul capacity," *EURASIP Journal on Advances in Signal Processing*, 2009.

[248] H. Skjevling, D. Gesbert, and A. Hjørungnes, "Low-complexity distributed multibase transmission and scheduling," *EURASIP Journal on Advances in Signal Processing*, 2008.

[249] M. Slater, "Lagrange multipliers revisited," Technical Report 403, Cowles Commission Discussion Paper, Mathematics, 1950.

[250] B. Song, Y.-H. Lin, and R. Cruz, "Weighted max-min fair beamforming, power control, and scheduling for a MISO downlink," *IEEE Transactions on Signal Processing*, vol. 7, no. 2, pp. 464–469, 2008.

[251] E. Song, Q. Shi, M. Sanjabi, R. Sun, and Z.-Q. Luo, "Robust SINR-constrained MISO downlink beamforming: When is semidefinite programming relaxation tight?," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2011.

[252] Q. Spencer, A. Swindlehurst, and M. Haardt, "Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels," *IEEE Transactions on Signal Processing*, vol. 52, no. 2, pp. 461–471, 2004.

[253] R. Stridh, M. Bengtsson, and B. Ottersten, "System evaluation of optimal downlink beamforming with congestion control in wireless communication," *IEEE Transactions on Wireless Communications*, vol. 5, no. 4, pp. 743–751, 2006.

[254] C. Studer, M. Wenk, and A. Burg, "MIMO transmission with residual transmit-RF impairments," in *Proceedings of ITG/IEEE Workshop on Smart Antennas (WSA)*, 2010.

[255] C. Studer, M. Wenk, and A. Burg, "System-level implications of residual transmit-RF impairments in MIMO systems," in *Proceedings of European Conference on Antennas and Propagation*, 2011.

[256] J. Sturm, "Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones," *Optimization Methods and Software*, vol. 11–12, pp. 625–653, 1999.

[257] A. Tajer, N. Prasad, and X. Wang, "Robust linear precoder design for multi-cell downlink transmission," *IEEE Transactions on Signal Processing*, vol. 59, no. 1, pp. 235–251, 2011.

[258] L. Tanner, "Selecting a text-processing system as a qualitative multiple criteria problem," *European Journal on Operational Research*, vol. 50, no. 2, pp. 179–187, 1991.

[259] A. Tarighat, M. Sadek, and A. Sayed, "A multi user beamforming scheme for downlink MIMO channels based on maximizing signal-to-leakage ratios," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, pp. 1129–1132, 2005.

[260] A. Tarighat and A. Sayed, "Joint compensation of transmitter and receiver impairments in OFDM systems," *IEEE Transactions on Wireless Communications*, vol. 6, no. 1, pp. 240–247, 2007.

[261] E. Telatar, "Capacity of multi-antenna Gaussian channels," *European Transactions on Telecommunications*, vol. 10, no. 6, pp. 585–595, 1999.

[262] A. Tölli, M. Codreanu, and M. Juntti, "Soft handover in adaptive MMIMO-OFDM cellular system with cooperative processing," in *Proceedings of IEEE International Symposium on Personal Indoor and Mobile Radio Communications*, 2006.

[263] A. Tölli, M. Codreanu, and M. Juntti, "Cooperative MIMO-OFDM cellular system with soft handover between distributed base station antennas," *IEEE Transactions on Wireless Communications*, vol. 7, no. 4, pp. 1428–1440, 2008.

[264] A. Tölli, H. Pennanen, and P. Komulainen, "On the value of coherent and coordinated multi-cell transmission," in *Proceedings of IEEE International Conference on Communications*, 2009.

[265] A. Tölli, H. Pennanen, and P. Komulainen, "Decentralized minimum power multi-cell beamforming with limited backhaul signaling," *IEEE Transactions on Wireless Communications*, vol. 10, no. 2, pp. 570–580, 2011.

[266] D. Tomecki and S. Stanczak, "On feasible SNR region for multicast downlink channel: Two user case," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, 2010.

[267] N. Tripathi, J. H. Reed, and H. Vanlandingham, *Radio Resource Management in Cellular Systems*. Springer, 2001.

[268] M. Trivellato, F. Boccardi, and H. Huang, "On transceiver design and channel quantization for downlink multiuser MIMO systems with limited feedback," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 8, pp. 1494–1504, 2008.

