Foundations of User-Centric Cell-Free Massive MIMO

Özlem Tuğfe Demir
KTH Royal Institute of Technology
and Linköping University
ozlemtd@kth.se

Emil Björnson
KTH Royal Institute of Technology
and Linköping University
emilbjo@kth.se

Luca Sanguinetti
University of Pisa
luca.sanguinetti@unipi.it

Version of record: Özlem Tuğfe Demir, Emil Björnson and Luca Sanguinetti (2021), "Foundations of User-Centric Cell-Free Massive MIMO", Foundations and Trends® in Signal Processing: Vol. 14, No. 3-4, pp 162–472. DOI: 10.1561/2000000109.

Simulation code and supplementary material:
https://github.com/emilbjornson/cell-free-book
Printed books: Available from now Publishers Inc., http://www.nowpublishers.com

This is the authors' version of the manuscript. See the above version of record for the final published manuscript. Date of this version: February 22, 2022.

Contents

Foundations of User-Centric Cell-Free Massive MIMO

Özlem Tuğfe Demir[1], Emil Björnson[2] and Luca Sanguinetti[3]

[1] *KTH Royal Institute of Technology and Linköping University; ozlemtd@kth.se*
[2] *KTH Royal Institute of Technology and Linköping University; emilbjo@kth.se*
[3] *University of Pisa; luca.sanguinetti@unipi.it*

ABSTRACT

Imagine a coverage area where each mobile device is communicating with a preferred set of wireless access points (among many) that are selected based on its needs and cooperate to jointly serve it, instead of creating autonomous cells. This effectively leads to a user-centric post-cellular network architecture, which can resolve many of the interference issues and service-quality variations that appear in cellular networks. This concept is called User-centric Cell-free Massive MIMO (multiple-input multiple-output) and has its roots in the intersection between three technology components: Massive MIMO, coordinated multipoint processing, and ultra-dense networks. The main challenge is to achieve the benefits of cell-free operation in a practically feasible way, with computational complexity and fronthaul requirements that are scalable to enable massively large networks with many mobile devices. This monograph covers the foundations of User-centric Cell-free Massive MIMO, starting from the motivation and mathematical definition. It continues by describing the state-of-the-art signal processing algorithms for channel estimation, uplink data reception,

Özlem Tuğfe Demir, Emil Björnson and Luca Sanguinetti (2021), "Foundations of User-Centric Cell-Free Massive MIMO", Foundations and Trends® in Signal Processing: Vol. 14, No. 3-4, pp 162–472. DOI: 10.1561/2000000109.

and downlink data transmission with either centralized or distributed implementation. The achievable spectral efficiency is mathematically derived and evaluated numerically using a running example that exposes the impact of various system parameters and algorithmic choices. The fundamental tradeoffs between communication performance, computational complexity, and fronthaul signaling requirements are thoroughly analyzed. Finally, the basic algorithms for pilot assignment, dynamic cooperation cluster formation, and power optimization are provided, while open problems related to these and other resource allocation problems are reviewed. All the numerical examples can be reproduced using the accompanying Matlab code.

1

Introduction and Motivation

The purpose of mobile networks is to provide devices with wireless access to a variety of data services anywhere in a wide geographical area. For many years, the main service of these networks was voice calls, but nowadays transmission of data packets is the dominant service [**EricssonMobility**]. Hence, the service quality of contemporary networks is mainly determined by the data rate (measured in bit per second) that can be delivered at different locations in the coverage area. The range of wireless transmission is determined by the propagation environment. Since the received signal power decays quadratically, or even faster, with the propagation distance, a traditional mobile network infrastructure consists of a set of geographically distributed transceivers that the connecting device can choose between. These are typically deployed at elevated locations (e.g., in masts and at rooftops) to provide unobstructed propagation to many places in the area. Each transceiver will be called an *access point (AP)* and each user device will be called a *user equipment (UE)* in this monograph.

