1 Introduction

1.1 Motivations

Remarkable driving forces motivated scientists and engineers to move their ideas beyond the contemporary conceptual wireless generations to long-term evolution (LTE) of mobile communications towards 5G with power of imagination to create a system that seems at the time to be far-off from possible. After a self-reflection on my experience and overview of the literature, it was compelling to investigate how these enchanting technologies could be implemented to the physical layer of multiple-input multiple-output (MIMO) wireless systems working in a non-ideal environment. Pre-LTE, cellular wireless generation advances mostly implies a partial change in the characteristics of the system to enhance data speed, and/or power efficiencies. Each generation has particular standards, capabilities, limitations, and technologies with modified architectures to differentiate it from the previous one. Operators will find 5G infrastructures are not depending on previous generations but will be based on generic roll-out enterprise, so they have to upgrade their networks infrastructure. 4G mobile services are still developing in both coverage and data speed capabilities. So why do we need 5G?

The key benefits of 5G over 4G will not only be the speed of data delivery, which is expected to be between 10 and 100 Gb/s, but with latency. Low latency provides the possibility of connecting to immediately send and receive data in real-time. In the first quarter of 2018, the average latency for 4G was 75 ms and for LTE–Advanced was 45 ms, both much lower than latency in 3G, which was 135 ms. However, LTE-Advanced latency is still not enough to support real-time response.

5Gs will be ultra-low-latency with a possible range between less than 1 and 10 ms. 5G capacity and performance enhancements attribute to a massive MIMO system and are also important for coping with rapid growth of the Internet of things (IoT). IoT will transform the traditional physical objects into smart objects with the capabilities to see, hear, think, and perform jobs by communicating to each other and sharing information to coordinate their decisions. Indeed, IoT services support our economy to grow and also improve our lives. The 3rd Generation Partnership Project (3GPP) standardized IoT as the narrowband-IoT. IoT needs 5G high capability to cope with the vast number of connected devices. So, the bid for bandwidth (BW) will continue to grow even though IoT is a narrowband system. It is predicted that by 2020, there will be at least 20 billion devices connected to the internet. There is another problem between 4G and IoT and it's the latency. Low latency is a key enabler for coping with the massive number of connected devices.

5G will provide connections to such colossal numbers of Internet-connected IoT smart objects. The amount of data transported by wireless mobile networks is expected to increase rapidly with 5G. As an example, consider the time it takes to download an 800 MB movie. On average, it takes 26 minutes and 40 seconds for 3G (speed 4 kb/s) and 4 minutes and 16 seconds

5G Physical Layer Technologies, First Edition. Mosa Ali Abu-Rgheff.

for 4G (speed 25 Mb/s). This will be reduced to a fraction of a second (0.3 second) with 5G (speed 20 Gb/s).

The current growth of data is mainly driven by devices such as smart phones, tablets, and video streaming. In addition, the number of connected devices and data rates will continue to increase exponentially. Consequently, it was clear to researchers and industry that the incremental approach to system improvements to these challenges would not achieve its targets by 2020. There are enormous demands on academia and industry to meet these challenges through innovative new technologies that are practical, smart, and efficient.

During the last five years, there have been a number of collaborative research projects between academia and industry to develop and test technologies that can enable 5G to deliver its expected performance. These technologies will be part of the 5G networks according to the general consensus of the research community.

5G is more than a mobile network generation. It will advance our economy by generating a substantial contribution to the growth in gross domestic product (GDP) and will enhance our life and society. 5G offers these benefits through a set of technologies connecting people to people and people to everything. The full economic benefits encompass education, automotive industry, transport, building, health care, and entertainment, to name a few. Additionally, 5G Gigabit wireless connectivity and computing capacity help the development of smart cities and smart municipalities across our country, thus providing a cleaner, safer, and healthier environment for citizens both today and tomorrow. Furthermore, industries such as automotive, health care, and IoT will notice the effects of the changeover to 5G. Indeed, IoT will change the way we think of the Internet as a human-to-human interface to a more machine-to-machine platform. In particular, automotive, health care, and the IoT will bring about exciting benefits to our daily lives.

5G will not be an incremental advance on 4G. Indeed, it will be a paradigm shift with massive bandwidth, high base-station (BS) and device densities, and a very large number of antennas. 5G's greatest difficulty of providing a uniform service to a large number of users is mitigated using the formation of heterogeneous (HetNet) networking or small-cell scenarios. Small cell (discussed in Chapter 2) is a key radio cell used in HetNet of coverage 20–400 m and transmitting power of 20 mW–2 W. In addition, 5G applications require a cloud-based architectural elements. Accordingly, 5G radio access has to be flexible, scalable, robust, reliable, and efficient in spectrum and energy.

The outstanding expectations for 5G systems are in *data rate*, *latency*, *energy,* and *cost*. The expected aggregate data rate (i.e. network area capacity) in bits/s is expected to increase by 1000X compared to the 4G current optimal data rate, while the *worst data rate* that a user can expect to receive (known as edge rate, or 5% rate) ranges from 100 Mb/s to as much as 1 Gb/s. 5G is expected to increase links data rate between 10X and 100X compared to the 4G data rate. 5G low latency enriches applications such as gaming, tactile internet, wearable devices. 5G is likely to reduce (or at least not increase) energy consumption, and for this reason it is expected to reduce the running cost of the system.

With the purpose of achieving such high provisions, 5G is enabled by a combination of new and well-known technologies. These sets of enabling technologies aim to provide unlimited access to information and the ability to share data anywhere, anytime, by anyone for the comfort of people and to the advantage of business and society.

Half-duplex transceivers (TRXs) are commonly implemented in LTE devices, which means users can either transmit or receive but not both operations simultaneously. This is because of the self-interference generated at the user's end when the user is transmitting and receiving at the same time. 5G will be enabled by technology that removes the self-interference and users will then enjoy full duplex services. This implies that 5G users should be able to transmit messages and receive data at the same time. This technology is explored comprehensively in Chapter 2. Furthermore, while LTE devices provide azimuth (horizontal) transmission and reception, 5G systems, using 3D beamforming, will provide users with services in horizontal and vertical dimensions. In practice, this implies that the LTE service would be limited to horizontal dimension while 5G service will cover users in multi-story building as well. This technology is known as full-dimension multiple-input multiple-output (FD-MIMO) technology and is investigated thoroughly in Chapter 2.

According to the United Nations (UN), as of 2016, more than 54.5% of the world's population live in urban settlements, and this is projected to increase to 60% by 2030. Furthermore, the UN habitat organization predicted that by 2017, cities will cover 'less than 2% of the earth surface and consumed 78% of world energy and produced 60% of CO_2 and greenhouse gas emissions'. Consequently, cities are facing a magnitude of challenges, not only because they generate most of the air pollution but also because they put a strain on available resources. Therefore, appropriate solutions are needed to resolve these challenges. Such solutions have to transform traditional cities into *smart and resilient cities* by integrating the communication technologies and the IoT to manage the city's assets.

EU smart city reference model is based on a city performing well in six characteristics: smart economy, smart mobility, smart environment, smart people, smart living, and smart governance. These characteristics are evaluated by 31 factors and 74 indicators. The reference comprised three versions: version 2.0 and 3.0 published in 2013 and 2014, respectively, and version 4.0 in 2015. Versions 2.0 and 3.0 EU smart cities model were defined for medium-sized cities and version 4.0 is defined for large EU cities. Since the publication of the reference model, many EU collaborative projects have been funded to investigate innovative schemes that brand EU cities smarter using the 5G information technologies (i.e. make traditional networks and services smart and efficient such as urban transport, water supply, waste disposal facilities, city administration, and city assets, to name few smart and efficient). A sample of these projects includes REViSITE (2010–2012), OPTIMS (2013–2016), BESOS (2013–2016), and CITYOPT (2014–2017).

In order for 5G to exploit the mmWave frequency bands, it is important to learn more about the channel characteristics and the radio propagation models. Contemporary propagation models were validated for frequencies up to about 2 GHz and are not adequate for small cells. In addition, such models are not suitable for hilly and heavily wooded terrains. Radio propagation in indoor environments is not influenced by terrain profiles as the outdoors propagation but can be affected by the layout inside a building and influenced by various building materials used. Furthermore, atmospheric gaseous, clouds, fog, snow, and hail have to be accounted for in the mmWave propagation.

Channel models are widely used to predict the performance of the propagated signal under various constraints. There is a need for new channel models to support mmWave 5G operation. These models have to support large bandwidths and be suitable to medium- and high-mobility speeds. In addition, 5G channel models have to support lD and 2D massive MIMO with dual polarization and are appropriate for beamforming design. Intensive research activities were carried out by a number of groups for channels modelling at mmWave bands, such as Mobile and wireless communications Enablers for the Twenty-twenty Information Society (METIS), International Telecommunication Union-Radio (ITU-R), COST, and 5G mmWave Channel Alliance, to name a few. So far, research investigations show two possible channel models: independent and identically distributed (i.i.d) Rayleigh channel is used widely for flat faded channels, and the correlation inspired models. The latter comprised two types: the Kronecker channel model, where correlation occurs at one end of the link, and the Weichselberger channel model, exhibiting joint correlation of both link ends.

The electromagnetic (EM) action by the array elements on each other instigates array mutual coupling. Such action takes the form of energy absorption by one antenna element when a nearby element is radiating EM energy. Energy absorption by mutual coupling reduces antenna performance efficiency.

Array coupling problems have attracted interest not only from antenna engineers and researchers but also from researchers in other disciplines, such as biomedical imaging, including magnetic resonance imaging (MRI) that operates antenna arrays. Array mutual coupling can be reduced by spacing the antennas as far apart as possible. In practice, due to constraints on the physical dimensions of available spaces, the distances between the multiple antennas are usually small and antenna elements interact with each other (i.e. they are mutually coupled). Consequently, to improve the MIMO system efficiency, we need to find ways to decouple transmit and/or receive antenna arrays.

Reliable mmWave communication depends decisively on 3D beamforming to aim the beam in the correct direction and the ability to separate the data into independent streams aimed at end users using a process called multiuser precoding. Beamforming needs the knowledge of the spatial channel state information (CSI) at the transmitter and receiver. Spatial CSI is obtained by an adequately accurate estimate of the channel. In time-division duplex (TDD) transmission protocol, pilot sequences are transmitted on the uplink to obtain UL channel estimates at the BS. The downlink (DL) is determined using the reciprocity property.

That is, in ideal circumstances the DL CSI are given by the reciprocal of the UL CSI, and theoretically, there is no need to estimate the DL channel. There are basically two possibilities for the spatial channel estimation in cellular networks, namely: channel estimation in noncooperative TDD networks and estimation using a coordinated cells MIMO system. In massive MIMO multiuser cellular systems, it will be necessary to code each user data before transmission in a process called multiuser precoding and decode the received signals at the receiver in a second process called equalization to equalize the coding at the transmitter.

The purpose of the book is to present the most recent knowledge about 5G enabling technologies as applied to the PHY layer in a single book and in a unified framework. Various techniques are analysed and evaluated. Details are cited in a reader-friendly writing approach in all chapters. The analysis and explanation rendered are concise and informative. As the case with any advanced engineering book, derivations and analysis apply advanced mathematics. To alleviate these difficulties from students and professionals, or any related skills that require advanced mathematical expertise, mathematical difficulties are resolved either in the main text or, if the solution is lengthy, in appendices.

1.2 Overview of Contemporary Cellular Wireless Networks

First-generation (1G) commercial network was used in the 1980s and built to provide only analogue voice services without security. Users commonly experienced the frustration of dropped calls and poor battery life.

Second-generation (2G) was commercialized in the 1990s and was an enormous improvement to the subscriber experiences compared with 1G. The motivation for 2G was to increase system capacity; add additional services such SMS and caller ID, etc.; improve security; and reduce cost. 2G, also known as GSM, or global system for mobile. The key features of the GSM cellular system specification include using a digital time-division multiple access (TDMA) approach. Both UL and DL frequency bands are in the 900 MHz and 1.8 GHz bands in Europe.

There are five different cell sizes in GSM network, namely: macro, micro, pico, femto, and umbrella cells. Clearly, a lot of microcells contribute to a large number of handoffs, especially when the speed of the mobile is too high. An umbrella cell covers several micro cells to reduce the number of handoffs. By adopting this technique, more users can be accommodated within the available bandwidth. The signalling waveform used was Gaussian minimum-shift keying (GMSK). GSM cellular network is a circuit-switched system adopting 200 KHz radio frequency (RF) channels, and these channels are time-division multiplexed to enable up to eight users to access each carrier. So, in a practical scenario, GSM is a TDMA/frequency-division multiple access (FDMA) system, and each user is transmitting at about 34 kb/s. The total channel data rate is 270.8 kb/s.

GSM provided the following services: digital voice, SMS, international roaming, conferencing, call waiting, call hold, call forwarding, call barring, caller number identification, closed user groups (CUGs), unstructured supplementary service data (USSD) services, and authentication billing based on the services provided to their customers (e.g. charges based on local calls, long-distance calls, discounted calls, real-time billing). GSM uses USSD to send data between a mobile system and an application in the network.

GSM was an exceptionally successful mobile communication system, initially intended for use within EU, but within a short time it became an internationally accepted system. Further development of GSM gave rise to the general packet radio service (GPRS) using packet-switched data rather that circuit-switched data. The maximum number of time slots that can be used is four, giving a possible maximum data transfer rate of 57.6 kbps (or 38.4 kbps on a GSM 900 network) per time slot. Two major concerns for operators are capital expenditure (CAPEX) and operational expenditure (OPEX) costs.

Another extension of GSM is the Enhanced Data rates for GSM Evolution (EDGE) technology that uses 8PSK modulation, which require small changes in the BS and could often be achieved by software upgrades. EDGE offers radio data rate per time slot equal 69.2 kb/s and maximum user data rate per time slot 59.2, which gives maximum user data when using eight time slots to be 473.6 kb/s. 2G standard was implemented using different technologies: EU standard implemented the GSM technology but other technologies were used as well – United States (IS-95) and Japan (JDC). The data transfer speeds of GSM systems were only 35–50 kb/s.

Third-generation (3G) systems aimed at increasing network capacity and data throughput to promote growth and support new applications. Universal mobile telecommunications system (UMTS) was the main establisher of the 3G, and its new standards in cellular system performance and functionalities. In 1992, WARC allocated 230 MHz of spectrum between 1885–2025 and 2110–2200 MHz for the next generation (3G). A number of international organizations accepted the demand for international standards for the next-generation mobile communication. EU ETSI support UMTS, ARIB in Japan; and many other proposals were submitted, but the ones that gained support were UMTS/code division multiple access (CDMA); CDMA2000 under Interim Standard IS-95 first deployed using CDMA technology; and TDS-CDMA (time division synchronous code division multiple access) developed by China.

A number of companies and standards organizations set the 3GPP to produce technical specifications for 3G with radio access including both TDD and frequency division duplex (FDD) technologies. 3GPP published releases 99, 4, 5, 6, and 7 in specifying 3G. So, 3G enrolled in the EU in 2001 and UK operator 3 launched 3G in 2003. Wideband code division multiple access (WCDMA) can adopt Alamouti 2×2 MIMO scheme, which is the only orthogonal space-time code with rate 1. However, for transmit antennas higher than two, several orthogonal codes have been found but with lower rates, so spectral efficiency can be degraded. 3G standards based on WCDMA technology do not support MIMO technology.

UMTS/WCDMA with radio interface based on WCDMA technology offered DL data speed 384 kb/s–2 Mb/s. High-speed packet access (HSPA) enhanced WCDMA speeds to 14.4 Mb/s in DL and 5.76 Mb/s in UL. An upgrade to HSPA (HSPA+) provided theoretical peak speeds of 168 Mb/s in DL and 22 Mb/s in UL. A second organization, 3GPP2, was created by North America and Asian operators to adopt CDMA2000 for 3G with the following technologies: One times radio transmission technology (1 × RTT) recommended DL speeds up to 144 kb/s; Evolution-Data Optimized (EV-DO) increased DL data speeds up to 2.4 Mb/s; Evolution-Data Optimized Revision-A (EV-DO Rev. A) increased the DL speeds up to 3.1 Mb/s and reduced latency; Evolution-Data Optimized (EV-DO Rev. B) offered 2–15 channels, each with DL peak rate at 4.9 Mb/s; and finally, Ultra Mobile Broadband (UMB) DL peak rate reached 288 Mb/s.

3G was launched in many countries, but the success of 2G was difficult to replicate because the advancement from 2G to 3G innovated only a few services – not sufficiently inspiring for users to change their operators. 3GPP incorporated the multimedia broadcast and multicast service (MBMS) in combination with the internet protocol multimedia system (IMS) with no adjustment to the access technology [1].

Since evolved HSPA (also known as HSPA+) was launched in 2010, data traffic driven by smart phones and tablets has grown rapidly. Nokia traffic model published in 2012 using two traffic assumptions predicted traffic increase of 1000x relative to the traffic in 2010 will come into being around 2020 [2]. Operators should be prepared to install networks that can cope with this momentous traffic upsurge. The development of 3GPP LTE technologies is based on a series of releases. Most current 4G networks are based on 3GPP Releases 8 and 9, offering peak DL speed of 100 Mb/s and peak UL 50 Mb/s. Nonetheless, so to appropriately respond to increasing traffic and capacity growth, the LTE radio access technology has continuously evolved. 3GPP LTE-Advanced introduced enhancement on the existing 4G LTE networks and raised the peak DL transmission speed to 1000 Mb/s and the UL peak speed to 500 Mb/s, as recommended by 3GPP Release 10.