[269] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.

[270] T.-L. Tung and K. Yao, "Optimal downlink power-control design methodology for a mobile radio DS-CDMA system," in *Proceedings of IEEE Workshop on Signal Processing Systems*, pp. 165–170, 2002.

[271] R. Tütüncü, K. Toh, and M. Todd, "Solving semidefinite-quadratic-linear programs using SDPT3," *Mathematical Programming*, vol. 95, no. 2, pp. 189–217, 2003.
[272] H. Tuy, "Global minimization of a difference of two convex functions," *Nonlinear Analysis and Optimization*, vol. 30, pp. 150–182, 1987.
[273] H. Tuy, *Convex Analysis and Global Optimization*. Kluwer Academic Publishers, 1998.
[274] H. Tuy, "Monotonic optimization: Problems and solution approaches," *SIAM Journal of Optimization*, vol. 11, no. 2, pp. 464–494, 2000.
[275] H. Tuy, F. Al-Khayyal, and P. Thach, "Monotonic optimization: Branch and cut methods," in *Essays and Surveys in Global Optimization*, (C. Audet, P. Hansen, and G. Savard, eds.), Springer US, 2005.
[276] W. Utschick and J. Brehmer, "Monotonic optimization framework for coordinated beamforming in multicell networks," *IEEE Transactions on Signal Processing*, vol. 60, no. 4, pp. 1899–1909, 2012.
[277] S. Vavasis, "Complexity issues in global optimization: A survey," in *Handbook of Global Optimization*, pp. 27–41, Kluwer, 1995.
[278] M. Vázquez, A. Pérez-Neira, and M. Lagunas, "A unifying approach to transmit beamforming for the MISO interference channel," in *Proceedings of ITG Workshop on Smart Antennas (WSA)*, 2012.
[279] S. Venkatesan, A. Lozano, and R. Valenzuela, "Network MIMO: Overcoming intercell interference in indoor wireless systems," in *Proceedings of IEEE Asilomar Conference on Signals, Systems, and Computers*, pp. 83–87, 2007.
[280] L. Venturino, N. Prasad, and X. Wang, "Coordinated linear beamforming in downlink multi-cell wireless networks," *IEEE Transactions on Wireless Communications*, vol. 9, no. 4, pp. 1451–1461, 2010.
[281] S. Verdú, *Multiuser Detection*. Cambridge University Press, 1998.
[282] E. Visotsky and U. Madhow, "Optimum beamforming using transmit antenna arrays," in *Proceedings of IEEE Vehicular Technology Conference*, pp. 851–856, 1999.
[283] P. Viswanath and D. Tse, "Sum capacity of the vector Gaussian broadcast channel and uplink-downlink duality," *IEEE Transactions on Information Theory*, vol. 49, no. 8, pp. 1912–1921, 2003.
[284] P. Viswanath, D. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Transactions on Information Theory*, vol. 48, no. 6, pp. 1277–1294, 2002.
[285] B. Vojčić and W. Jang, "Transmitter precoding in synchronous multiuser communications," *IEEE Transactions on Communications*, vol. 46, no. 10, pp. 1346–1355, 1998.
[286] N. Vučić and H. Boche, "Robust QoS-constrained optimization of downlink multiuser MISO systems," *IEEE Transactions on Signal Processing*, vol. 57, no. 2, pp. 714–725, 2009.
[287] S. Wagner, R. Couillet, M. Debbah, and D. Slock, "Large system analysis of linear precoding in MISO broadcast channels with limited feedback," *IEEE Transactions on Information Theory*, vol. 58, no. 7, pp. 4509–4537, 2012.

[288] J. Wallace and M. Jensen, "Measured characteristics of the MIMO wireless channel," in *Proceedings on IEEE Vehicular Technology Conference-Fall*, pp. 2038–2042, 2001.

[289] K.-Y. Wang, T.-H. Chang, C.-Y. C. W.-K. Ma, and A. So, "Probabilistic SINR constrained robust transmit beamforming: A Bernstein-type inequality based conservative approach," in *Proceedings on IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 3080–3083, 2011.

[290] R. Weber, A. Garavaglia, M. Schulist, S. Brueck, and A. Dekorsy, "Self-organizing adaptive clustering for cooperative multipoint transmission," in *Proceedings of IEEE Vehicular Technology Conference-Spring*, 2011.