Current mobile networks are built as *cellular networks*, which means that each UE connects to one AP, namely the one that provides the strongest signal. The UE locations for which a particular AP is selected

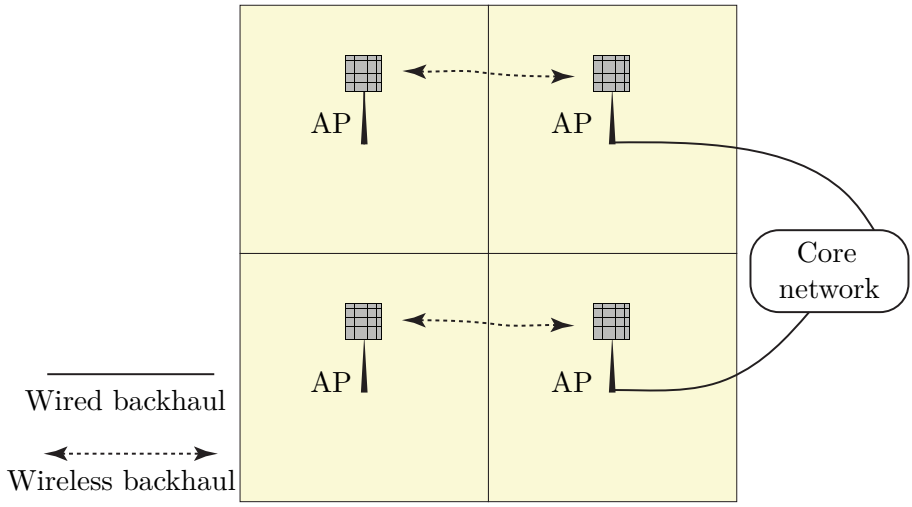


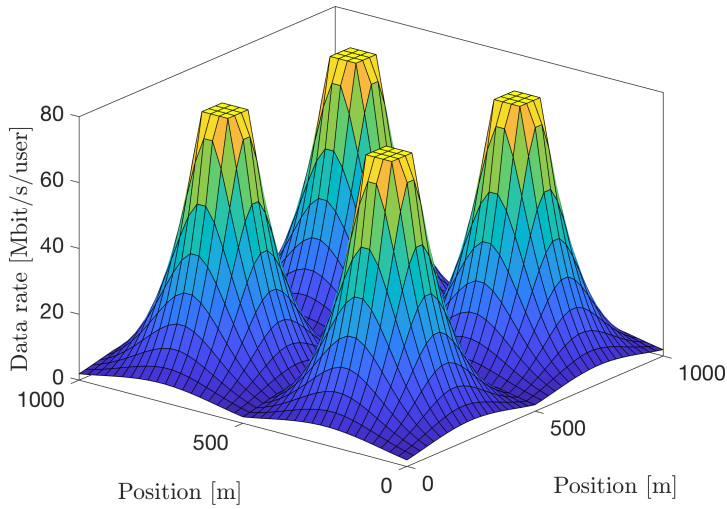
Figure 1.1: Cellular network with four APs, which are all connected to the core network via wired backhaul. Some APs are also interconnected by wireless backhaul.

is called a *cell*. Figure 1.1 shows the basic infrastructure of a cellular network with four APs, each equipped with a planar antenna array containing both the antenna elements and the associated radio units (also known as transceiver chains). The antenna elements emit and receive radio frequency (RF) waves, while the radios generate the analog RF signals to be emitted and process the received RF signals. The radios are connected to a baseband unit that processes the transmitted and received signals in the digital domain. This monograph is focused on the digital signal processing associated with the baseband, thus we will simply refer to each radio and its associated antenna element(s) as *an antenna*. The exact hardware implementation is thereby abstracted away. It is the number of such antennas that determines the dimensionality of the signals that will be generated and processed in the baseband.

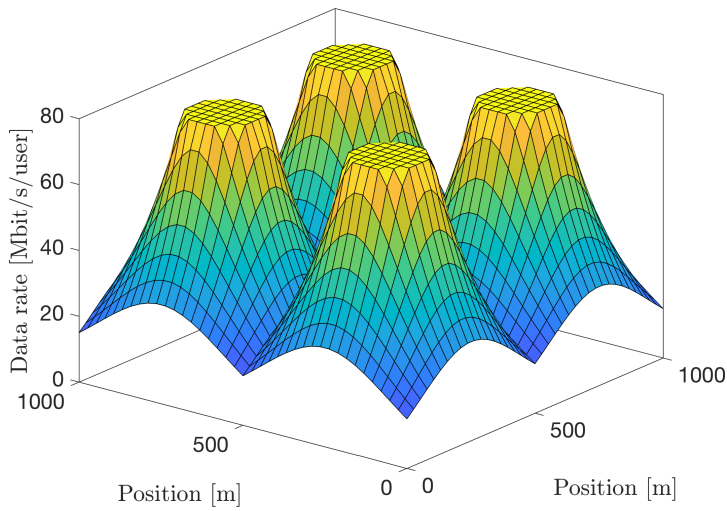
The square area around each AP illustrates the cell that the AP provides service to. In reality, the cells will not have symmetric shapes (such as squares, triangles, or hexagons), but it is commonly illustrated like that when describing the fundamentals. The infrastructure of contemporary cellular networks can be divided into two parts: an edge and a core. The edge consists of APs and other hardware units that