Macro cells bear almost the entire data traffic in contemporary mobile network, complemented by small cells in hot zones. A key performance enhancer to macro cells capacity is to partition each cell into smaller cells. Networks comprise an overlay of cells of different sizes. For example, outdoor users are likely to be served by a combination of macro, micro, and pico cells. Low-power remote radio head (RRH) and pico cells may provide both outdoor and indoor coverage. They can be employed in hotspots/hot zones such as shopping districts, train stations, or shopping malls, with typical cell radius of up to 200 m. In addition, pico and femto cells can be used indoors in cells of no more than 10–25 m. The tendency for multilayer placement (i.e. small-cell densification) is forced by the need to maintain better service quality both indoors and outdoors.

Conventionally, small cells in pre-LTE operate independently of the overlaid macro layer and transmit all the cell signals comprising cell-specific reference signals, synch signals, and the system full information. In addition, a user communicates with small cells and macro cells in both UL and DL directions. Coverage provided by the macro cell can be enhanced by operating the macro- and small-cell layers in a more integrated approach. Most importantly, *dual connectivity* requires the user terminal to be simultaneously connected to the two nodes [3]. Dual connectivity can provide multiple benefits in addition to throughput aggregation of the two nodes. The macro node is an *anchor node* to provide a mobile internet protocol (IP) link to communicate with the mobile user, as shown in Figure 1.1.

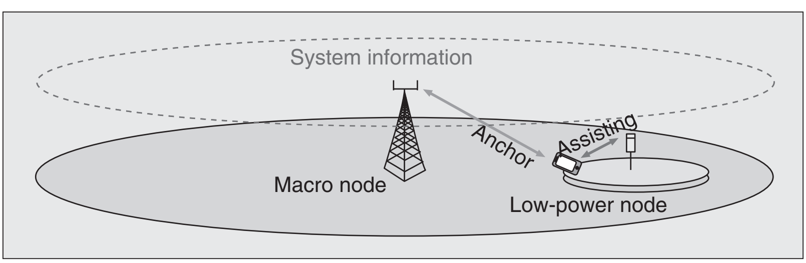

Figure 1.1 Soft cell: dual connectivity to the anchor node and assisting node [3].

Essentially, there exist two techniques to split a macro cell, in the horizontal (azimuth) domain and in the vertical domain. Combination of both sectorization approaches arrange for narrow beam targeting of single or more user(s).

1.3 Evolution of Wireless Communications in 3GPP Releases

3G technology is based on the novelty of the spread spectrum concept, a technology developed as an anti-jamming capability by the military establishments. The energy-spreading technique reduces the level of transmitted information below background noise level to provide data protection against any third-party intruder. The BW expansion is achieved using a coding process that is independent of the information being transmitted or the modulation being applied.

Commercial spread spectrum scheme generates procession gain to express the bandwidth expansion factor. For example, for 20 dB processing gain and bit rate of 1 Gb/s, the spreading clock has to generate over 100G chip per sec (Gc/s). 3G data rate can be increased by efficient spreading waveforms, adopting inter-cell interference cancellation and adaptive coding and modulation. These enhancements to HSPA eventually increased its data rates. These considerations conclude that there is an upper limit on the capacity of 3G networks enforced by the availability of suitable spreading clocks.

3GPP was set up in 1998 to work out specifications for future advanced mobile communications. The public and private collaborative project comprises seven standards organizations: market associations and many industrial companies. The aims of the collaboration were to develop a vision for an all IP wireless network covering radio access/core networks: service capabilities and interworking with Wi-Fi networks. The 3GPP led the campaign for LTE of wireless communications, and 3GPP standards are arranged in Releases. Release 14 was published in 2017; Release 15 was available at the end of 2018, and Release 16 is due at the end of 2019. Releases 14 and 15 concur with 5G trial service tests and 5G commercial service enrolled in 2020. The following section briefly describes some the key evolve technologies of each 3GPP release. 3GPP LTE continues developing enabling technologies for 5G, with releases shown in Figure 1.2.

1.3.1 3GPP Release 8

Release 8 was published in Dec. 2008. 3GPP introduced LTE technology for the very first time. Most of the subsequent releases added enhancements to the previously released technologies. Evolved Packet Core (EPC) is the IP-based core network defined by 3GPP in Release 8 and so circuit-switching data mode is rendered inoperative. EPC for LTE networks was reported by a group of venders in February 2009 to permit operators to adopt a range of access types using a common core network. LTE networks' deployments commenced in 2010 with a theoretically download speed of 100 Mb/s. However, initial speeds tended to be closer to 10 Mb/s. LTE

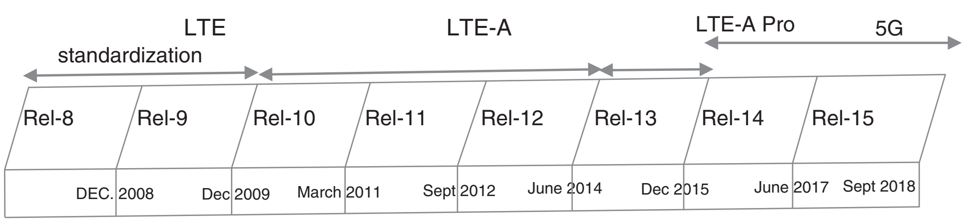

Figure 1.2 3GPP Long-term evolution development towards 5G.

networks adopt 4 × 4 MIMO system and flexible BWs 5–20 MHz in 5 MHz steps together with orthogonal frequency-division multiple access (OFDMA) in DL and single carrier frequency-division multiple access (SC-FDMA) in UL.

1.3.2 3GPP Release 9

Published in Dec 2009, Release 9 brought enhancements to LTE reported in Release 8 specifications. Three positioning schemes were specified namely: Assisted Global Positioning System (A-GPS); Observed Time Difference of Arrival (OTDOA): and Enhanced Cell ID (E-CID). These schemes can be employed to accurately determine a user's location during a natural disaster and emergency cases. Improvements were also added to the concept of self-organized network (SON), which was introduced in Release 8 with a focus on evolved Node B (eNB) self-configuration, while Release 9 added self-optimization as well. Evolved multimedia broadcast Multicast services (eMBMS) were outlined at physical layer in Release 8 but Release 9 completed the specifications for network layer and higher layer aspects. Support for single-layer (single-user) beamforming was described in Release 8 and Release 9 extended the beamforming to multi-layers (multi-users). In addition, Release 9 introduced Commercial Mobile Alert System (CMAS) in addition to the Earth and Tsunami Warning System (ETWS) introduced in Release 8.

1.3.3 3GPP Release 10

3GPP Release 10 described what is generally known as LTE-Advanced systems. Published in March 2011, it implements a combination of large (macro) cells as well as small cells, commonly known as heterogeneous networks (HetNet). In addition, it added enhancements in SON features for self-healing procedures. Release 10 increased the number of antennas in the MIMO system to up to 8 × 8 MIMO in DL and 4 × 4 in UL (user side). Furthermore, it provided worldwide roaming with high spectral efficiency. Release 10 introduced the clustered SC-FDMA in UL and permited selected carriers scheduling in UL, while Release 8 only allowed adjacent block of carriers for scheduling.

A key feature introduced in Release 10 is carrier aggregation (CA), which operators can use to employ their fragmented spectrum scattered across different or same bands to enhance user throughput by transporting data simultaneously over two or more carriers. LTE-Advanced handles BW of 100 MHz composed of five 20 MHz carriers. Another important feature described in Release 10 is the enhanced inter-cell interference coordination (eICIC) to manage the interference issues in HetNet. The eICIC alleviates interference on both data and control channels. Intra-frequency interference is mitigated using power, frequency, and time domain. The peak throughput in LTE-Advanced in DL is 1 Gb/s and 500 Mb/s in the UL. Release 10 recommended the relay nodes connected to a donor eNB to extend the coverage of the system eNB.

1.3.4 3GPP Release 11

Published in September 2012, Release 11 added further improvements to LTE-Advanced, such as introducing the coordinated multipoint (CoMP), which permits the transmitters to share data load transmission and reception even if the remotely hosted are connected by fibre link. An Enhanced Physical Downlink Control Channel (ePDCCH) was proposed to increase control channel capacity. The ePDCCH consumes some of the resources allocated to Physical Downlink Shared Channel (PDSCH) for sending control information, contrary to Release 8, which allows only the use of control region subframes.

Release 11 recommended new positioning schemes for UL added by employing sounding reference signals for time difference measured by a number of neighbouring eNBs. In addition, the release defines a method where user device (UE) informs the network whether it needs to operate in battery saving mode or normal mode. So, based on the user equipment (UE) request network, it can adjust its discontinuous reception (DRX) parameters. Further, Release 11 introduced radio access network (RAN) overload control, where in some scenarios of massive communications from a group of devices, the network can blockade these devices to send connection requests to the network. In addition, Release 11 recommended further enhancements to CA, including multiple time advances (TAs) for UL CA, nonadjacent intra band CA, and the CA support in TDD LTE physical layer.

1.3.5 3GPP Release 12

Published in June 2014, Release 12 recommended further enhancements to LTE-Advanced. Small cells were supported in Release 10 with structures that comprise eICIC. Release 12 recommended the optimization and enhancements of small cells, including their deployment in dense areas. In addition, the release introduced inter-site CA between macro and small cells, also known as dual connectivity. Release 12 focused on machine type communication (MTC) that created huge network signalling and capacity concerns, so a new UE category was defined for optimized MTC operations. One study item in Release 12 recommended operating in unlicensed spectrum, as it delivers many benefits to operators such as increased network capacity and improved performance.

1.3.6 3GPP Release 13

Release 13, also known as initial LTE-Advanced Pro release, was published in December 2015. Release 13 introduced three important technology categories. The first was FD-MIMO. The key idea of FD-MIMO was to employ a large number of antennas arranged in a 2D antenna array panel and additional frequency resources by adopting CA of up to 32 component carriers (CCs) compared to only up to 5 CCs in Release 10. Further, it supported up to 8 antenna MIMO as per previous releases but promised to examine high-order MIMO systems with up to 64 antenna MIMO. Release 13 proposed possible enhancements for DL multiuser transmission using superposition coding. The third were new enhancements recommended to MTC, including a low-complexity UE category, which was defined to support reduced BW, power, and a long battery life. Narrowband Internet of Things (NB-IoT) was standardized for providing wide-area connectivity for massive machine-type communications (mMTCs). Release 13 focuses on a combination of primary cell from licenced spectrum with a secondary cell from an unlicensed spectrum to accommodate the increasing data traffic demand.

1.3.7 3GPP Release 14

Published in June 2017, Release 14 is known as the initial start of 5G standardization. Some of the main technology areas include latency reduction to 1 ms or less for improved end-user experience compared to 5 ms in Release 8 for IP packets in ideal radio conditions. In addition, Release 14 provides full support for UL transmissions in unlicenced spectrum to build on the dual-connectivity framework recommended in Release 13.

Further, Release 14 discussed new use cases such as intelligent transport systems (ITSs) inclusive of vehicular-to-vehicular and vehicular-to-infrastructure communication. LTE technology greatly improves traffic safety on the roads and also makes it possible to include a wide range of road users, e.g. pedestrians and cyclists, in an overall traffic safety work. The existing device-to-device framework can serve as a basis for the work on vehicular-to-vehicular communication. Release 14 introduced more improvements to the MTC capacity and added new features for MTC devices, such as MBMS support for delivering software upgrades and device-to-device relaying for coverage extension. Release 14 focused on extending the MIMO framework to even larger number of antennas. Release 14 recommended a new radio access named as 'NR' to be standardized. The LTE evolution and the 'NX' technologies will build 5G radio access.

1.3.8 3GPP Release 15 (5G phase 1)

3GPP 5G New Radio (NR) specifications included in Release 15, on which work began in June 2016 and was set to complete in September 2018. 5G NR is a new air interface, i.e. the part of the circuit between the mobile device (UE) and the active BS. Active BS is used to imply the fact that BSs change as the UE moves and the BSs change accompanied by a handoff process. The non-standalone (NSA) mode of the NR was approved by the 3GPP in December 2017. The standalone (SA) mode is to be completed by September 2018 and entails user and control planes capability using the new 5G core network architecture.

Achieving the 5G expectation, it is essential that 5G NR must be able to deliver numerous and varied services across a different set of devices with different performance and latency needs. Optimized orthogonal frequency-division multiple access (OFDM) or OFDM variants are to be used for signalling waveforms and multiple access in 5G NR design and should enable lower latency. 5G NR should support a range of deployment models such as macro cells, small cells, and device to device (D2D) connections, with extremely low device cost and high levels of power and deployment efficiencies. The core 5G NR will comprise three key elements: enhanced mobile broadband (eMBB) including technologies Gb/s LTE, massive MIMO, mmWave technologies, and advanced channel coding; ultra-reliable low-latency communications (uRLLC) for services like autonomous deriving with technologies involved such cellular vehicle-to-everything (C-V2X) and cellular drone communications; and mMTC technologies involve such over-the-air broadcast firmware (software that makes hardware work) updates. Release 15 was released in late 2018.

1.3.9 3GPP Release 16 (5G phase 2)

The target completion date for Release 16 is 2020, when complete 5G standards will be published.

1.4 Multiuser Wireless Network Capacity Regions

Wireless channels often hold more than one user transmission. These channels are referred to as multiuser channels. At the network physical channel, there can be three categories of channels: multiple access, broadcast, and interference multiuser channels. The concept of a user is clear in a single user communication, so what constitutes a user in multiuser communication? A user in a multiuser channel can be defined as the amount of information component that can be reliably detected by one or more receivers in the multiuser channel. In other words, a user is channel and information. Furthermore, a single user can be decomposed into a number of subusers. Each subuser in the multiuser channel is described by a channel and stream of subuser data, so a single user rate is the sum of the rates of the streams.

The multiple access channel (MAC) draws signals from a number of individual users transmitted from physically separated transmitters, each with its own message vector. A single receiver processes the single channel output as a vector MAC. MAC in wireless communications is developing when several user transmitters share a common frequency band in transmitting UL to an access point or BS. Let us denote the i^{th} user transmits a symbol vector x_i where $i = 1, 2, \ldots., K$. The receiver output a single vector $\mathbf{y}$. The MAC is shown in

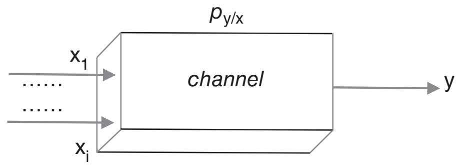

Figure 1.3 The multiple-access channel (MAC).

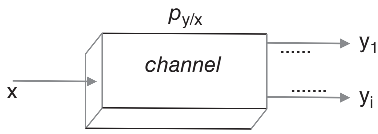

Figure 1.4 The broadcast channel (BC).

Figure 1.5 The interference channel (IC).

Figure 1.3 where $p_{\mathbf{y}/\mathbf{x}}$ is the conditional probability for **x** conditioned to observations **y** and the channel = $[\mathrm{h}_K \ldots\ldots \mathrm{h}_2\ \mathrm{h}_1]$.

The dual of the MAC is the broadcast channel (BC) as shown in Figure 1.4. The input to the BC is generated by a single transmitter and the BC output has intended to physically separate users. The BCs are the opposite direction of the MACs. However, some transmission systems may have BC but no reverse MAC like channel such TV and radio broadcasting. When a certain user is decomposed into more than one user with cooperating receivers, then the channel is no longer a BC. If the user is decomposed into two or more isolated receivers then the channel remains BC with increased users. The BC is shown in Figure 1.4 as the channel = $[\mathrm{h}_K \ldots\ldots \mathrm{h}_2\ \mathrm{h}_1]^T$.

The K-user interference channel (IC) is depicted in Figure 1.5 supporting K transmitters and K receivers. It is assumed that the receivers have independent additive complex white Gaussian noise with zero mean and given variance.

Each transmitter/receiver is equipped with a single antenna and it is assumed the transmitted signal is transported over a frequency flat channel. The flat channel gain received at the i^{th} receiver coming from the j^{th} transmitter is denoted as $\mathrm{h}_{i,j}$ and the channel gain of the i^{th} desired signal is $\mathrm{h}_{i,i}$, where $i = j$. The interfering channel gain is $\mathrm{h}_{i,j}$ where $i \neq j$. Accordingly, the interfering channel **H** can be mathematically modeled as K x K square matrix presented in (1.1a). The diagonal elements of the **H** matrix give the channel gain of the desired signals, and the non-diagonal elements are the interfering signal gain.

$$\mathbf{H} = \begin{bmatrix} \mathrm{h}_{K,K} - - - - - - - - \mathrm{h}_{K,1} \\ \mathrm{h}_{K-1,K} - - - - - - - - \mathrm{h}_{K-1,1} \\ - - - - - - - - - - - - - \\ - - - - - - - - - - - - - \\ - - - - - - - - - - - - - \\ - - - - - - - - - - - - - \\ \mathrm{h}_{2,K} - - - - - - - - - - - \mathrm{h}_{2,1} \\ \mathrm{h}_{1,K} - - - - - - - - - - - \mathrm{h}_{1,1} \end{bmatrix} \tag{1.1a}$$

When the interference channel operates in an additive Gaussian noise environment, the received K signal vector **y** can be expressed as,

$$\mathbf{y} = \mathbf{H}\,\mathbf{x} + \mathbf{n} \tag{1.1b}$$

where the K-elements vector **x** represents the transmitted signals and the K elements vector **n** represents the K-receivers complex white noise.