[291] P. Weeraddana, M. Codreanu, S. Joshi, and M. Latva-aho, "Multicell downlink weighted sum-rate maximization: A distributed approach," in *Proceedings of Asilomar Conference on Signals, Systems, and Computers*, pp. 1368–1375, 2011.

[292] P. Weeraddana, M. Codreanu, M. Latva-aho, and A. Ephremides, "Weighted sum-rate maximization for a set of interfering links via branch and bound," *IEEE Transactions on Signal Processing*, vol. 59, no. 8, pp. 3977–3996, 2011.

[293] P. Weeraddana, M. Codreanu, M. Latva-aho, A. Ephremides, and C. Fischione, "Weighted Sum-Rate Maximization in Wireless Networks: A Review," *Foundations and Trends in Networking*, vol. 6, no. 1–2, pp. 1–163, 2012.

[294] W. Weichselberger, M. Herdin, H. Özcelik, and E. Bonek, "A stochastic MIMO channel model with joint correlation of both link ends," *IEEE Transactions on Wireless Communications*, vol. 5, no. 1, pp. 90–100, 2006.

[295] H. Weingarten, Y. Steinberg, and S. Shamai, "The capacity region of the Gaussian multiple-input multiple-output broadcast channel," *IEEE Transactions on Information Theory*, vol. 52, no. 9, pp. 3936–3964, 2006.

[296] A. Wiesel, Y. Eldar, and S. Shamai, "Linear precoding via conic optimization for fixed MIMO receivers," *IEEE Transactions on Signal Processing*, vol. 54, no. 1, pp. 161–176, 2006.

[297] A. Wiesel, Y. Eldar, and S. Shamai, "Zero-forcing precoding and generalized inverses," *IEEE Transactions on Signal Processing*, vol. 56, no. 9, pp. 4409–4418, 2008.

[298] A. Wyner, "The wire-tap channel," *Bell System Technical Journal*, vol. 54, pp. 1355–1387, 1975.

[299] A. Wyner, "Shannon-theoretic approach to a Gaussian cellular multiple-access channel," *IEEE Transactions on Information Theory*, vol. 40, no. 6, pp. 1713–1727, 1994.

[300] J. Xu, J. Zhang, and J. Andrews, "On the accuracy of the Wyner model in cellular networks," *IEEE Transactions on Wireless Communications*, vol. 10, no. 9, pp. 3098–3109, 2011.

[301] Y. Xu, T. Le-Ngoc, and S. Panigrahi, "Global concave minimization for optimal spectrum balancing in multi-user DSL networks," *IEEE Transactions on Signal Processing*, vol. 56, no. 7, pp. 2875–2885, 2008.

[302] J. Yang, E. Björnson, and M. Bengtsson, "Receive beamforming design based on a multiple-state interference model," in *Proceedings of IEEE International Conference on Communications*, 2011.

[303] W. Yang and G. Xu, "Optimal downlink power assignment for smart antenna systems," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 3337–3340, 1998.

[304] R. Yates, "A framework for uplink power control in cellular radio systems," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 7, pp. 1341–1347, 1995.

[305] T. Yoo and A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 3, pp. 528–541, 2006.

[306] K. Yu, M. Bengtsson, B. Ottersten, D. McNamara, P. Karlsson, and M. Beach, "Modeling of wide-band MIMO radio channels based on NLoS indoor measurements," *IEEE Transactions on Vehicular Technology*, vol. 53, no. 3, pp. 655–665, 2004.

[307] W. Yu, T. Kwon, and C. Shin, "Multicell coordination via joint scheduling, beamforming and power spectrum adaptation," in *Proceedings of IEEE INFOCOM*, 2011.

[308] W. Yu and T. Lan, "Transmitter optimization for the multi-antenna downlink with per-antenna power constraints," *IEEE Transactions on Signal Processing*, vol. 55, no. 6, pp. 2646–2660, 2007.

[309] S. Zabell, "Alan Turing and the central limit theorem," *The American Mathematical Monthly*, vol. 102, no. 6, pp. 483–494, 1995.

[310] R. Zakhour and D. Gesbert, "Coordination on the MISO interference channel using the virtual SINR framework," in *Proceedings of ITG Workshop on Smart Antennas (WSA)*, 2009.