are directly involved in the physical-layer communication with the UEs. The core network facilitates all the services requested by the UEs, including routing of data packages and connection to the Internet. The connections between the edge and core are called *backhaul* links and can either be fully wired (e.g., using fiber cables) or partially wireless (e.g., using fixed microwave links). Figure 1.1 shows an example where the APs to the right are connected via wired backhaul links to the core network. The APs to the left are connected wirelessly to the APs to the right, thus their backhaul traffic flows over both wireless and wired links.

An important consequence of the fact that the received signal power rapidly decays with the propagation distance is that the UEs that happen to be close to an AP (i.e., in the cell center) will experience a higher signal-to-noise ratio (SNR) than those that are close to the edge between two cells. A 10 000 times (40 dB) difference is common between the cell center and cell edge. Moreover, UEs at the cell edge are also affected by interference from neighboring APs, thus the signal-to-interference-plus-noise ratio (SINR) can be substantially lower than the SNR at these locations. The data rate is an increasing function of the SINR, thus there are large rate variations in each cell. Figure 1.2(a) exemplifies this behavior by showing the data rate achieved in the downlink by a UE at different locations, when each AP uses a traditional fixed-gain antenna and transmits with maximum power. When the UE is close to one of the APs, it achieves the maximum rate that is supported by the system, which is 80 Mbit/s in this example. In contrast, UEs at the cell edges achieve rates below 1 Mbit/s. This is insufficient for many data services but is nevertheless enough for making voice calls. Depending on the codec, a voice call requires as little as 10-100 kbit/s and this is supported everywhere in this example. Cellular networks were initially designed with this property in mind; we needed the SNR to be above a threshold everywhere in the coverage area to prevent dropped calls, but there was no benefit from being far above that threshold. This basic property has changed entirely when we started using cellular technology for data transmission. Since the UEs request the same data services everywhere in the coverage area, cell-center UEs only need to be connected part of the time, while the cell-edge UEs must be turned



(a) Each AP has a 9 dBi fixed-gain antenna.



(b) Each AP is equipped with 64 omni-directional antennas.

Figure 1.2: Example of the downlink data rate achieved by a UE at different locations in the cellular network in Figure 1.1, assuming each AP transmits with full power. The cell-edge SNR is 0 dB in (a) and the power is assumed to decay as the distance to the power of four. The bandwidth is 10 MHz, and the maximum spectral efficiency (SE) is 8 bit/s/Hz. The key observation is that the rates vary substantially in the network.

on for a much larger fraction of time (if the requested service can even be provisioned). Hence, at a given time instance, the majority of active UEs are at the cell edges and their performance will determine how the customers perceive the service quality of the network as a whole.

The large data rate variations are inherent to the cellular network architecture and remain even if the APs are equipped with advanced hardware, such as *Massive multiple-input multiple-output (MIMO)* [**massivemimobook**], [**Marzetta2010a**], [**Marzetta2016a**]. The MIMO technology enables each AP to use an array of antennas (with integrated radios) to serve multiple UEs in its cell by directional transmission, which also increases the SNR and reduces inter-cell interference. More precisely, in the uplink, multiple UEs transmit data to the APs in the same time-frequency resource. The APs exploit the massive number of channel observations (made on the receive antennas) to apply linear receive combining, which discriminates the desired signal from the interfering signals using the spatial domain. In the downlink, the UEs are coherently served by all the antennas, in the same time-frequency resource, but separated in the spatial domain by receiving very directive signals. Figure 1.2(b) shows the downlink data rate achieved by a UE at different locations when each AP has an array of 64 antennas. The data rates are generally higher than in Figure 1.2(a). The cell-center area where the maximum data rate is delivered grows and large improvements are also seen at the cell-edge UEs, since beamforming from the antenna array at the AP can increase the SNR without increasing the inter-cell interference. Despite these gains, there are still substantial rate variations in each cell. Each AP could, in principle, optimize its transmit power to even out the differences (e.g., by reducing the power when serving UEs in the cell center) but this is undesirable since it results in serving all the UEs using the relatively low rates that can be delivered at the cell edge.