A key merit in multiuser channels [4] is the maximum available data rates for each user. Obviously, the latter is a function of the data rates used by the other users sharing the channel. The MAC and BC achievable rates are described in the so-called *capacity regions*.

1.4.1 The Capacity Region for Multiuser Channel

The multiuser capacity region can be determined by a search algorithm. However, the algorithm has to search over all $(K!)^K$. The search complexity can be managed for two or three users but for sizable number of users it demands a lengthy and complex computation. The MAC input is the collection of individual users' inputs made into a single input vector. The MAC and mutual information for the MAC channel from **x** to **y**, i.e. $\mathbf{I}(\mathbf{x};\mathbf{y})$ brings together the sum of users' rates. The sum-rate has no priority to the distribution of the rates among the users on the MAC. Using the chain rule of mutual information, we get Eq. (1.2):

$$\mathbf{I}(\mathbf{x};\mathbf{y}) = \mathrm{I}(\mathrm{x}_1;\mathbf{y}) + \mathrm{I}(\mathrm{x}_2;\mathbf{y}/\mathrm{x}_1) + \dots\dots\dots\dots + \mathrm{I}(\mathrm{x}_K;\mathbf{y}/[\mathrm{x}_1,\mathrm{x}_2,\dots\dots\dots,\mathrm{x}_{K-1}])$$

$$\mathbf{I}(\mathbf{x};\mathbf{y}) = \sum_{k=1}^{K} \mathrm{I}(\mathrm{x}_k;\mathbf{y}/[\mathrm{x}_1,\mathrm{x}_2,\dots\dots\dots,\mathrm{x}_{k-1}]) \tag{1.2}$$

The BC is a multiuser system; its rates limit is given by a *capacity region*. The latter is the region that contains all user rates that are simultaneously achievable. The boundary of the capacity region corresponds to the maximum sum-rates that a transmitter can transmit messages to the receivers. In other words, the capacity region is the union of all the achievable rate vectors. Clearly, there should be a coding scheme for the receivers to be able to correctly decode their messages. An efficient decoding system is that the decoding errors go minimal as the coder block length increases.

In single-antenna BC, the sum-rate can be achieved by transmitting only to the 'strongest' receiver. Such a transmission scheme is called an 'opportunistic transmission'. A capacity region is for a general BC call for arranging the different users in the BC from the strongest to the weakest channel. Such a BC is referred to as a degraded channel, and users can be ordered according to their signal power to noise power ratio (SNRs, better known as signal-to-noise ratio). The commonly used coding scheme in the degraded Gaussian BC is the superposition coding.

In superposition, the BC transmits to all users intended receivers simultaneously and the receivers' messages are superimposed on top of one another. The receivers decode the signals starting from the weakest receiver to the strongest receiver. In other words, the weakest receiver decodes only its own message, while the strongest receiver decodes all the messages.

Let us consider a Gaussian noise BC system comprising a transmitter provided with M antennas and there are n users (receivers), each with N antennas. The channels to each user are assumed with coherence transmission interval T, i.e. the channels remain constant for T channel uses. After T uses, the channels change to an independent set of values and remain constant for another coherence interval. During the coherence interval, the signal received by the i^{th} user is

$$\mathbf{y}_i(t) = \sqrt{P}\,\mathbf{H}_i\,\mathbf{s}(t) + \mathbf{n}_i \tag{1.3}$$

where $\mathbf{y}_i(t) \in \mathbb{C}^{\mathrm{N\,x\,1}}$ is the vector of received signals at time $t = 1, \dots, \mathrm{T}$, P is the total transmit power, $\mathbf{H}_i \in \mathbb{C}^{\mathrm{N\,x\,M}}$ is the channel matrix constant during coherence interval, $\mathbf{s}(t) \in \mathbb{C}^{\mathrm{M\,x\,1}}$ is the transmit symbol, $\mathbb{E}\|\mathbf{s}(t)\|^2 = 1$, and $\mathbf{n}_i \in \mathbb{C}^{\mathrm{N\,x\,1}}$ is additive Gaussian noise with random variable entries with zero mean and unit variance.

1.4.2 Analysis of Degraded BC with Superposition Coding

We assume that there is only a single transmit antenna (M = 1), during each coherence transmission interval T. In this scenario, then the users can be ordered according to their SNRs [5]. The channel to the i^{th} user can be expressed as $\mathbf{h}_i \in \mathbb{C}^{N \times 1}$ where $\mathbf{h}_i = [\mathrm{h}_{1,i}\ \mathrm{h}_{2,i} \ldots \ldots \ldots \ldots \mathrm{h}_{N,i}]^T$ and $|\mathbf{h}_i|^2$ is two-norm of the vector, i.e. $|\mathbf{h}_i|^2 = \mathrm{h}_{i,1}^2 + \mathrm{h}_{i,2}^2 + \cdots + \ldots \mathrm{h}_{i,N}^2$. Assuming the receiver Gaussian noise variance is unit, the $SNR_i = |\mathbf{h}_i|^2$. Ordering the users according to their SNR can be expressed as

$$|\mathbf{h}_1|^2 \geq |\mathbf{h}_2|^2 \geq \ldots \ldots \ldots \ldots \geq |\mathbf{h}_n|^2 \tag{1.4}$$

Since the BC system is degraded, the capacity region can be achieved by *superposition coding and interference cancellation*. The key idea here is to divide the transmit signal into n independent components containing the messages of each user:

$$\sqrt{P}\, \boldsymbol{s}(t) = \sum_{i=1}^{n} \sqrt{P_i}\ \ s_i(t) \tag{1.5}$$

such that $\sum_{i=1}^{n} P_i = P$. Each $s_i(t)$ is constructed from an independent Gaussian codebook. Assuming the channels are not fading during the coherence interval and therefore $\mathbf{h}_i$s are all fixed, we arrange the powers and rates such that the weakest user can only decode its own message, whereas stronger users can decode their own and all weaker messages. This means that the set of rates that can be achieved are

$$R_1 < \log(1 + P_1 |\mathbf{h}_1|^2) \tag{1.6a}$$

$$R_2 < \log\left(1 + \frac{P_2 |\mathbf{h}_2|^2}{1 + P_1 |\mathbf{h}_2|^2}\right) \tag{1.6b}$$

$$\ldots \ldots \ldots \ldots \ldots \ldots$$

$$\ldots \ldots \ldots \ldots \ldots \ldots$$

$$R_j < \log\left(1 + \frac{P_j |\mathbf{h}_j|^2}{1 + \sum_{i=1}^{j-1} P_i |\mathbf{h}_j|^2}\right) \tag{1.6j}$$

$$\ldots \ldots \ldots \ldots \ldots \ldots$$

$$\ldots \ldots \ldots \ldots \ldots \ldots$$

$$R_n < \log\left(1 + \frac{P_n |\mathbf{h}_n|^2}{1 + \sum_{i=1}^{n-1} P_i |\mathbf{h}_n|^2}\right) \tag{1.6n}$$

When the channels are fading, the users' ordering becomes a random process. In this case**,** coding over large block lengths that are adapted to the fading states on a block by block basis is required. We now investigate the ergodic capacity region when the channels are fading. The BC ergodic capacity region is defined in [6] as the set of all users' average rates that can be achieved with negligible probability of error when the average is taken with respect to all fading states. The ergodic (Shannon) sum-rate capacity is

$$\sum_{i=1}^{n} R_i \leq C_{sum} = \mathbb{E}\left\{ \max_{P_i \geq 0, \sum_{i=1}^{n} p_i = P} \log(1 + P_i |\mathbf{h}_i|^2) \right\}$$

$$\sum_{i=1}^{n} R_i \leq C_{sum} = \mathbb{E}\left\{ \log(1 + P \max_{1 \leq i \leq n} |\mathbf{h}_i|^2) \right\} \tag{1.7}$$

where the expectation is over the fading channels $\mathbf{h}_i$. It is thus clear that the scheme that maximizes the sum rate is the one that, during each coherence interval, transmits only to the user with the best SNR (i.e. only one of the $P_i i$'s is nonzero). This is referred to as *opportunistic transmission*.

1.4.3 The Capacity Region for Multiuser MIMO Channel

Unlike the single antenna system model we considered in deriving the channel matrix expressed in (1.1a), we now explore a cellular system with the BS equipped with M antennas and each of the K (mobile) users has N antennas. We examine the two-essential multiuser MIMO channels in the cellular system: the MIMO multiple access channel (MAC) and the MIMO broadcast channel (BC). The downlinks (DLs) of the system constitute the MIMO BC and the uplinks (ULs) constitute the MIMO MAC. The DL channel matrix from the BS to the i^{th} user is denoted as $\mathbf{H}_i$. Assuming reciprocity hold for TDD transmission protocol, then the UL matrix of the i^{th} user is $\mathbf{H}_i^H$.

Let $\mathbf{u}_k \in \mathbb{C}^{\mathrm{N \times 1}}$ represents a typical k^{th} user transmitted signal on the MAC, i.e. $\mathbf{u}_k$ is a vector of N elements, each element symbolizes the signal component transmitted from each of the user N antennas. Denote the received signal $\mathbf{r} \in \mathbb{C}^{\mathrm{M \times 1}}$ and the noise vector $\mathbf{w} \in \mathbb{C}^{\mathrm{M \times 1}}$ where $\mathbf{w} \sim \tilde{\mathcal{N}}(0, \mathbf{I})$ is circularly symmetric complex Gaussian with identity covariance. The received signal at the BS is,

$$\mathbf{r} = \mathbf{H}_1^H \, \mathbf{u}_1 + \mathbf{H}_2^H \, \mathbf{u}_2 + \dots \dots \dots \dots + \mathbf{H}_K^H \, \mathbf{u}_K + \mathbf{w} \tag{1.8}$$

where the conjugate transpose of the channel is denoted as $\mathbf{H}^H$. Expression (1.8) can be re-written as,

$$\mathbf{r} = [\mathbf{H}_1^H \;\; \mathbf{H}_2^H \;\; \dots \dots \dots \dots \mathbf{H}_K^H \;] \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ - \\ - \\ \mathbf{u}_K \end{bmatrix} + \mathbf{w} \tag{1.9}$$

Each user transmit power is constraint to P_k. The transmit covariance matrix of user k is defined to be $\mathbf{Q}_k$:

$$\mathbf{Q}_k \triangleq \mathbb{E}[\mathbf{u}_k \, \mathbf{u}_k^H] \tag{1.10}$$

The power constraint can be expressed as

$$\mathrm{tr}(\mathbf{Q}_k) \leq \mathrm{P}_k \text{ for } k = 1, \dots \dots \dots, \mathrm{K}.$$

On the other hand, let us denote the transmitted signal from the BS towards the user over the BC be $\mathbf{x} = \epsilon\mathbb{C}^{1 \times \mathrm{M}}$. Let $\mathbf{y}_k \; \epsilon\mathbb{C}^{\mathrm{N \times 1}}$ be the received signal at the k^{th} user's receiver. The noise at receiver k is represented by $\mathbf{n}_k \epsilon\mathbb{C}^{\mathrm{N \times 1}}$ and is assumed to be circularly symmetric complex Gaussian noise $\mathbf{n}_k \sim \mathcal{N}(0, \mathbf{I})$. The received signal of the k^{th} user is given in Eq. (1.11):

$$\mathbf{y}_k = \mathbf{H}_k \, \mathbf{x} + \mathbf{n}_k \tag{1.11}$$

The transmit covariance matrix of the input signal is

$$\Sigma_{\mathbf{x}} \triangleq \mathbb{E}[\mathbf{x} \, \mathbf{x}^H] \tag{1.12}$$

The BS average power is constraint to P, which is expressed as $\mathrm{tr}(\Sigma_{\mathbf{x}}) \leq P$.

1.4.4 The MIMO MAC Capacity Region

We assume the MAC is nonvarying during the analysis. The capacity region of an MAC can be modelled as the convex closure of the union of rate regions corresponding to the product of every input distribution $p(u_1) \dots \; \dots \; \dots . . \, p(u_K)$ satisfying the user-by-user power constraints.

However, considering Gaussian MIMO MAC, it is appropriate to consider the Gaussian elements so the entire convex operation is not needed. Accordingly, for any set of powers

$$\mathbf{P} = (P_1, \ldots \ldots \ldots .., P_K) \tag{1.13}$$

The capacity of the MIMO MAC is given as

$$C_{MAC}(\mathbf{P}; \mathbf{H}^H) \triangleq \cup_{\{\mathbf{Q}_i \geq 0, \mathrm{tr}(\mathbf{Q}_i) \leq P_i \; for\ all\ i\}} \left\{ \begin{array}{c} (R_1, \ldots \ldots ., R_K) : \\ \sum_{i \in S} R_i \leq \frac{1}{2} \log|\mathbf{I} + \mathbf{H}_i^H \mathbf{Q}_i \mathbf{H}_i| \\ for\ all\ S \subseteq \{1, \ldots ., K\} \end{array} \right\} \tag{1.14}$$

The i^{th} user transmits a zero-mean Gaussian signal with spatial covariance matrix $\mathbf{Q}_i$. Each set of covariance matrices $(\mathbf{Q}_1, \ldots \ \ldots \ \ldots \ \ldots .., \mathbf{Q}_K)$ corresponds to a K-dimensional polyhedron. In geometry, a polyhedron is simply a 3D solid that consists of a collection of polygons, usually joined at their edges.

The MAC capacity region is equal to the union over all covariance matrices satisfying the power constraints of all such polyhedrons. The corner points of each pentagon can be achieved by successive decoding where users' signals are successively decoded and subtracted out of the received signal (i.e. successive interference cancellation). To explain the MAC capacity region in depth, let us consider a simple two-users scenario. Each set of covariance matrices corresponds to a pentagon, similar in form to the capacity region of the scalar Gaussian MAC.

The corner point B corresponds to decoding user 2 first (i.e. in the presence of interference from user 1) and decoding user 1 last (without interference from user 2). The performance of user 1 can achieve a single user rate as

$$R_1 = \log|\mathbf{I} + \mathbf{H}_1^H \mathbf{Q}_1 \mathbf{H}_1| \tag{1.15}$$

The slope A-B corresponds to the sum rate:

$$R_1 + R_2 \leq \log|\mathbf{I} + \mathbf{H}_1^H \mathbf{Q}_1 \mathbf{H}_1 + \mathbf{H}_2^H \mathbf{Q}_2 \mathbf{H}_2| \tag{1.16a}$$

At corner point B, the rate is

$$R_1 + R_2 = \log|\mathbf{I} + \mathbf{H}_1^H \mathbf{Q}_1 \mathbf{H}_1 + \mathbf{H}_2^H \mathbf{Q}_2 \mathbf{H}_2| \tag{1.16b}$$

$$R_2 = \log|\mathbf{I} + \mathbf{H}_1^H \mathbf{Q}_1 \mathbf{H}_1 + \mathbf{H}_2^H \mathbf{Q}_2 \mathbf{H}_2| - R_1 \tag{1.16c}$$

$$R_2 = \log\left|\frac{\mathbf{I} + \mathbf{H}_1^H \mathbf{Q}_1 \mathbf{H}_1 + \mathbf{H}_2^H \mathbf{Q}_2 \mathbf{H}_2}{\mathbf{I} + \mathbf{H}_1^H \mathbf{Q}_1 \mathbf{H}_1}\right| \tag{1.16d}$$

$$R_2 = \log\left|\mathbf{I} + \frac{\mathbf{H}_2^H \mathbf{Q}_2 \mathbf{H}_2}{\mathbf{I} + \mathbf{H}_1^H \mathbf{Q}_1 \mathbf{H}_1}\right| \tag{1.16e}$$

$$R_2 = \log|\mathbf{I} + (\mathbf{I} + \mathbf{H}_1^H \mathbf{Q}_1 \mathbf{H}_1)^{-1} \mathbf{H}_2^H \mathbf{Q}_2 \mathbf{H}_2| \tag{1.16f}$$

Corner point A concurs with decoding user 1 first (in the presence of interference from user 2) and decoding user 2 last in interference free signal. The performance of user 2 achieves a single user rate as

$$R_2 = \log|\mathbf{I} + \mathbf{H}_1^H \mathbf{Q}_1 \mathbf{H}_1| \tag{1.17}$$

Using similar method to above to derive user 1 rate as

$$R_1 = \log|\mathbf{I} + (\mathbf{I} + \mathbf{H}_2^H \mathbf{Q}_2 \mathbf{H}_2)^{-1} \mathbf{H}_1^H \mathbf{Q}_1 \mathbf{H}_1| \tag{1.18}$$

Accordingly, successive interference cancellation reduces a complex multiuser detection problem into a series of single-user detection processes. The capacity region of an MIMO MAC for users with single transmit antenna each when two users sharing the band is shown in Figure 1.6. In a scenario N = 1, the covariance matrix of each transmitter is a scalar quantity equal to the transmitted power, and each user transmits at full power. Thus, the capacity

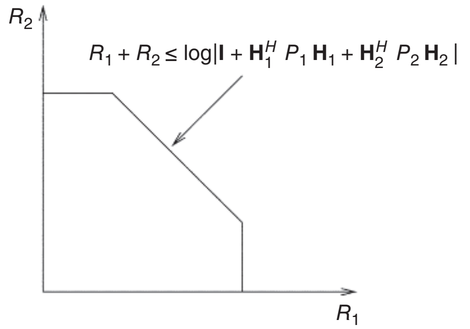

Figure 1.6 Capacity region of MIMO MAC for N = 1 [7].