[311] R. Zakhour and D. Gesbert, "Distributed multicell-MISO precoding using the layered virtual SINR framework," *IEEE Transactions on Wireless Communications*, vol. 9, no. 8, pp. 2444–2448, 2010.

[312] R. Zakhour and D. Gesbert, "Team decision for the cooperative MIMO channel with imperfect CSIT sharing," in *Proceedings of Information Theory and Applications Workshop*, 2010.

[313] R. Zakhour and D. Gesbert, "Optimized data sharing in multicell MIMO with finite backhaul capacity," *IEEE Transactions on Signal Processing*, vol. 59, no. 12, pp. 6102–6111, 2011.

[314] J. Zander, "Performance of optimum transmitter power control in cellular radio systems," *IEEE Transactions on Vehicular Technology*, vol. 41, no. 1, pp. 57–62, 1992.

[315] J. Zander and M. Frodigh, "Comment on "Performance of optimum transmitter power control in cellular radio systems"," *IEEE Transactions on Vehicular Technology*, vol. 43, no. 3, p. 636, 1994.

[316] J. Zander and S.-L. Kim, *Radio Resource Management for Wireless Networks*. Artech House, 2001.

[317] A. Zanella, M. Chiani, and M. Win, "Performance of MIMO MRC in correlated rayleigh fading environments," in *Proceedings of IEEE Vehicular Technology Conference-Spring*, pp. 1633–1637, 2005.

[318] B. Zarikoff and J. Cavers, "Coordinated multi-cell systems: Carrier frequency offset estimation and correction," *IEEE Journal on Selected Areas in Communications*, vol. 28, no. 9, pp. 1490–1501, 2010.

[319] P. Zetterberg, "Experimental investigation of TDD reciprocity-based zero-forcing transmit precoding," *EURASIP Journal on Advances in Signal Processing*, January 2011.

[320] P. Zetterberg and B. Ottersten, "The spectrum efficiency of a base station antenna array system for spatially selective transmission," *IEEE Transactions on Vehicular Technology*, vol. 44, no. 3, pp. 651–660, 1995.

[321] H. Zhang and H. Dai, "Cochannel interference mitigation and cooperative processing in downlink multicell multiuser MIMO networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2, pp. 222–235, 2004.

[322] H. Zhang, N. Mehta, A. Molisch, J. Zhang, and H. Dai, "Asynchronous interference mitigation in cooperative base station systems," *IEEE Transactions on Wireless Communications*, vol. 7, no. 1, pp. 155–165, 2008.

[323] J. Zhang, R. Chen, J. Andrews, A. Ghosh, and R. Heath, "Networked MIMO with clustered linear precoding," *IEEE Transactions on Wireless Communications*, vol. 8, no. 4, pp. 1910–1921, 2009.

[324] Q. Zhang and H. Li, "MOEA/D: A multiobjective evolutionary algorithm based on decomposition," *IEEE Transactions on Evolutionary Computation*, vol. 11, no. 6, pp. 712–731, 2007.

[325] R. Zhang and S. Cui, "Cooperative interference management with MISO beamforming," *IEEE Transactions on Signal Processing*, vol. 58, no. 10, pp. 5450–5458, 2010.

[326] R. Zhang, Y.-C. Liang, and S. Cui, "Dynamic resource allocation in cognitive radio networks," *IEEE Signal Processing Magazine*, vol. 27, no. 3, pp. 102–114, 2010.

[327] X. Zhang and Y.-C. Liang, "Exploiting multi-antennas for opportunistic spectrum sharing in cognitive radio networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 1, pp. 88–102, 2008.

[328] G. Zheng, K.-K. Wong, and T.-S. Ng, "Robust linear MIMO in the downlink: A worst-case optimization with ellipsoidal uncertainty regions," *EURASIP Journal on Advances in Signal Processing*, 2008.

[329] G. Zheng, K.-K. Wong, and T.-S. Ng, "Throughput maximization in linear multiuser MIMO-OFDM downlink systems," *IEEE Transactions on Vehicular Communications*, vol. 57, no. 3, pp. 1993–1998, 2008.

[330] J. Zyren and W. McCoyh, "Overview of the 3GPP long term evolution physical layer," in *Freescale Semiconductor*, 2007.