Current cellular networks can achieve high peak data rates in the cell centers, but the large variations within each cell make the service quality unreliable. Even if the rates are sufficiently high at, say, 80% of the locations in a cell, this is not sufficient when we are creating a society where wireless access is supposed to be ubiquitous. When payments, navigation, entertainment, and control of autonomous vehicles are all

relying on wireless connectivity, we must raise the uniformity of the data service quality. In summary, the primary goal for future mobile networks should not be to increase the peak rates, but the rates that can be guaranteed to the very vast majority of the locations in the geographical coverage area. The cellular network architecture was not designed for high-rate data services but for low-rate voice services, thus it is time to look beyond the cellular paradigm and make a clean-slate network design that can reach the performance requirements of the future. This monograph considers the cell-free network architecture that is designed to reach the aforementioned goal of uniformly high data rates everywhere.

The cell-free concept for wireless communication networks is defined in Section 1.1, which briefly describes how to operate such networks. Section 1.2 puts the new technology into a historical perspective. Section 1.3 describes three basic benefits that cell-free networks have compared to cellular networks. The key points are summarized in Section 1.4.

1.1 Cell-Free Networks

We will now describe the basic architecture and terminology of a cell-free network. The system and channel propagation models, including the mathematical notation, will be introduced in Section 2 on p. 203.

A cell-free network consists of L geographically distributed APs that are jointly serving the UEs that reside in the area. Each AP is connected via a *fronthaul* to a central processing unit (CPU), which is responsible for the AP cooperation. There can be multiple CPUs all connected via fronthaul links, which can be wired or wireless. An illustration of a cell-free network with single-antenna APs is provided in Figure 1.3. A cell-free network can be divided into an edge and a core, just as cellular networks. The APs and CPUs are at the edge and the connections between them are called fronthaul links, while the connections between the edge and core are still called backhaul links. Hence, the CPUs are connected to the core network via backhaul links, which are used to send/receive data from the Internet and other sources, to facilitate various data services. In contrast, the fronthaul links can be used for: 1) sharing physical-layer signals that will be transmitted in the

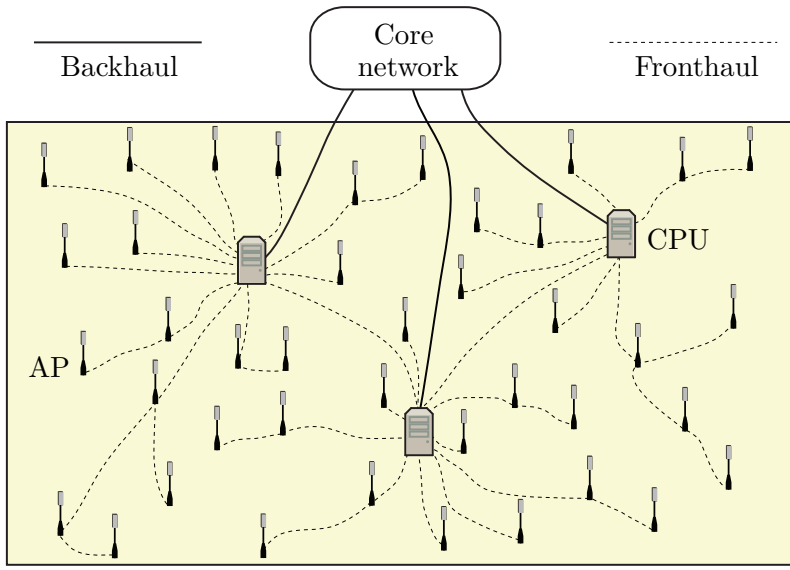


Figure 1.3: Illustration of a cell-free network with many geographically distributed APs connected to CPUs via fronthaul links. The CPUs are connected to the core network via backhaul links. The APs are jointly serving all the UEs in the coverage area.

downlink; 2) forwarding received uplink data signals that are yet to be decoded; and 3) sharing channel state information (CSI) related to the physical channels. The fronthaul also facilitates phase-synchronization between geographically distributed APs, for example, by providing a common phase reference.