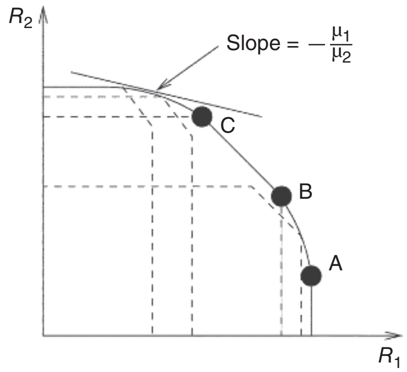

Figure 1.7 Capacity region of MIMO MAC for N > 1 [7].

region for a K-user MAC for N = 1 is the set of all rate vectors $(R_1, \dots \dots ., R_K)$ satisfying equation (1.19):

$$\sum_{i \varepsilon S} R_i \leq \frac{1}{2} \log|\mathbf{I} + \mathbf{H}_i^H \boldsymbol{p}_i \, \mathbf{H}_i| \, for\ all\ S \subseteq \{1, \dots ., K\} \tag{1.19}$$

This reduces the two-user capacity region to a simple pentagon as shown in Figure 1.7. While N > 1, a union must be taken overall covariance matrices. Evidently, different sets of covariance matrices maximize users' rates and are different from the set of covariance matrices that maximize the sum rate. So, let us examine a MAC capacity region when users are equipped with N > 1. As we already emphasised, the capacity region is equal to the union of pentagons. Each pentagon corresponding to a different set of transmit covariance matrices as depicted in Figure 1.7. A few are shown with dashed lines in the figure.

The generic boundary of the capacity region is curved, and only straight lines at the sum rate point. Each point on the curved part of the boundary is achieved by a different set of covariance matrices.

At point A, user 1 is decoded last and achieves single-user capacity, choosing $\mathbf{Q}_1$ as a water-fill of the channel $\mathbf{H}_1$ independent of $\mathbf{H}_2$ or $\mathbf{Q}_2$. User 2 is decoded first in the presence of interference from user 1, $\mathbf{Q}_2$ is chosen as a water-fill of the channel $\mathbf{H}_2$ and the decoding is degraded by the presence of interference from user 1.

The corner points B and C are the two corner points of the pentagon corresponding to the sum rate with optimal covariance matrices $\mathbf{Q}_1^{\text{sum}}$ and $\mathbf{Q}_2^{\text{sum}}$. At point B, user 1 is decoded last, hence achieving a single user rate while at point C user 2 is decoded last. Thus, points B and C are achieved using the same optimal covariance matrices but different decoding orders.

Next, we investigate the features of the optimal covariance matrices $(\mathbf{Q}_1, \dots \dots \dots \dots ., \mathbf{Q}_K)$ that achieve different points on the capacity region boundary of the MIMO MAC. Since the

MAC capacity region is convex, it is well known that the boundary of the capacity region can be maximized by the function $\mu_1 R_1 + \ldots\ldots\ldots\ldots + \mu_K R_K$ overall capacity region rate vectors and for all nonnegative priorities $(\mu_1, \ldots\ldots\ldots, \mu_K)$ such that $\sum_{i=1}^{K} \mu_i = 1$ where μ is the water-fill level. Let us define a fixed set of priorities $(\mu_1, \ldots\ldots\ldots, \mu_K)$, then a point on the capacity region boundary is corresponds to the tangent to a line whose slope is defined by the priorities as shown in Figure 1.7.

The boundary of the MAC capacity region indicates that all points of the capacity region are corner points of polyhedrons corresponding to different sets of covariance matrices. Additionally, the corner point should correspond to successive decoding with interference cancellation in order of *increasing* priority, i.e. the user with the highest priority should be (interference free) decoded last. Accordingly, the problem of locating the capacity region boundary point associated with priorities assumed to be in descending order (arbitrarily renumbering the users to satisfy this condition) can be written as [7].

$$\max_{\mathbf{Q}_1,\ldots\ldots\ldots,\mathbf{Q}_K} \left(\mu_K \log\left|\mathbf{I} + \sum_{l=1}^{K} \mathbf{H}_{l=1}^{H}\, \mathbf{Q}_l\, \mathbf{H}_l\right| \right) + \sum_{i=1}^{K-1} (\mu_i - \mu_{i+1}) \log\left|\mathbf{I} + \sum_{l=1}^{i} \mathbf{H}_{l=1}^{H}\, \mathbf{Q}_l\, \mathbf{H}_l\right| \tag{1.20}$$

Maximizing the above function subject to satisfying the power constraints defines the optimal covariance matrices.

The most interesting aspect of the optimization problem is that the objective function is concave in the covariance matrices. An efficient numerical technique to find the sum rate maximizing covariance matrices, is called iterative water-filling. The technique is based on the Karush–Kuhn–Tucker (KKT) optimality conditions for the sum rate maximizing covariance matrices. The KKT optimization method is introduced in the tutorials on theory and techniques of optimization mathematics, Appendix 1.C at the end of the chapter. These conditions indicate that the sum rate maximizing covariance matrix of any user in the system should be the single-user water-filling covariance matrix of its own channel with noise equal to the actual noise plus the interference from the other $K-1$ transmitters.

1.4.5 The MIMO BC Capacity Region [6, 8, 9]

When the transmitter (i.e. BS) has more than one antenna, the Gaussian BC is typically nondegraded. This is because users receive different strength signals from different transmit antennas, which make it difficult to order the users according to their SNRs. In this section, we assume the MIMO BC is a constant channel and BS supported with M antennas, and each receiver is equipped with N antennas. We consider the optimum coding scheme, i.e. dirty paper coding (DPC) [10, 11] applied at the MIMO transmitter that achieves the sum rate capacity region of BC. The basic concept of DPC is that if the transmitter has perfect knowledge of additive Gaussian interference in the channel, then the capacity of the channel is the same as if there was no additive interference. The DPC allows a known interference to be pre-subtracted at the transmitter through selecting codewords for different receivers, keeping the transmit power unchanged.

Let us consider the MIMO BC with DPC applied to two users: user 1 is referred to single antenna receiver 1 and user 2 to single antenna receiver 2. First, the transmitter arbitrary selects a codeword (i.e. x_1) for receiver 1. The transmitter then chooses a codeword for receiver 2 (i.e. x_2) with full knowledge of the codeword intended for receiver 1.

Therefore, the codeword of user 1 can be pre-subtracted and receiver 2 does not see the codeword intended for receiver 1 as interference. Similarly, when three users are considered, the codeword for receiver 3 is chosen such that receiver 3 does not see the signals intended for receivers 1 and 2 (i.e., presubtract $\mathbf{x}_1 + \mathbf{x}_2$) as interference. This presubtraction process continues for all K receivers.

If user $u(1)$ is encoded first, followed by user $u(2)$, etc., the rate vector can be achieved:

$$R_{u(i)} = \log \frac{\left| \mathbf{I} + \mathbf{H}_{u(i)} \left(\sum_{k \le u(i)} \boldsymbol{\Sigma}_k \right) \mathbf{H}_{u(i)}^H \right|}{\left| \mathbf{I} + \mathbf{H}_{u(i)} \left(\sum_{k < u(i)} \boldsymbol{\Sigma}_k \right) \mathbf{H}_{u(i)}^H \right|} \qquad i = 1, \ldots \ldots \ldots \ldots, K \tag{1.21}$$

In deriving Eq. (1.21), we use the identical method that is used in deriving Eq. (1.16). The dirty paper region is defined as the convex hull, (i.e., convex boundary in which a given set of rate points is contained) of the union of all such rates vectors over covariance matrices $\boldsymbol{\Sigma}_1, \ldots \ldots \ldots \ldots \ldots, \boldsymbol{\Sigma}_K$ such that $\mathrm{tr}(\boldsymbol{\Sigma}_1 + \ldots \ldots + \boldsymbol{\Sigma}_K) \le P$ and overall permutations $(u(1), \ldots \ldots \ldots \ldots, u(K))$, the DPC capacity region of BC C_{DPC} $(P, \mathbf{H})$ is given by (1.22) and $\mathbf{R}_{u(k)}$ is given by equation (1.21),

$$C_{\mathrm{DPC}}(P, \mathbf{H}) \triangleq Co \left(\bigcup_{u, \boldsymbol{\Sigma}_k} \mathbf{R}_{u(k)} \right) \tag{1.22}$$

where Co implies a coordinated joint detection system. The DPC rate region for two-user channel $\mathbf{H}$ with power constraint $P = 10$ is depicted in Figure 1.8. The transmitter operates with two antennas and each receiver has a single antenna. The channel $\mathbf{H}$ matrix is

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_1 \\ \mathbf{H}_2 \end{bmatrix} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

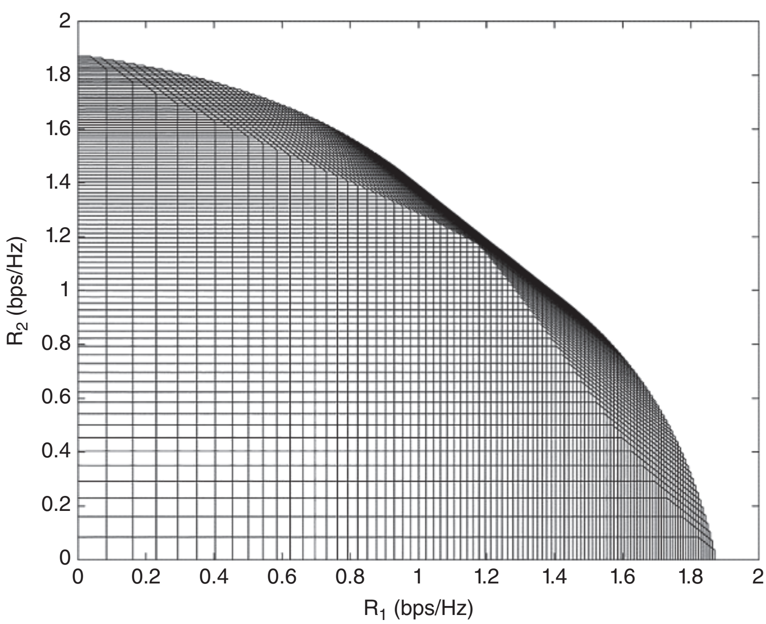

Figure 1.8 Dirty paper capacity region, $\mathbf{H}_1 = [1\ 0.5]$, $\mathbf{H}_2 = [0.5\ 1]$, $P = 10$ [7].

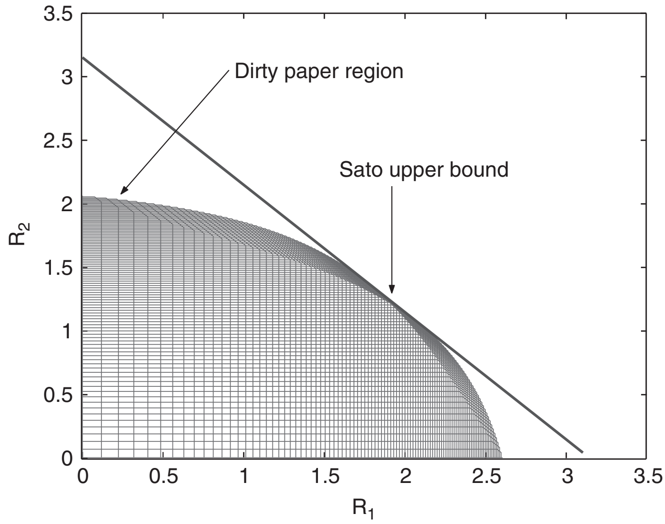

Figure 1.9 Dirty paper broadcast region – achievability and converse [13].

It is worth noting that Sato [12] presents an upper bound on the capacity region of general BC. This bound utilizes the capacity of the coordinated system as defined above. The capacity of the coordinated system is an upper bound on the BC sum rate $C_{DPC}(P, \mathbf{H})$.

The DPC BC capacity regions for a two single antenna receivers BC and two antenna transmitters as well as the Sato upper bound and the power constraint $P = 10$ are shown in Figure 1.9. The BC channel matrix H is given by

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_1 \\ \mathbf{H}_2 \end{bmatrix} = \begin{bmatrix} 1 & 0.5 \\ 0.2 & 0.8 \end{bmatrix}$$

1.5 Fading Wireless Channels

In wireless (mobile) communications, the transmitted signal strength is varying due to various variables such as time, location, signal obstruction, and so on. The signal variation causes the signal in most cases to randomly change, and this is called fading. There are different classifications of fading such as flat fading; fast fading, slow fading; and frequency selective fading. Generally, the signal energy propagation can be of two types: direct line of sight energy propagation between transmitter and receiver (LOS), but the most common type of propagation is due to scattering of the energy by objects in the way and the scattering mechanism generates multiple copies of the signal with different amplitude and phase (i.e. multipath signal) arriving at the receiver. The signal power level range over time is called *multipath intensity profile,* which is defined by multipath spread. The reciprocal of the multipath spread is called the channel coherence bandwidth. When the signal bandwidth is much less than the coherence bandwidth of the channel, the amplitudes of the signal frequency components are flat, and such channels are called frequency nonselective or flat fading. The flat fading is of two types: slow fading occurs if the transmitted symbol duration is much smaller than the coherence time of the channel (channel coherence time is the reciprocal of channel Doppler spread). On the contrary, a rapidly fading channel is one in which the symbol duration is much greater than the channel coherence time. When the transmitted signal bandwidth is greater than the channel coherence bandwidth, the frequency components exceed the channel coherence bandwidth, and are exposed to different gains and phase shifts, and the channel is said to be a frequency selective fading channel [14].

In most publications, researchers assume the transmission channel is subjected to flat fading when considering the analysis/simulations, and this is what we have assumed in this book. The capacity region analysis for flat-fading channels demands a number of assumptions regarding the perfect knowledge of CSI at transmitter / receiver or both. The analysis of the sum rate regions is dealt with in detail in the following chapters.

1.6 Multicell MIMO Channels

In Section 1.4, a single cell is assumed for analysing the MAC and BC system. Nonetheless, a cellular system, by definition, entails many cells [15]. Additionally, the basic characteristic of wireless energy propagation in a cellular system is not limited to within a single cell. This means that users as well as BSs in adjacent cells encounter interference from each other. However, BSs are not mobile, they can communicate through optical fibre links capable of high data rates. This gives a good opportunity for BSs to cooperate on processing different users' signals. However, capacity of the cellular network, with multiple users and multiple antennas, is a very difficult problem. The single cell multiuser capacity is important and can be used as a benchmark to assess the efficiency of a practical multicell system in similar approach to the capacity of a single user link serves to measure the performance merits of practical multiusers scheme. The assumption of perfect BSs cooperation is the key to the extension of a single cell with results to multiple cells systems. Theoretically, multiple BSs cooperation can be treated as composite BS with physically distributed antennas.

Explicitly, let use consider a group of B coordinated cells, each with M antennas and serving K mobiles, each with N antennas. Let us also define $\mathbf{H}_{i,j} \in \mathbb{C}^{\mathrm{N \times M}}$ to be the *DL* channel of user i from base station j, then the *composite* DL channel of user i is $\mathbf{F}_i = [\mathbf{H}_{i,1} \ldots \; \ldots . \mathbf{H}_{i,B}]$ and the composite UL channel is given as $\mathbf{F}_i^H$. The received signal of user i can then be written as $\mathbf{Y}_i$,

$$\mathbf{Y}_i = \mathbf{F}_i \mathbf{W} + \mathbf{n}_i \tag{1.23}$$

where $\mathbf{W}$ is the composite transmitted signal defined as $\mathbf{W}^T = [\mathbf{W}_1^T \ldots \ldots \ldots \mathbf{W}_B^T]$. Here, we let $\mathbf{W}_j$ *to* represent the transmit signal from base j.

On the DL, since the base stations can cooperate perfectly, DPC can be used over the entire network transmit signals (i.e. across base stations) in a straightforward manner. This requires perfect data and *power cooperation* between the base stations. Denote $\mathbf{x}_{i,j}$ to represent the transmitted vector for user i from base station j, the composite transmit vector intended for user i is $\mathbf{X}_i^T = [\mathbf{x}_{i,1}^T \ldots \ldots \ldots . \mathbf{x}_{i,B}^T]$. Thus, the composite covariance of user i is defined as $\Sigma_i = \mathbb{E}[\mathbf{X}_i \, \mathbf{X}_i^T]$. The covariance matrix of the entire transmitted signal is $\Sigma_{\mathbf{x}} = \sum_{i=1}^{K} \Sigma_i$.