A particular fronthaul topology is illustrated in Figure 1.3, where some APs are directly connected to a CPU while other APs are connected via a neighboring AP. We stress that this is only for illustration purposes. No specific assumption on the topology will be made in this monograph, except that the fronthaul links exist, have infinite capacity, negligible latency, and introduce no errors. This allows us to quantify the ultimate physical-layer performance of the cell-free network architecture. Practical constraints on the fronthaul infrastructure are briefly reviewed in Section 7.6 on p. 433. We also note that a CPU may not be a separate physical unit but may be viewed as a logical entity;

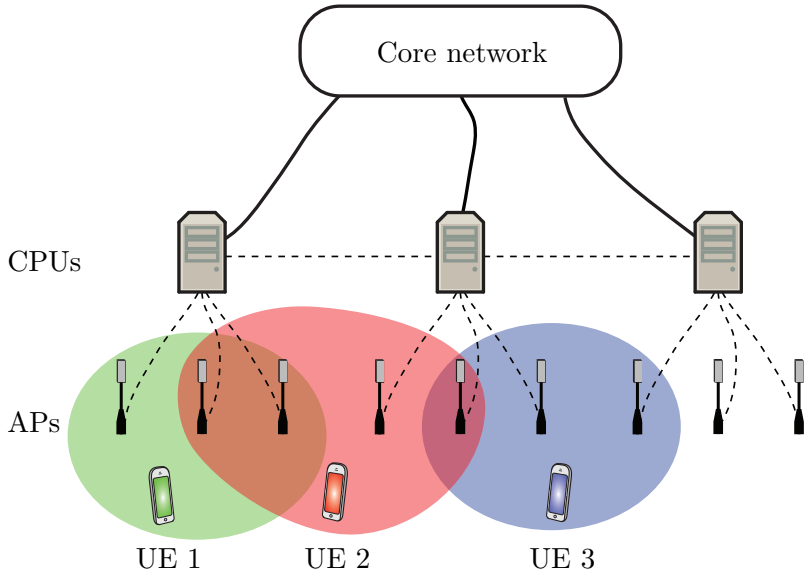


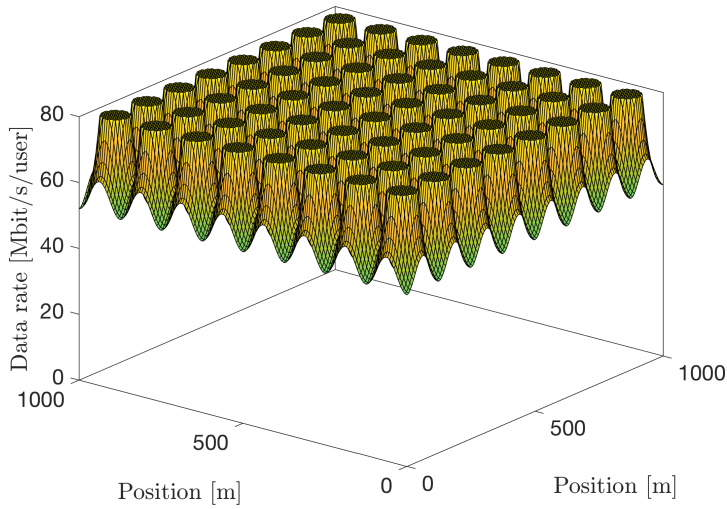
Figure 1.4: Illustrations of the different layers in a cell-free network. Each UE connects to a subset of the APs, which is illustrated by the shaded regions. Each AP is connected to one CPU via fronthaul. The CPUs are interconnected either directly or via the core network.

for example, the CPUs may represent a set of local processors that can be either located at a subset of the APs or at other physical locations, and which are connected via fronthaul links. Aligned with the ongoing cloudification of wireless networks [**6882182**], [**Peng2016a**], known as *cloud radio access network (C-RAN)*, the CPU-related processing tasks can be distributed between the local processors in different ways [**Bjornson2013d**].