Although data cooperation is a justifiable assumption, in practice each BS has its own power constraint. The *per-BS power constraint* can be expressed as $\mathbb{E}[\mathbf{W}_j^H \, \mathbf{X}_j] \leq P_j$, where P_j is the power constraint at BS j. Consequently, power cooperation, or pooling the transmit power for all the BSs over the BC to have one overall transmit power constraint, is not realistic. On the UL the BSs are only receiving signals and, therefore no power cooperation is required.

1.7 Green Wireless Communications for the Twenty-First Century

Practically every human on this planet uses mobile communications. This growth in mobile phone subscribers comes with enormous energy consumption due to mobile networks.

The intensive concerns for the environment and the impacts of global warming on the climate necessitate an appropriate action on the energy efficiency of these systems. It is estimated that 80% of the energy needed for the operation of a wireless cellular network is consumed at the BS sites. Furthermore, to increase the energy efficiency in mobile communication systems there is a need to reduce the BS power consumption at instants of low network loads. Researchers around the world put in great efforts to find a solution to climate change challenge. One of these efforts is the EU funded research project Energy Aware Radio and network technologies (EARTH) [16].

When considering green energy for wireless communications, the first few questions that come to mind are how much energy is required to operate a wireless network and how to model the *energy efficiency evaluation framework* (E^3F). The EARTH E^3F suggested a power model for various BS types in addition to large-scale long-term traffic models. The BS power model relates the RF energy radiated at the antenna elements to the total supply power of a BS site. The proposed traffic model mirrors the spatial distribution of the traffic demands over large geographical regions in addition to urban and rural areas, as well as time-based variations between peak and off-peak hours. The performance of a wireless network at the system level can be evaluated for system throughput, quality of service (QoS) metrics, and cell edge user throughput.

The EARTH E^3F is depicted in Figure 1.10**.** The EU project considered for the performance modelling comprising various types of BSs, wireless channels, mobile devices and their power consumption.

The global E^3F is illustrated in three steps. The system evaluations over small-scale short-term scenarios models the statistical traffic as a File Transfer Protocol (FTP) file download or Voice over Internet Protocol (VoIP) calls. In addition, small-scale deployment includes

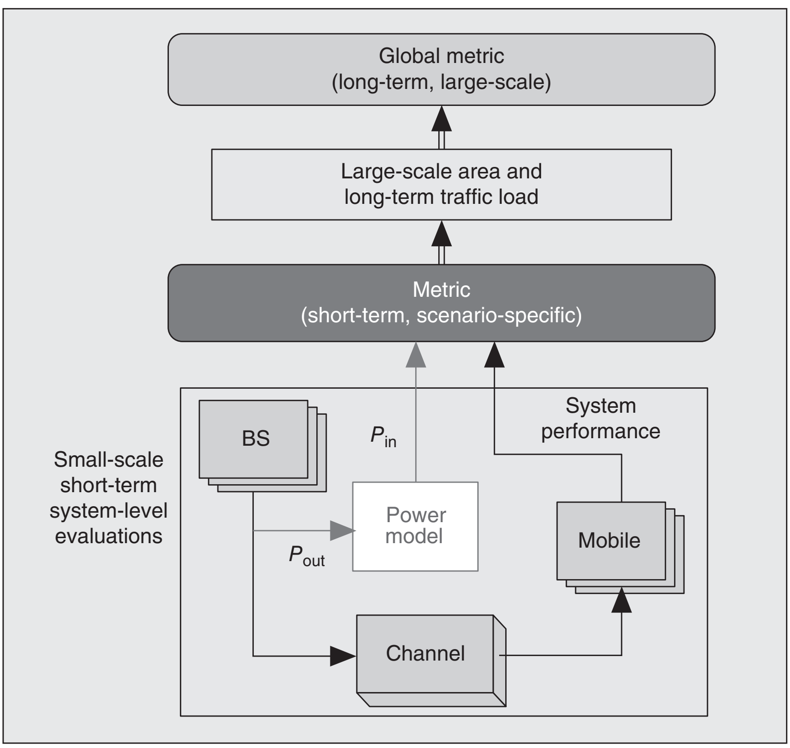

Figure 1.10 EARTH Energy Efficiency Evaluation Framework (E^3F) [16].

urban macro-cell network incorporating 57 cells with uniformly distributed users to enable short-term system-level evaluations. These evaluations are assisted by a BS power consumption model. The small-scale short-term evaluations are broadened to a global scale to cover countrywide geographical areas and ranging over a full day or week. The small-scale short-term evaluations cover cities, suburbs, and villages, and a set of network loads. The evaluations' results include energy consumption, throughput, QoS over a range of traffic loads. For a given daily/weekly traffic profile of each deployment, performance metrics over a day/week are determined from the short-term evaluation results and extended to large-scale deployment over large geographical areas.

1.7.1 Network Power Consumption Model

The power consumption model creates a connection between the system components and system levels that enable the calculation of the energy saving on specific components required to enhance the energy efficiency at the network level. System components are characterized by the BS type in terms of radiated power (BS output power), constraints on cost, and BS size. So, we need a power model designed for an individual BS type. However, we can account for the power consumption based on a generalized infrastructure of all BS types, including macro, micro, pico, and femto BSs. A BS consists of a number of transceivers (TRXs) equal to the number of transmit antennas. A block diagram of a BS transceiver is shown in Figure 1.11. A TRX is composed of a power amplifier (PA), RF TRX module, BS engine encompassing UL receiver and DL transmitter section, DC-DC power supply, an active cooling system, and an AC-DC unit for connection to the main supply. In the following section, we analyse the power consumption of each component of the BS TRX.

1.7.2 Antenna Interface Losses

The influence of the antenna interface on the network power efficiency is modelled as a certain amount of losses that include antenna feed cable, antenna filters, duplexer, and matching circuits. The loss from small BS feeder losses is mostly insignificant. Generally, macro BS and its antennas are located at different geographical locations and the feeder losses typically amount to −3 dB. Nonetheless, when a RRH is deployed, the PA can be placed at the same location at the transmit antenna, and so the feeder losses can be reduced or diminished.

1.7.3 Power Amplifier (PA)

With the aim to be power-efficient, a PA average input mains power has to be maintained as close as possible to the power required for transmitting the data. There are two attainable

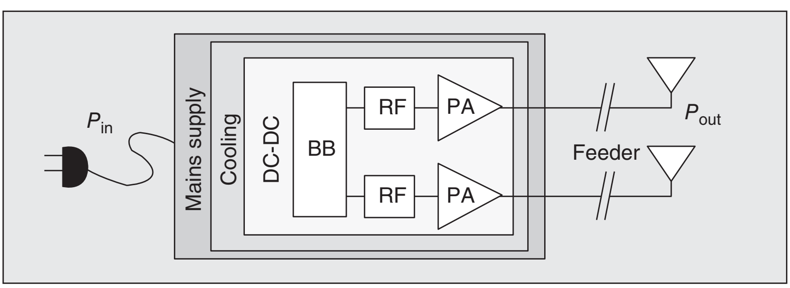

Figure 1.11 Block diagram of a base station transceiver [16].

solutions to this problem: either introduce digital predistortion (DPD) to the signal before transmission or dynamically adapt the amplitude of the input voltage of the amplifier to the transmission power requirement. A linearization scheme is commonly associated with the DPD to activate a compensation for the nonlinear part of the amplifier output. Actually, the compensation covers the adjacent channels interference (ACI) and all the precoding changes introduced with time and temperature signal characteristics. The linearization may use a Doherty PA technique, which applies two PAs driven in parallel, and the final power output is the combined output power of the two amplifiers. One amplifier is operated in the linear region in class A/B continuously until its nonlinear region. The second amplifier operates in class C to provide additional power when the amplifying device enters its the nonlinear region. Theoretically, Doherty amplifiers are capable of high efficiency. Nonetheless, when used in cellular networks working at high frequencies and obliged to output high power, Doherty amplifiers achieved an average efficiency of 25–30%. In addition, Doherty amplifiers are narrowband due to the matching required between the two internal amplifiers. DPD and Doherty techniques are applied by Huawei to construct low-power GSM BSs to operate with 500 W, which reduces CO_2 emissions considerably (claims to be equivalent to burning 1.7 tons of coal per year per BS) [12]. When considering there are 100 000 Huawei BSs operating worldwide, then the reduction of CO_2 emissions is equivalent to burning 170 000 tons of coal per year.

The second solution to enhance the PA efficiency is to adapt the amplifier voltage amplitude to the required transmit power, which was widely known as envelope-tracking technology proposed by Bell Labs in the late 1930s, but it did not take off until recently with the availability of low-power transistor amplifiers. The key concept of envelop tracking can be explained as follows: adapting the amplitude of the input voltage of the amplifier to the transmission requirement is accomplished by dynamically changing the supply voltage of the PA to match the signal to be amplified. Accordingly, envelope tracking guarantees the output of the PA is always in its most efficient operating region. The amplifier inefficiency when a constant supply of voltage applied is demonstrated in Figure 1.12. With no envelope tracking, the difference between the constant supply voltage and the RF waveform amplitude is consumed as heat by the power transistor. Figure 1.13 shows the situation when envelope tracking is used. The supply voltage (i.e. input power) is dynamically adjusted to be close to the RF output power, which makes an

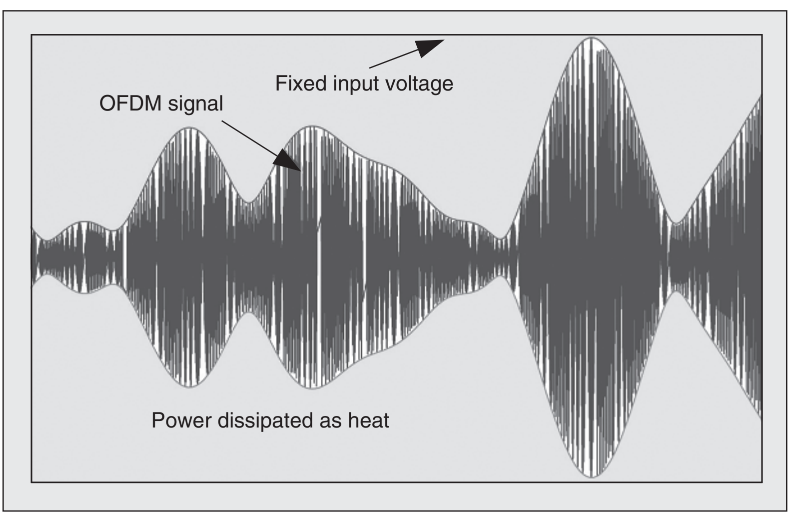

Figure 1.12 Power amplifier with constant input voltage supply [17].

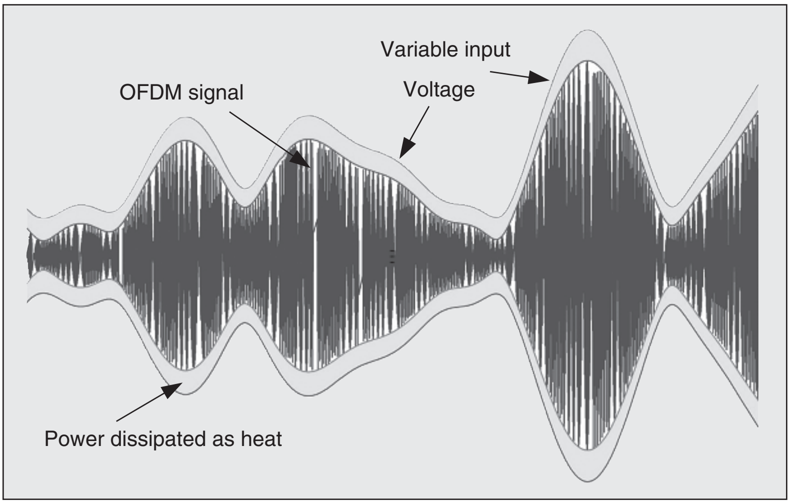

Figure 1.13 Power amplifier with input voltage supply using envelope tracking [17].

Table 1.1 Efficiency, costs, and environmental impact of a 20 000-BS network with different power amplifier technologies [17].

	Traditional technology	Doherty technology	Envelope tracking (HAT)
Power ampl. effic. (%)	15	25	45
Power consumption (MW)	51.7	27.2	16.1
Power cost (M)	\$54.3	\$28.6	\$17.0
CO_2 emission (tons)	194 600	102 400	60 800

impressive reduction in the energy converted to heat. Naturally, envelope tracking comes with CPU cost and a few watts of extra power consumption, which is negligible compared to the tens or even hundreds of watts saving when using high PAs. Table 1.1 displays the PA efficiency and its power consumption, power cost, and CO_2 emission in equivalent tons of burning coal for a large network consisting 20, 000 base stations when using traditional technology, Doherty technology, and envelop-tracking design technologies.

In practical systems, the energy-efficient solutions are related to energy consumption and *traffic load*. To achieve this goal, wireless operators have proposed the adaptation of smart software platforms to enable energy saving at BSs. For example, Nokia Siemens network employs a service quality manager (SQM) known as NetAct software traffic load capacity management at the BS. NetAct commands BS consumption automatically based on a certain setting. NetAct schedules system's capacity tuning every short interval and the actual capacity is chosen as a function of the current traffic, traffic load history, and the estimate of the traffic in the next interval. Smart software called dynamic power save is initiated by Alcatel-Lucent and claims power consumption reduction up to 30% using the possibility of turning off the amplifier power in the BS TRX based on traffic activity monitoring by BS. Ericsson employs standby operational modes for power saving estimates of 10–20%, depending on the traffic patterns. China mobile proposed a green action plan based on a low-carbon economy. The strategy is defined by the following six Rs: *Recovery & Recycle* implies recycling materials that need low-energy cost for their recycle life. *Right & Reduce* refers to small and lightweight carton design to reduce packaging and transport costs. *Returnable & Reuse* denotes the use of efficient recycling system to extend packaging material life cycle. Huawei announced a cut of over 6000 tons in CO_2 emissions when applying the six Rs strategy for a greener environment.

1.8 BS Power Model

PAs operating in the nonlinear region with input OFDM modulation (nonconstant envelope signals) generate a high level of ACI that causes considerable performance degradation at the receiver. Accordingly, the PA has to operate in the linear region (i.e. 6–12 dB below the saturation point). Nonetheless, such high operating back-off induces a reduction in the power efficiency, which converts to an increase in the PA power consumption. The objective of the analysis is to derive a power model that links the RF output power radiated at the array elements P_{out} to the total supply power at BS site P_{in}, as shown in Figure 1.10. Let us denote the power efficiency η_{PA}, PA power consumption P_{PA} and the BS feeder loss σ_{feed}. BS feeder loss can be as high as 3 dB for macro cells and can be considered negligible for smaller cells. The PA power consumption can be expressed as

$$P_{PA} = \frac{P_{out}}{\eta_{PA}\,(1 - \sigma_{feed})} \tag{1.24}$$

As we highlighted in Section 1.7.3, DPD combined with Doherty PAs can improve the power efficiency while keeping ACI closely controlled. However, extra feedback is necessary for the DPD, and additional signal processing is assumed to be essential in macro and micro BSs. In smaller BSs, the PA power consumption account for a smaller percentage of the total BS power consumption and advanced PA architectures are not required.

1.8.1 Small-Signal RF Transceiver

The transceiver consists of a transmitter and a receiver for the UL and DL communications. Linearity and blocking of the RF modules differ greatly depending on the BS type and therefore affect the RF architecture. For example, low intermediate frequency (IF) or even super-heterodyne design can be chosen for macro/micro BSs and zero IF architecture is acceptable for pico/femto BSs. The parameters that have the highest impact on the RF power consumption are: appropriate bandwidth, accepted signal to noise and distortion (SiNAD) and A/D resolution.

1.8.2 Baseband (BB) Unit

At the baseband (BB) unit, the following signal processing normally carried out: filtering, modulation / demodulation, predistortion processing, signal detection including synchronization, channel estimation, equalization, and channel coding /decoding. For large BSs, the digital BB unit encompasses the power consumed by the link to the backbone network and the MAC operation consumption. Denote the BB power consumption as P_{BB}.

1.8.3 Power Supply and Cooling

Losses encountered by the DC-DC supply, mains supply, and active cooling (i.e. using fans to reduce the heat of components) are only enforced to macro BSs. The losses may be approximated by loss factors σ_{DC}, σ_{Mains}, and σ_{cool}, respectively. Assume the BS power consumption

Table 1.2 Base station power consumption at maximum load for different LTE BS types.

		Macro	**RRH**	**Micro**	**Pico**	**Femto**
BS	Max Tx power, dBm	43.0	43.0	38.0	21.0	17.0
	(average) P_{max}, W	20.0	20.0	6.3	0.13	0.05
	Feeder loss σ_{feed}, dB	−3	0	0	0	0
PA	Back-off, dB	8.0	8.0	8.0	12.0	12.0
	Max PA out (peak), dBm	54.0	51.0	46.0	33.0	29.0
	PA eff. η_{PA}, %	31.1	31.1	22.8	6.7	4.4
Total PA, $\frac{P_{max}}{\eta_{PA}\,(1-\sigma_{feed})}$, W		128.2	64.4	27.7	1.9	1.1
RF	P_{TX}, W	6.8	6.8	3.4	0.4	0.2
	P_{RX}, W	6.1	6.1	3.1	0.4	0.3
Total RF, P_{RF}, W		12.9	12.9	6.5	1.0	0.6
BB	Radio (inner Rx/Tx), W	10.8	10.8	9.1	1.2	1.0
	Turbo code (outer Rx/Tx), W	8.8	8.8	8.1	1.4	1.2
	Processors, W	10.0	10.0	10.0	0.4	0.3
Total BB, P_{BB}, W		29.6	29.6	27.3	3.0	2.5
DC-DC, σ_{DC}, %		7.5	7.5	7.5	9.0	9.0
Cooling, σ_{cool}, %		10.0	0.0	0.0	0.0	0.0
Mains supply, σ_{ms}, %		9.0	9.0	9.0	11.0	11.0
Total per TRX chain, W		225.0	125.8	72.3	7.3	5.2
Sectors		3	3	1	1	1
Antennas		2	2	2	2	2
Carriers		1	1	1	1	1
Total N_{TRX} *chains*, P_{in}, W		1350	754.8	144.6	14.7	10.4

An LTE system with 10 MHz bandwidth and 2 × 2 MIMO configuration is assumed [16].

will grow proportionally with the number of transceivers N_{TRX}. Accordingly, the BS power consumption at maximum load (i.e. $P_{out} = P_{max}$) can be expressed as [16]

$$P_{in} = N_{TRX} \frac{\frac{P_{out}}{\eta_{PA}(1-\sigma_{feed})} + P_{RF} + P_{BB}}{(1-\sigma_{DC})(1-\sigma_{mains})(1-\sigma_{cool})} \tag{1.25}$$

The power efficiency is defined as

$$\eta \triangleq P_{out}/P_{in} \tag{1.26}$$

while the loss factor is defined:

$$\sigma \triangleq 1-\eta \tag{1.27}$$

The power consumption at maximum load for different LTE BS types as of 2010 are given in Table 1.2.

1.8.4 BS Power Consumption at Variable Load

In an LTE system DL, the BS load defined by $\frac{P_{out}}{P_{max}}$ is proportional to the amount of consumed resources, including both data and control signals. Furthermore, the BS load also depends on power control settings in terms of the spectral power density on an LTE UL transmission. Power consumption for various BS types as a function of the variable load is displayed in Figure 1.14.

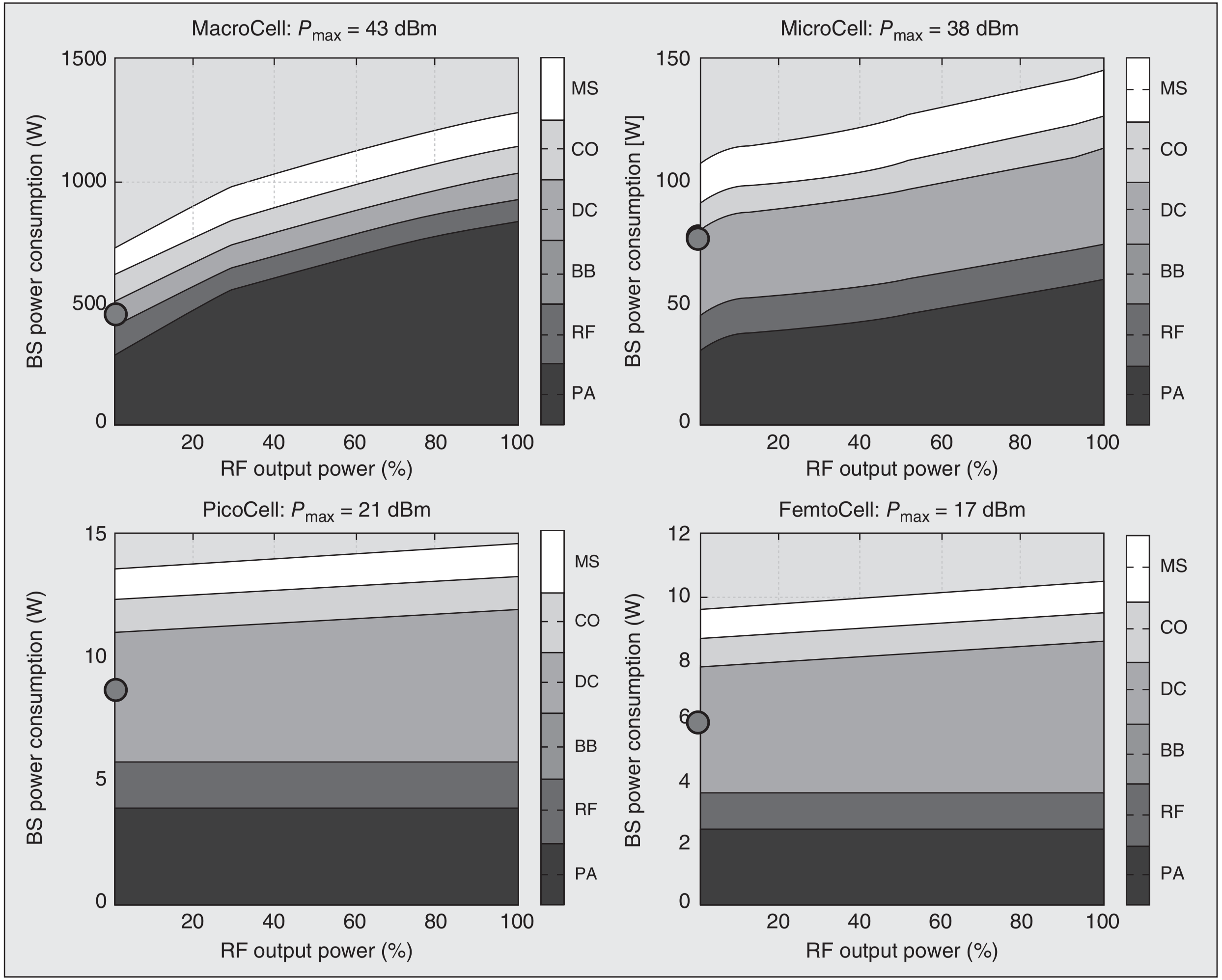

Figure 1.14 Power consumption for various BS types as a function of the load [16]. PA: power amplifier, RF: small signal RF transceiver, BB: baseband processor, DC: DC-DC converters, CO: macro BS cooling, MS: mains power supply.

Inspection of the power consumption versus BS load affirms that the PA power consumption mainly grows with BS load and the amount of growth largely depends on the BS type. Accordingly, in macro BSs the power consumption is load-dependent, but to a lesser extent for micro BSs, and load dependency is negligible for pico and femto BSs. Arguably, for low-power BSs the PA power consumption account is less than 30% of total power consumption, while for macro BSs the PA power consumption amounts to about 60% of total consumption.

1.9 Green Cellular Networks

Increasing carbon emissions brought about by mobile communication is related to energy consumption, which contributes to global warming. Thence, greening cellular networks is essential to reducing the carbon emission due to information and communication technology. Global carbon footprint of mobile communications is analysed in [18]. It predicts an increase in carbon emissions considerably more than the mobile communication traffic rates. Extraordinary research efforts in both academia, information, and communications industry are invested in reducing carbon emissions. It is well known that multicell cooperation (MCC) can be employed to considerably enhance cellular performance (i.e. throughput and coverage). Energy consumption of cellular networks is mainly due to BSs, as we discussed, and accounts for more than 50% of the network energy consumption. Accordingly, enhancing the energy efficiency of BSs is fundamental to green cellular networks. Exploiting the MCC, cellular networks energy efficiency can be improved from three aspects. First, reducing the number of active BSs needed to serve mobile subscribers in a given coverage area. The basic idea is to adapt network layout to traffic demands and switch off BSs for a certain time when users' traffic demands are below a certain threshold. When BSs are switched off, their users' services are taken care of by their neighbouring cells. Second, associate certain users with green BSs (i.e. powered by renewable energy). Subsequently, on-grid BSs reduce their service area and the off-grid BSs increase their service area. Special algorithms are designed to guide users to associate with BSs powered by renewable energy. This aspect not only enhances energy efficiency but also the operation costs. Third, exploit CoMP to raise the energy efficiency of cellular networks. This aspect creates a double whammy effect. On the one hand, with the help of MCC, energy efficiency of BSs serving cell edge users increased. On the other hand, adopting MCC expands coverage and hence reduces the number of active BSs.

Figure 1.15 depicts a number of cellular network layout adaptations according to user demands. Each BS has three sectors. We consider the central cell in Figure 1.15 to explain the layout adaptation. If most of the traffic demands in the central cell originated from sector three, and traffic demand from sectors one and two are lower than the predefined threshold, then central BS switched-off sectors one and two, and users in the switched-off sectors are served by neighbouring BSs. Figure 1.15a shows the original network layout where all BSs are active. Figure 1.15b shows the network layout where central BS switched-off sectors one and two, and users in these sectors are served by the two cells below (i.e. central BS partially switched off). In a situation when sector three traffic demands in the central cell also decreases below threshold, the central BS is completely switched off and the network layout is changed as in Figure 1.15c.

When BS is switched off, energy consumed by its elements (i.e. transceivers, processing circuits, cooling, etc.) are saved. Accordingly, network layout adaptation is capable of saving a considerable amount of energy in cellular networks. Nonetheless, there are certain constraints that have to be satisfied such as coverage requirement and QoS requirements of all users. These constraints can be guaranteed with MCC, or else the service suffers from high blocking probability and severe QoS degradation.

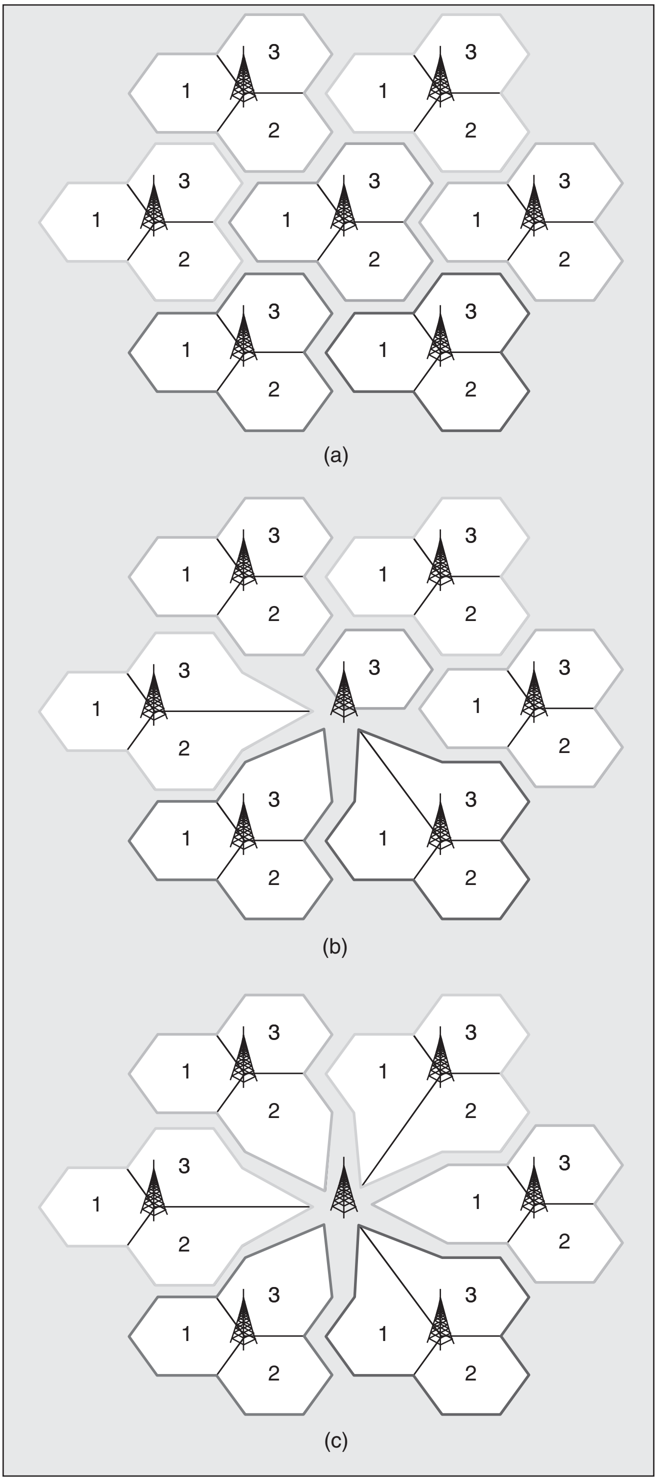

Figure 1.15 Block diagram showing network layout adaptations for macro cellular sites: (a) original network layout; (b) BS partially switched off; (c) BS entirely switched off [19].

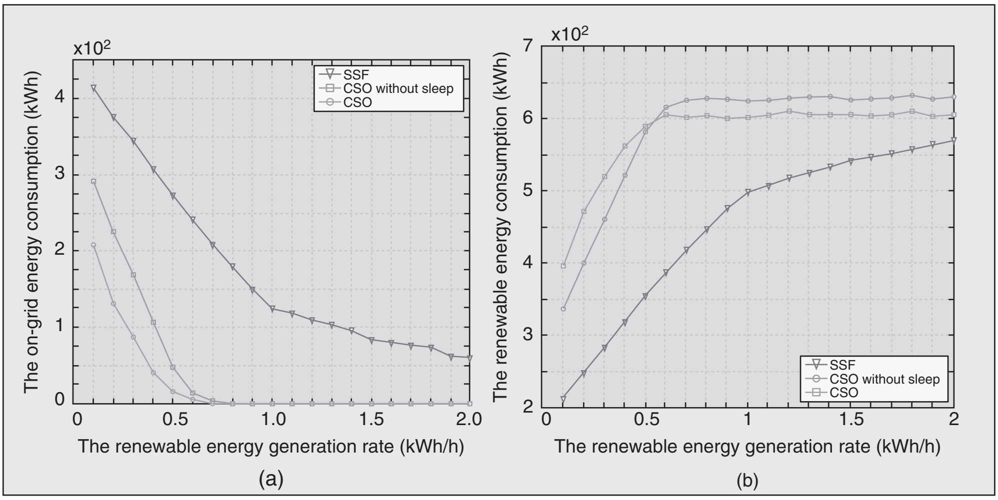

Figure 1.16 Energy consumption of the cellular network: (a) on-grid energy consumption; (b) renewable energy consumption [19].

Figure 1.16 depicted comparisons of on-grid and renewable energy consumptions at different renewable energy generation rates using two algorithms: cell size optimization (CSO) and strongest signal first (SSF) algorithms. The network serving 400 mobile users. SSF associates a user with the BS with strongest received signal strength. Figure 1.16a shows as the renewable energy rates increases, the on-grid energy consumption is reduced, and as the rate increases more, more renewable energy is generated and more BSs use renewable energy rather on-grid. When *renewable energy rate* ≥0.8 kwh/h, the CSO algorithm shows zero on-grid consumption due to CSO compared to SSF algorithm, which shows considerable on-grid energy consumption. Note that kwh is a unit of energy. The renewable energy rate is 0.8 kwh/h is 0.8 unit of energy per hour.

Figure 1.16b displays the renewable energy consumption at different renewable energy generation rates. When the renewable energy generation rate is <0.7 kwh/h, the renewable energy consumption, for both algorithms, increases as the generation rate increases. Additionally, CSO algorithm consumes more renewable energy than SSF algorithm because CSO optimizes the cell size to reduce the on-grid energy consumption. When the renewable energy generation rate is >0.7 kwh/h, the renewable energy consumption is almost flat while the SSF keeps increasing because the on-grid energy consumption of the CSO algorithm is zero.

1.10 Green Heterogeneous Networks

Heterogeneous networks (HetNets) integrating macro cells with small cells (low power nodes) can be used to reduce network energy consumption and the reason is that increasing number of users' equipment is likely to be served by a small cell in the neighbourhood and small energy radiation is needed as well as small-cell BS consumes less power than macro-cell BS. Nonetheless, the number of small-cell BSs are far more than the number of macro cell BSs.

Multi-tier HetNet composed of different tiers of devices such IoT, small cells, macro cell, etc., which makes the interference management and mitigation more complicated. A promising solution is based on cooperative transmission, which improves intercell interference but will not achieve energy efficiency. What is needed to perform cooperative transmission that enhances signal strength while it minimizes the energy consumption. Such technologies comprised transmit power control; BSs cooperation; BS sleep mode, etc. Since user-generated traffic varies with time and a great part of the energy consumption associated with BSs operations, it seems sensible to switch off underutilized BSs to minimize network consumption. BS sleep mode with dynamic clustering of CoMP provides a complete solution for performance improvements. In addition, hybrid energy resources (on-grid and renewable energy harvesting by BSs) can be used to enhance the green HetNets.

Green multicell cooperation (GMC) in HetNets with hybrid energy sources was investigated in [20]. The goal is to reduce the grid energy consumption in HetNets as much as possible. Accordingly, we need an efficient solution to deal with the spatial and temporal characteristics of the GMC problem simultaneously. A greedy decomposition is posed to solve the GMC problem by simplifying it by means of two independent subproblems: MCC and green energy planning subproblems. Further, the authors derived an upper bound for the greedy grid energy consumption achieved by the integrated solution and evaluated the time complexity of the proposed scheme.

1.11 Summary

In the introductory chapter, Section 1.1 we made a case for the 5G and motivation for writing the book and described the theme of the book. Most importantly, considering that 4G LTE development in coverage and speed is continuing, why do we need 5G now?