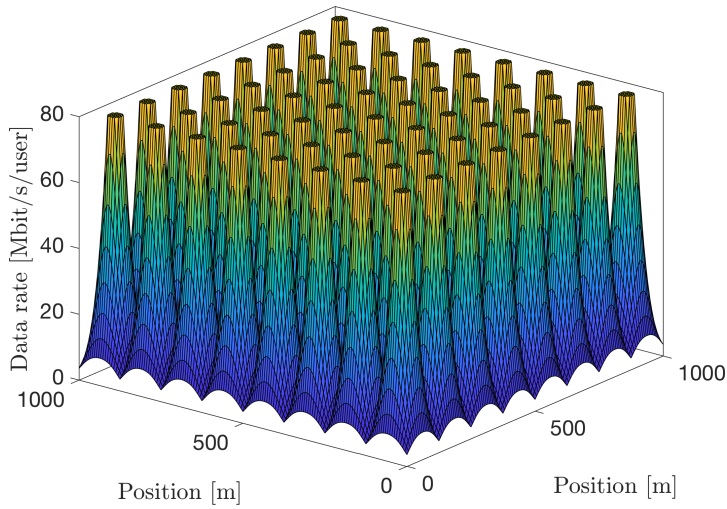
Generally speaking, C-RAN is a network deployment architecture where a group of APs is connected to the same CPU, which carry out most of the APs' baseband processing. By sharing computational resources, the total computational capacity can be reduced since it is unlikely that all APs need the maximum capacity simultaneously. One can also make use of general-purpose hardware and open protocols. Recently, the C-RAN abbreviation has started to stand for *centralized RAN*, since the word "cloud" gives the impression that the CPU is

owned by another vendor than the wireless network and can be located anywhere in the world. However, to meet the latency constraints of baseband processing, the CPU is rather an edge-cloud processor located in the same geographical area as the APs. Many different physical-layer technologies can be implemented using the C-RAN architecture. So far, it has mainly been used for cellular networks but it is also the foundation for cell-free networks. Figure 1.4 gives a schematic view of a cell-free network that uses the C-RAN architecture. It is divided into different layers: the core network, the CPU layer, the AP layer, and the UE layer. Each UE is served by a subset of the APs, for example, all the neighboring ones. These subsets are illustrated by the shaded regions in Figure 1.4. For each UE, one of the selected APs is the so-called *Master AP* that is responsible for serving the UE and appointing a CPU where the uplink data decoding and downlink data encoding will be carried out. That CPU delivers the downlink data to all APs that are transmitting to the UE and combines/fuses the uplink received signals obtained at those APs in a final decoding step. A UE can be served by APs connected to different CPUs; there exists a fronthaul link between every pair of APs even if it might go via other entities. The signal processing required for communication can be divided between the APs and CPU in different ways, which will be explored in later sections of this monograph. As the UE moves around, the Master AP assignment, selection of CPU, and selection of cooperating APs may change dynamically.

The word "cell-free" signifies that no cell boundaries exist from a UE perspective during uplink and downlink transmission since all APs that affect a UE will take an active part in the communication. For example, when a UE transmits an uplink data signal then all APs that receive it, with an SNR that is above a threshold, will collaborate in decoding the signal. The partially overlapping shaded regions in Figure 1.4 can be created in that way. The network is jointly serving all the K UEs that are active in the coverage area of the network, even if not all APs might serve every single UE. The differences between cellular and cell-free networks exist at the infrastructure and signal processing side, but can be transparent to the UEs. It should be possible for the same UE to connect to both types of networks without upgrading its software.



(a) Cell-free network.



(b) Cellular network with the same AP locations.

Figure 1.5: Example of the downlink data rate achieved by a UE at different locations in a network with 64 APs with omni-directional antennas deployed on a square grid and jointly transmitting to the UE. The propagation parameters are otherwise the same as in Figure 1.2. A cell-free network operation is considered in (a), while a cellular network operation is considered in (b). The key observation is that only the cell-free operation can provide almost uniformly high data rates in the entire network.

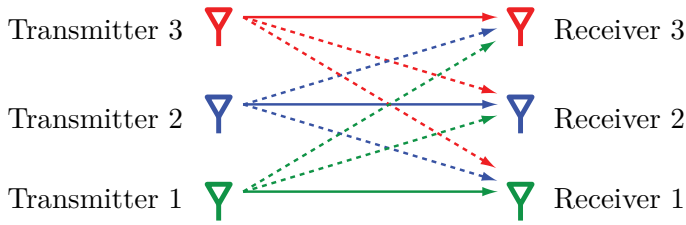
To give a first impression of the goal of creating cell-free networks, Figure 1.5(a) shows the downlink data rate achieved by a UE at different locations in a setup that resembles the cellular example in Figure 1.2. For simplicity, an ideal deployment with 64 APs deployed on an 8×8 square grid is considered. The figure shows that the rates vary between 52 and 80 Mbit/s everywhere in the coverage area. One contributing factor is the denser deployment, which greatly reduces the average propagation distance between a UE and the closest AP. However, the main reason is that all the surrounding APs are jointly transmitting to the UE, thereby alleviating the inter-cell interference issue that is one of the main causes of the large rate variations in cellular networks. This is evident when comparing Figure 1.5(a) with Figure 1.5(b), where a cellular network with the same AP locations is considered. The inter-cell interference then gives rise to large rate variations.