It is worth remembering the essential role of LTE technologies on enabling 5G services. 4G will not have the capabilities to support IoT and many other projected services for 2020. The key distinction for 5G services is not the expected high data delivery but latency. Current LTE latency is around 45 ms, compared to less than 1 ms in future 5G services.

5G is not an incremental advance of 4G. 5G will have a paradigm shift with massive BW, high BS and terminal devices densities, and massive MIMO systems. 5G will be formed as heterogeneous network (HetNet) comprising two layers of cells, macro cells, and small cells. A single macro cell comprises many small cells. In such a system, a dual-connectivity is normally employed: simultaneous connection of terminal to macro cell and small cell.

Contemporary channel and propagation models are not adequate for 5G small cells. In addition, 5G channel models have to support lD and 2D massive MIMO with dual polarization and should be appropriate for beamforming design.

Section 1.2 presents an overview of the contemporary cellular networks. Since the 1980s, we have witnessed four generations of mobile communications. 2G, also known as GSM, is a TDMA/FDMA system. During its long service duration, GSM was developed to GPRS to use packet-switched data and EDGE to increase data rates.

The motivation of 3G was the aspiration for higher data rate delivery for multimedia and improved robustness. Few factors help to advance W-CDMA technology in the 3G systems. A key attribute was how a wideband has the ability to resolve multipath components and with the help of rake receiver combines the signals of each component. Another key factor was the ability to spread data with orthogonal variable spreading factors (OVSF).

Nokia traffic model published in 2012 predicted that an increase of 1000X relative to the traffic in 2010 will come into being around 2020. So, to appropriately respond to increasing traffic and capacity growth, the LTE radio access technology has continuously evolved. Most current 4G LTE networks are based on 3GPP Releases 8 and 9, offering peak DL speed of 100 Mb/s and peak UL 50 Mb/s. 3GPP LTE-advanced introduced enhancement on the existing 4G LTE networks and raise the peak DL transmission speed to 1000 Mb/s and the UL peak speed to 500 Mb/s as recommended by 3GPP Release 10.

Section 1.3 summarized a few of the key developments in 3GPP LTE through Releases 8–15. The aims of the 3GPP were to develop a vision for an all IP wireless network covering radio access/core networks; service capabilities and interworking with Wi-Fi networks. 3GPP introduced the EPC to Release 8 specifications as IP-based core network for 4G LTE and also grants access to legacy networks such as 2G, 3G, and non-3GPP and fixed-access networks. So, circuit-switching data mode was operative no longer. EPC is adequate for fast processing for high data rates with low latency. The *v*EPC moves the core network components that traditionally run on a dedicated hardware to software run on low-cost servers standardized by ETSI. Release 9 defined three positioning schemes to accurately determine user location in emergency cases and further improvements were added to SON. In addition, Release 9 introduced CMAS added to ETWS introduced in Release 8. Release 10 introduced LTE-Advanced and defined the HetNet, where the coverage is achieved by a combination of large (macro) cells and small cells and improved the MIMO system to up to 8×8 in DL and 4×4 in UL. Releases 11 and 12 added in further improvement to LTE-Advanced. Release 13 introduce three important technologies: full-dimension MIMO; CA of up to 32 CCs compared to only 5 in release 10; and new enhancement to MTC including narrow BW low complexity user devices with long battery life for NB-IoT. Release 14 is known as the initial 5G standardization. Releases 15 and 16 are identified as 5G phase 1 and 2, respectively.

The capacity regions of multiuser cellular network are introduced in Section 1.4 by defining three basic channels in the multiuser network, namely: MAC, BC, and IC, followed by the analysis of ergodic sum-rate capacity of the degraded BC with superposition coding. The aim is presenting an overview of basic technologies in communication systems to the readers. We then consider the multiuser channels in an MIMO network to analyse the MIMO MAC capacity region and MIMO BC capacity region for $N = 1$ and $N > 1$ and we compared their regions with the achievable rate region of the optimum coding scheme, i.e. DPC.

In Section 1.5, we present a brief overview of the different types of fading channels, which are most commonly used in practical mobile communications, and in Section 1.6 we elucidated a possible extension of the single cell analysis carried out in Section 1.4 to multicell, multiuser analysis. In Section 1.7, we investigated in some detail an important issue associated with mobile communications, that is the impact of CO_2 emissions by active mobile networks. There are two key global initiatives: the EU energy consumption project and China green plan action. We consider in detail the EU project EARTH, which derived a mathematical model for computing the energy consumption and then briefly outlined China's mobile green plan action achievements. The main element in the network that consumed most of the energy is the BS. However, contemporary networks comprise various BS types in addition to different types of traffic models. The large-cell BS is the main consumer of energy in the system. In particular, key elements that may consume noticeable amounts of energy are PAs, transceivers (equal to the number of antennas), DC-DC power supply, active cooling system, AC-DC units, and the power loss in antenna feed cables. The network energy consumption was analysed in depth.

In Section 1.8, we derived a BS power consumption model that includes contribution from PA, small signal RF transceivers, BB units, power supply, and system active cooling. We then considered the power consumption at variable traffic load. We found that the consumption grows with load, but the growth is dependent of the BS type.

In Section 1.9, we considered the carbon emissions due to cellular networks. Using MCC, cellular networks can improve their energy efficiencies from three aspects: switching off some BSs when traffic demands have fallen below a certain threshold; using renewable energy to drive some BSs; and adopting MCC to expand coverage and hence reduce the number of active BSs.

In Section 1.10, we examined the possibilities of green heterogeneous network. A promising solution comprising BS sleep mode with dynamic clustering of CoMP is likely to provide a complete solution for performance improvements. In addition, hybrid energy resources (on-grid and renewable energy harvested by BSs) can be used to enhance the green HetNets.

1.A Tutorials on Theory and Techniques of Optimization Mathematics: Basics

Optimization of a mathematical function is a process of choosing the appropriate input to maximize/minimize the value of the function. In real wireless networks, the optimizations of resources are frequently required to satisfy certain conditions. In general, there are two types of optimizations: *unconstrained optimization* and *constrained optimization*. Before we proceed further into the mathematics of the optimization, let us define certain vocabularies that we need to use in what follows:

- Critical point: Any point in the function domain such that the function is not differentiable or its derivative is zero.
- Local minimum: At local minimum $x^*, f(x^*) < f(x)$ when $|x - x^*| < \in$ and $\in > 0$.
- Local maximum: At local maximum $x^*, f(x^*) > f(x)$ when $|x - x^*| < \in$ and $\in > 0$.
- Global minimum: At global minimum $x^*, f(x^*) \leq f(x)$ for all x.
- Global maximum: At global maximum $x^*, f(x^*) \geq f(x)$ for all x.
- Stationary point: The value of the function $f(x^*)$ calculated at the critical point x^*.
- Hessian matrix: A square matrix of second-order partial derivatives of a function of multiple variables (assuming the derivatives exist). Hessian matrix describes the curvature of the function.

The illustration in Figure 1.A.1 shows $f(x)$ on the y-axis and x on the x-axis and global max and min and the local max and min.

1.A.1 Optimization of Unconstrained Function with a Single Variable

Let get started with unconstrained optimization by considering a function $f(x)$ of a single independent variable such that

$$f(x) = \sin x \tag{1.A.1}$$

A plot of $f(x)$ vs x is shown in Figure 1.A.2. At the critical point x^*, we have

$$\frac{df}{dx} = 0 \tag{1.A.2}$$

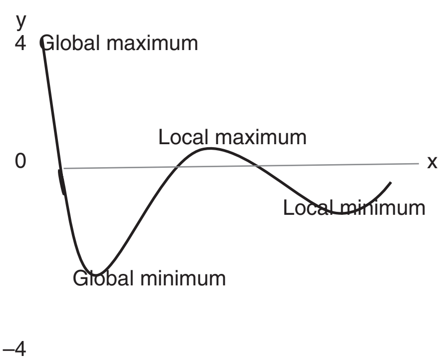

Figure 1.A.1 Illustration showing the global and local max and min.

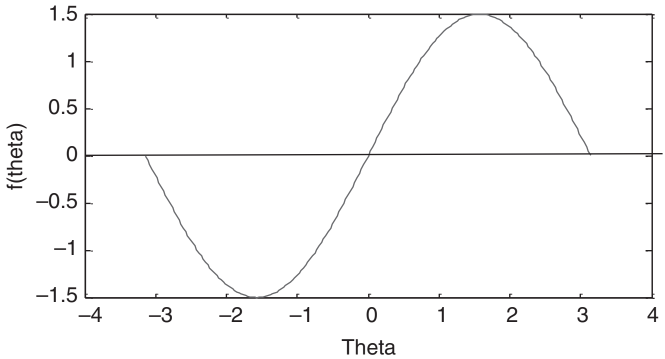

Figure 1.A.2 Plot of sine (theta) verses theta.

The gradient of $f(\mathrm{x})$ is changing from positive to negative before and after the maximum point. In other words, the gradient of $f(\mathrm{x})$ is decreasing. Consequently, the second-order derivative of the function at the critical point is negative. Similarly, it can be shown that at the minimum, the second-order derivative of the function $f(\mathrm{x})$ at the critical point is positive.

Worked Example 1 Investigate the maximum/minimum of the following single-variable function:

$$f(\mathrm{x}) = 5\mathrm{x}^3 + 6\mathrm{x}^2 - 3\mathrm{x} + 7$$

The first-order derivative of the function is

$$\frac{df}{d\mathrm{x}} = 15\mathrm{x}^2 + 12\mathrm{x} - 3$$
$$= 3(5\mathrm{x}^2 + 4\mathrm{x} - 1) = 3(5\mathrm{x} - 1)(\mathrm{x} + 1)$$

Thus, at

$$\frac{df}{d\mathrm{x}} = 0 \quad \mathrm{x}_1^* = \frac{1}{5} \quad and \quad \mathrm{x}_2^* = -1$$

Now

$$\frac{d^2f}{d\mathrm{x}^2} = 30\,\mathrm{x} + 12$$

At $\mathrm{x}_1^* = \frac{1}{5}$, $\frac{d^2f}{dx^2} = 18 > 0$; hence $\mathrm{x}_1^* = \frac{1}{5}$ is a relative minimum point and at the minimum point, $(\mathrm{x}_1^*) = 6.68$.

At $\mathrm{x}_2^* = -1$, $\frac{d^2f}{dx^2} = -18 < 0$; hence $\mathrm{x}_2^* = -1$ is a relative maximum point and at the maximum point, $f(\mathrm{x}_2^*) = 11$.

1.A.2 Optimization of Unconstrained Function with Multiple Variables

Instead of a one-variable function, we now consider a function of multiple variables and we wish to find its relative extremes. Let $\mathbf{x}$ be a vector of $\mathrm{n} \times 1$ elements such that

$$\mathbf{x} = [\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3, \dots\dots\dots\dots\dots.., \mathrm{x}_\mathrm{n}]^T \tag{1.A.3}$$

We assume that $f(\mathbf{x})$ is differentiable and that the derivatives are continuous. Let the Gradient of $f(\mathbf{x})$ denoted $\nabla f(\mathbf{x})$ such that

$$\nabla \mathrm{f}(\mathbf{x}) = \begin{bmatrix} \frac{\partial \mathrm{f}(\mathbf{x})}{\partial \mathrm{x}_1} \\ \frac{\partial \mathrm{f}(\mathbf{x})}{\partial \mathrm{x}_2} \\ \frac{\partial \mathrm{f}(\mathbf{x})}{\partial \mathrm{x}_n} \end{bmatrix} \tag{1.A.4}$$

1.A.3 The Hessian Matrix

The Hessian matrix (or simply Hessian) is a square matrix of second-order partial derivatives of a function (assuming it exists) used in optimization to determine whether the optimal point $\mathbf{x}^*$ is a maximum or minimum or *saddle point*. The Hessian of $f(\mathbf{x})$ is $\mathrm{n} \times \mathrm{n}$ matrix denoted by $\mathbf{H}_{f(\mathbf{x})}$ or simply $\mathbf{H}_f$ where

$$\mathbf{H}_f = \left[\frac{\partial^2 f(\mathbf{x})}{\partial \mathrm{x}_i \mathrm{x}_j}\right] \qquad i, j = 1, 2, \ldots \ldots \ldots \mathrm{n} \tag{1.A.5}$$

Using (1.A.4) and (1.A.5), the $\mathrm{n} \times \mathrm{n}$ Hessian of $f(\mathbf{x})$ is $\mathbf{H}_f$

$$\mathbf{H}_f = \begin{pmatrix} \frac{\partial^2 f(\mathbf{x})}{\partial \mathrm{x}_1{}^2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial \mathrm{x}_1 \partial \mathrm{x}_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\mathbf{x})}{\partial \mathrm{x}_n \partial \mathrm{x}_1} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial \mathrm{x}_n{}^2} \end{pmatrix} \tag{1.A.6}$$

Matrix $\mathbf{H}_f$ is symmetric such that $\frac{\partial^2 f(\mathbf{x})}{\partial \mathrm{x}_i \partial \mathrm{x}_j} = \frac{\partial^2 f(\mathbf{x})}{\partial \mathrm{x}_j \partial \mathrm{x}_i}$. We can now summarize the conditions for a maximum or a minimum critical point for differentiable multiple variable function $f(\mathbf{x})$ as follows:

- For maximum or minimum of $f(\mathbf{x})$ then $\nabla f(\mathbf{x}^*) = \mathbf{0}.$
- For maximum $\mathbf{x}^*$: leading principle minor of $\mathbf{H}_f$, order k has sign $(-1)^k$.
- For minimum $\mathbf{x}^*$: leading principle minor of $\mathbf{H}_f$, order k is >0.
- For saddle $\mathbf{x}^*$ when neither of the above conditions hold.

Note: To create leading principal minors of $\mathbf{H}_f$ ($\mathbf{h}_1, \mathbf{h}_2, \ldots \ldots \ldots .., \mathbf{h}_n$), start with upper left (1,1) element of $\mathbf{H}_f$ for $\mathbf{h}_1$, gradually add more rows and columns to the right of selected ones, and compute the determinant of the new matrix for ($\mathbf{h}_2, \ldots \ldots \ldots .., \mathbf{h}_n$). Consider a general 3×3 matrix $\mathbf{A}$:

$$\mathbf{A} = \begin{bmatrix} \mathrm{a}_{11} & \mathrm{a}_{12} & \mathrm{a}_{13} \\ \mathrm{a}_{21} & \mathrm{a}_{22} & \mathrm{a}_{23} \\ \mathrm{a}_{31} & \mathrm{a}_{32} & \mathrm{a}_{33} \end{bmatrix}$$

1st order principal minors: $|\mathrm{a}_{11}|$ formed by deleting the last two rows and cols. $|\mathrm{a}_{22}|$ formed by deleting the first and third rows and cols. $|\mathrm{a}_{33}|$ formed by deleting the first two rows and cols.

2nd order principal minors: $\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}$ formed by deleting col 3 and row 3; $\begin{vmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{vmatrix}$ formed by deleting col 2 and row 2; $\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix}$ formed by deleting col 1 and row 1; 3rd order principal minors $|\mathbf{A}|$ formed by determinant of matrix.

Worked Example 2 Consider the optimization problem of function $f(\mathbf{x})$ and $\mathbf{x}$ is 2×1 vector:

$$f(\mathbf{x}) = 2x_2^3 - 3x_1^3 + 7\,x_1x_2$$

$$\frac{\partial f(\mathbf{x})}{\partial x_1} = -9x_1^2 + 7x_2$$

$$\frac{\partial f(\mathbf{x})}{\partial x_2} = 6\,x_2^2 + 7\,x_1$$

At $\nabla f(\mathbf{x}) = 0$, the critical points $(x_{1,}^* \, x_2^*)$ are:

$$(0,0), (-0.8903, 1.0191)$$

We now investigate whether the critical points are maximum or minimum, so we have to compute the Hessian matrix $\mathbf{H}_f$ of $f(\mathbf{x})$ as follows:

$$\nabla^2 f(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f(\mathbf{x})}{\partial x_1^2} & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_2^2} \end{bmatrix} = \begin{bmatrix} -18x_1 & 7 \\ 7 & 12x_2 \end{bmatrix}$$

We compute the leading principal minors (two of these minors): First, the leading principal minor $= -18x_1$ and the second leading principle minor $= -18 * 12\,x_1x_2 - 49$.

At critical point (0,0), leading principal minors = 0 and −49. At critical point (0.8903, 1.0191), leading principal minors = 16.0254, 146.9778. Hence critical point (0, 0) a saddle of $f(\mathbf{x})$ and critical point (−0.8903, 1.0191) is a local minimum of $f(\mathbf{x})$.

Since at the critical point (−0.8903, 1.0191), both leading principal minors are positive definite, in fact (−0.8903, 1.0191) is a global minimum.

1.B Theory of Optimization Mathematics

Although the optimization theory and techniques cover a large area of mathematical analysis, in its simple definition in wireless networks, optimization proposes methods of finding the *best available* resource that is distributed fairly among users to make the network more efficient. Consequently, optimization is a process of choosing a level of resources for individual users, subject to certain constraints, for the total available resources. This involves finding maximum or minimum resource allocations that satisfies the given criteria.

Constraints have to be satisfied by the solution. There are two types of constraints namely: equality constraints and inequality constraints. In the following deliberations we present some aspects of the theory of optimization relevant to wireless networks, and then we proceed to use it to optimize wireless networks.

1.B.1 Constrained Optimization

Optimization problems may have constraints in diverse ways such as variables are integer or the signal may need to be positive. A generic constrained optimization may look like given a function of a variable **x** which has n values,

$$\min f(\mathbf{x}) \tag{1.B.1}$$

The optimization must satisfy the following conditions,

$$\text{subject to (s.t)} \quad \mathbf{h}(\mathbf{x}) = \mathbf{0} \tag{1.B.2a}$$

$$\mathbf{g}(\mathbf{x}) \leq \mathbf{0} \tag{1.B.2b}$$

where $\mathbf{h} = [\mathrm{h}_1 \ \mathrm{h}_2 \ ---- \ \mathrm{h}_\mathrm{m}]$ is m-element vector that is function of the n-element variable $\mathbf{x} = [\mathrm{x}_1 \ \mathrm{x}_2 ----, \mathrm{x}_\mathrm{n}]$, and $\mathbf{g} = [\mathrm{g}_1 \ \mathrm{g}_2 ----- \mathrm{g}_\mathrm{p}]$ is p-element vector that is function of the n-values variable **x**. The constraint $\mathbf{h}(\mathbf{x}) = \mathbf{0}$ is called equality constraint and $\mathbf{g}(\mathbf{x}) \leq \mathbf{0}$ is inequality constraint.

The optimization problem can be articulated as: find that values of the variable **x** that minimize $f(\mathbf{x})$. The equality constraints $\mathbf{h}(\mathbf{x}) = \mathbf{0}$ generates a set of m constraints, $\mathrm{h}_1(\mathbf{x}) = 0$; $\mathrm{h}_2(\mathbf{x}) = 0, \ldots\ldots \mathrm{h}_\mathrm{m}(\mathbf{x}) = 0$.

Let $\mathbf{x}^*$ be a regular point of these constraints and then there are m real values denoted by an m-element vector $\lambda \in \mathbb{R}^\mathrm{m}$ such that,

$$\nabla \mathrm{f}(\mathbf{x}^*) + \nabla \mathbf{h}(\mathbf{x}^*)\boldsymbol{\lambda} = 0 \tag{1.B.3}$$

where the symbol ∇ characterizes the gradient expressed as a vector differential operator and $\lambda = [\lambda_1 \ \lambda_2 \ ---- \lambda_\mathrm{m}]$ is called Lagrange multiplier provide an indication of how much the constraint cost when $\mathbf{x}^*$ is in the neighbourhood of the optimum solution.

1.B.2 Bordered Hessian Matrix H^B

A bordered Hessian matrix $\mathbf{H}^\mathrm{B}$ comprises the Hessian of a multivariate function $\mathbf{H}_f$ bordered by the gradient of a multivariate function **h** and is used in certain constrained optimization problems. When the function $f(\mathbf{x})$ is continuous so the border Hessian matrix is square and symmetric, the Border Hessian form looks like

$$\mathbf{H}^\mathrm{B} = \begin{bmatrix} 0 & \frac{\partial \mathbf{h}(\mathbf{x})}{\partial \mathrm{x}_1} & \frac{\partial \mathbf{h}(\mathbf{x})}{\partial \mathrm{x}_2} & \cdots\cdots\cdots & \frac{\partial \mathbf{h}(\mathbf{x})}{\partial \mathrm{x}_\mathrm{n}} \\ \frac{\partial \mathbf{h}(\mathbf{x})}{\partial \mathrm{x}_1} & \frac{\partial^2 f}{\partial^2 \mathrm{x}_1} & \frac{\partial^2 f}{\partial \mathrm{x}_1 \partial \mathrm{x}_2} & \cdots\cdots\cdots & \frac{\partial^2 f}{\partial \mathrm{x}_1 \partial x_\mathrm{n}} \\ \frac{\partial \mathbf{h}(\mathbf{x})}{\partial \mathrm{x}_2} & \frac{\partial^2 f}{\partial \mathrm{x}_2 \partial \mathrm{x}_1} & \cdots\cdots\cdots & & \frac{\partial^2 f}{\partial \mathrm{x}_2 \partial x_\mathrm{n}} \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ \frac{\partial \mathbf{h}(\mathbf{x})}{\partial \mathrm{x}_\mathrm{n}} & \frac{\partial^2 f}{\partial \mathrm{x}_\mathrm{n} \partial \mathrm{x}_1} & \frac{\partial^2 f}{\partial \mathrm{x}_\mathrm{n} \partial \mathrm{x}_2} & \cdots\cdots\cdots & \frac{\partial^2 \mathrm{f}}{\partial^2 \mathrm{x}_\mathrm{n}} \end{bmatrix} \tag{1.B.4}$$

Worked Example 3 Consider variable $\mathbf{x}$ as a vector of four elements: $\mathbf{x} = [x_1\ x_2\ x_3\ x_4]$ and function $f(\mathbf{x})$expressed as,

$$f(\mathbf{x}) = 2x_1^2 + 5{x_2}^2 + 4{x_3}^2 + 6{x_4}^2$$

We wish to find the optimal elements $\mathbf{x}^*$, λ^* that maximize $f(\mathbf{x})$ subject to the constraint

$$\mathbf{h}(\mathbf{x}) = x_1 + 3x_2 - 5x_3 - 8x_4 - 10$$

Solution

As there is just a single constraint, we use only one Lagrange multiplier. The Lagrange function $L(\mathbf{x}, \lambda)$ is

$$\mathbf{L}(\mathbf{x}, \lambda) = f(\mathbf{x}) - \mathbf{h}(\mathbf{x})\,\lambda$$

$$\mathbf{L}(\mathbf{x}, \lambda) = 2x_1^2 + 5{x_2}^2 + 4{x_3}^2 + 6{x_4}^2 + \lambda(-x_1 - 3x_2 + 5x_3 + 8x_4 + 10)$$

Using (1.B.4), we get

$$\frac{\partial L}{\partial x_1} = 4\,x_1 - \lambda = 0$$
$$\frac{\partial L}{\partial x_2} = 10\,x_2 - 3\,\lambda = 0$$
$$\frac{\partial L}{\partial x_3} = 8x_3 + \lambda(5) = 0$$
$$\frac{\partial L}{\partial x_4} = 12x_4 + \lambda(8)$$
$$\frac{\partial L}{\partial \lambda} = (x_1 + 3x_2 - 5x_3 - 8x_4 - 10) = 0$$

We now have to solve the five equations above to find the five unknowns. Solving these five equations, we get

$$\lambda^* = -1200/127$$
$$x_1^* = -300/127$$
$$x_2^* = -360/127$$
$$x_3^* = 750/127$$
$$x_4^* = -800/127$$
$$f(\mathbf{x}^*) = 425.3368$$
$$h(\mathbf{x}^*) = 0$$
$$h(\mathbf{x}) = x_1 + 3x_2 - 5x_3 - 8x_4 - 10$$
$$\frac{\partial h}{\partial x_1} = 1$$
$$\frac{\partial h}{\partial x_2} = 3$$
$$\frac{\partial h}{\partial x_3} = -5$$
$$\frac{\partial h}{\partial x_4} = -8$$

The bordered Hessian matrix for the example is

$$\mathbf{H}^{\mathrm{B}} = \begin{bmatrix} 0 & 1 & 3 & -5 & -8 \\ 1 & 4 & 0 & 0 & 0 \\ 3 & 0 & 10 & 0 & 0 \\ -5 & 0 & 0 & 8 & 0 \\ -8 & 0 & 0 & 0 & 12 \end{bmatrix}$$

Hence, the leading principle minors are:

$$\mathrm{H}_1^{\mathrm{B}} = -1$$
$$\mathrm{H}_2^{\mathrm{B}} = -46$$
$$\mathrm{H}_3^{\mathrm{B}} = -1368$$
$$\mathrm{H}_4^{\mathrm{B}} = -36896$$

Hence $\mathbf{x}^* = (0.2602, 0.3122, 0.6505, 0.6939)$ is local minimum.

1.C Karush–Kuhn–Tucker (KKT) Conditions

In mathematics, the KKT conditions (also known as Kuhn–Tucker) conditions are essential conditions for determining a solution to optimum problem provided certain constraints are satisfied. KKT scheme generalizes the Lagrange method, which is applicable to only equality constraints. Let $\mathbf{x}^*$be a local minimum and assume that has to satisfy both equality and inequality constraints, then there is a Lagrange multiplier vector $\lambda \in \mathbb{R}^m$ and a Lagrange multipliers vector $\boldsymbol{\mu} \in \mathbb{R}^p$ with

$$\boldsymbol{\mu} \geq \mathbf{0} \tag{1.C.1}$$

Consider the situation when there are inequality constraints. The optimization problem can be expressed as follows:

$$\begin{aligned} &\text{Optimize} && f(\mathbf{x}) \\ &\textit{subject to} && \mathbf{h}(\mathbf{x}) = \mathbf{0} \\ &\mathbf{g}(\mathbf{x}) \leq \mathbf{0} \end{aligned}$$

where

$$\mathbf{g}(\mathbf{x}) = [\mathrm{g}_1(\mathbf{x}) \quad \mathrm{g}_2(\mathbf{x}) \quad \mathrm{g}_3(\mathbf{x}) \quad \ldots\ldots\ldots\ldots\ldots .\mathrm{g}_\mathrm{p}(\mathbf{x})]$$

As in a previous consideration of equality constraints, we assume $f(\mathbf{x})$, $\mathbf{h}(\mathbf{x})$ and $\mathbf{g}(\mathbf{x})$ are differentiable and that their derivatives are continuous. Additionally, $\mathbf{g}(\mathbf{x})$ provides p inequality constraints. If the equality and inequality constraints problem is transformed to the corresponding unconstrained problem using KKT conditions, then there exists a solution $\mathbf{x}^*$ if and only if there is $\boldsymbol{\mu}$ such that:

$$\boldsymbol{\mu} \geq \mathbf{0} \tag{1.C.2}$$

$$\mathbf{g}(\mathbf{x}^*)^{\mathrm{T}} \boldsymbol{\mu} = \mathbf{0} \tag{1.C.3}$$

$$\nabla f(\mathbf{x}^*) + \nabla \mathbf{h}(\mathbf{x}^*)^T \boldsymbol{\lambda} + \nabla \mathbf{g}(\mathbf{x}^*)^T \boldsymbol{\mu} = \mathbf{0} \tag{1.C.4}$$

where $\mathbf{x}^*$ is a regular point, λ and $\boldsymbol{\mu}$ are Lagrange multipliers for equality and inequality constraints.

Worked Example 4 In this example we have one equality constraint and one inequality constraint.

Consider the function

$$f(\mathbf{x}) = x_1^2 + x_2^2 + x_3^2 + x_4^2$$

We are asked to find the minimum of $f(\mathbf{x})$ subject to the following two constraints:

$$\min f(\mathbf{x}) \tag{1.C.5}$$

subject to

$$\mathbf{h}(\mathbf{x}) = x_1 + x_2 + x_3 + x_4 - 1 \tag{1.C.5a}$$

$$x_4 \leq g(\mathbf{x}) \tag{1.C.5b}$$

Solution

The Lagrange function is

$$\mathbf{L}(\mathbf{x}, \lambda, \mu) = x_1^2 + x_2^2 + x_3^2 + x_4^2 - \lambda(x_1 + x_2 + x_3 + x_4 - 1) - \mu(x_4 - g(\mathbf{x}))$$

We now apply the KKT conditions:

$$\begin{aligned} &\frac{\partial \mathbf{L}}{\partial \mathbf{x}} = \mathbf{0} \\ &x_4 \leq g(\mathbf{x}) \\ &\mu \geq 0 \\ &\mu(x_4 - g(\mathbf{x})) \leq 0 \\ &\frac{\partial \mathbf{L}}{\partial \mathbf{x}} = \mathbf{0} \end{aligned} \tag{1.C.5c}$$

$$\frac{\partial \mathbf{L}}{\partial \mathbf{x}} = \begin{pmatrix} 2x_1 - \lambda \\ 2x_2 - \lambda \\ 2x_3 - \lambda \\ 2x_4 - \lambda + \mu \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

Therefore,

$$x_1 = x_2 = x_3 = \frac{\lambda}{2}; x_4 = \frac{\lambda - \mu}{2}$$

$$x_1 + x_2 + x_3 + x_4 = 4\left(\frac{\lambda}{2}\right) - \frac{\mu}{2} = 1$$

Thus,

$$\lambda = \frac{\mu + 2}{4}$$

and

$$x_1 = x_2 = x_3 = \frac{\mu + 2}{8}$$

$$x_4 = \frac{\lambda - \mu}{2} = \frac{1}{4} - \frac{3\mu}{8} \tag{1.C.5d}$$

Accordingly,

$$\frac{1}{4} - \frac{3\mu}{8} \leq g(\mathbf{x}),$$

that is,

$$\frac{3\mu}{8} \geq \frac{1}{4} - g(\mathbf{x}) \qquad (1.C.5e)$$

There are three possible cases for (1.C.5c)

Case 1 $g(\mathbf{x}) > \frac{1}{4}$ and since $\mu \geq 0$ so (1.C.5d) is satisfied.

Case 2 $g(\mathbf{x}) = \frac{1}{4}$ and (1.C.5c) is also satisfied.

Case 3 $g(\mathbf{x}) < \frac{1}{4}$ then (1.C.5e) is satisfied, but (1.C.5c) require $\mu = 0$ and (1.C.5d) gives $x_4 = \frac{1}{4}$ which violates $x_4 \leq g(\mathbf{x})$. Therefore $g(\mathbf{x}) = \frac{1}{4}$.

References

1 Frattasi, S., Fathi, F., Fitzek, F.H.P. et al. (2006). Defining 4G technology from the users prospective. *IEEE Network* 20 (1): 35–41.

2 Deployment Strategies for Heterogeneous Networks – White Paper- Nokia Solutions and Networks, 2014.

3 Astely, D., Dahlman, E., Fodor, G. et al. (2013). LTE release 12 and beyond. *IEEE Communications Magazine* 51 (7): 154–160.

4 Tse, D. and Viswanath, P. (2012). *Fundamentals Wireless Communication*. UK: Cambridge University Press.

5 Vanka, S., Srinivasa, S., Gong, Z. et al. (2012). Superposition coding strategies: design and experimental evaluation. *IEEE Transactions on Wireless Communications* 11 (7): 2628–2638.

6 Caire, G. and Shamai, S. (2003). On the achievable throughput of a multiantenna Gaussian broadcast channel. *IEEE Transactions on Information Theory* 49 (7): 1691–1706.

7 Goldsmith, A., Jafar, S.A., Jindal, N., and Vishwanath, S. (2003). Capacity limits of MIMO channels. *IEEE Journal on Selected Areas in Communications* 21 (5): 684–702.

8 Spencer, Q.H., Peel, C.B., Swindlehurs, A.L., and Haardt, M. (2004). An introduction to the multi-user MIMO downlink. *IEEE Communications Magazine* 42 (10): 60–67.

9 Hassibi, B. and Sharif, M. (2007). Fundamental limits in MIMO broadcast channels. *IEEE Journal on Selected Areas in Communications* 25 (7): 1333–1344.

10 Lin, S.-C. and Su, H.-J. (2007). Practical vector dirty paper coding for MIMO Gaussian broadcast channels. *IEEE Journal on Selected Areas in Communications* 25 (7): 1345–1357.

11 Costa, M.H.M. (1983). Writing on dirty paper. *IEEE Transactions on Information Theory* 29 (3): 439–441.

12 Sato, H. (1978). An outer bound on the capacity region of the broadcast channel. *IEEE Transactions on Information Theory* 24 (3): 374–377.

13 Vishwanath, S., Jindal, N., and Goldsmith, A. (2002). On the capacity of multiple input multiple output broadcast channels. *IEEE International conference on communications (ICC)* 3: 1444–1450.

14 Biglieri, E., Proakis, J., and Shamai, S. (1998). Fading channels: information: theoretic and communications aspects. *IEEE Transactions on Information Theory* 44 (6): 2619–2692.

15 Li, L. and Goldsmith, A. (2001). Capacity and optimal resource allocation for fading broadcast channels. I. Ergodic capacity. *IEEE Transactions on Information Theory* 47 (3): 1083–1102.

16 Auer, G., Giannini, V., Desset, C. et al. (2011). How much energy is needed to run a wireless network? *IEEE Wireless Communications* 18 (5): 40–49.

17 Mancuso, V. and Alouf, S. (2011). Reducing costs and pollution in cellular networks. *IEEE Communications Magazine* 49 (8): 63–71.

18 Fehske, A., Fettweis, G., Malmodin, J., and Biczok, G. (2011). The global footprint of mobile communications: the ecological and economic perspective. *IEEE Communications Magazine* 49 (8): 55–62.

19 Han, T. and Ansari, N. (2013). On greening cellular networks via multicell cooperation. *IEEE Wireless Communications* 20 (1): 82–89.

20 Chiang, Y.-H. and Liao, W. (2016). Green multicell cooperation in heterogeneous networks with hybrid energy sources. *IEEE Transactions on Wireless Communications* 15 (12): 7911–7925.