1.2 Historical Background

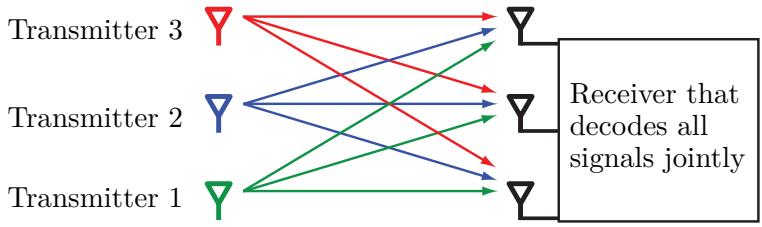
The cellular architecture has played a key role in enabling mobile communications, from the early concepts developed in the 1950s and 1960s [**Bullington1953a**], [**Frefkiel1970a**], [**Schulte1960a**] to the first commercial deployment in 1979 [**Kinoshita2018a**]. The motivating factor of building a cellular network was to make efficient use of the limited frequency spectrum by enabling many concurrent transmissions in the geographical area covered by the network. To control the interference between the transmissions, the coverage area was divided into predefined geographical zones, known as cells, where a fixed AP takes care of the service. In the beginning, a predefined frequency plan was utilized so that adjacent cells use different frequency resources, thereby limiting the inter-cell interference. Over the years, commercial cellular networks have been densified by deploying more APs per area unit [**Cooper2010a**]. By using steerable multi-antenna panels at each AP, instead of fixed-beam antennas, the interference between adjacent cells can be partially controlled so that the traditional frequency plans can be alleviated. Depending on the deployment scenario (e.g., indoor/outdoor, frequency band, coverage area, and distance from the AP to the closest UE location), different types of AP hardware are

utilized [**Kamel2016a**]. The resulting parts of the cellular networks are sometimes categorized as microcells, picocells, and femtocells. We will use the overarching term *small cells* when referring to such networks [**Hoydis2011c**]. The use of smaller and smaller cells has been an efficient way to increase the network capacity, in terms of the number of bits per second that can be transferred in a given area. Ideally, the network capacity grows proportionally to the number of APs (with active UEs), but this trend gradually tapers off due to the increasing inter-cell interference [**Andrews2017a**], [**Zhou2003a**]. After a certain point, further network densification can actually reduce rather than increase the network capacity. This is particularly the case in the *ultra-dense network* regime [**Hwang2013a**], [**Kamel2016a**], [**Stefanatos2014a**], where the number of APs is larger than the number of *simultaneously active* UEs. Even if each AP would have a handful of antennas, this is not enough to suppress all the interference in such a dense scenario. A cell-free network is an attempt to move beyond those limits [**Chen2016a**], [**Interdonato2018**], [**Ngo2017b**], [**Zhang2020a**], [**Zhang2019a**]. Before explaining how that can be achieved, we will give a detailed historical background.

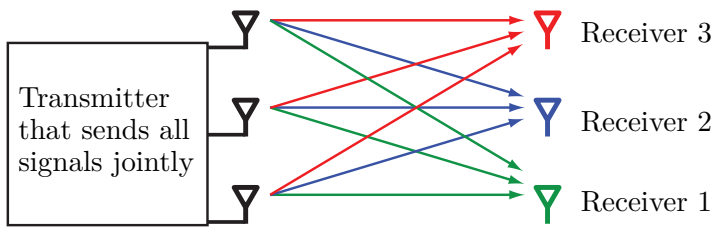
As mentioned earlier, a key property of conventional cellular networks is that each UE is assigned to one cell and only served by its AP. This is known as an *interference channel* in information theory and is illustrated in Figure 1.6(a) for the case of three single-antenna transmitters and three single-antenna receivers. Each receive antenna obtains a signal containing the information sent from one desired transmitter (solid line) plus two interfering signals (dashed lines) sent from the undesired transmitters. Even in the absence of noise, identifying the desired signal is like solving an ill-conditioned linear system of equations with three unknowns but only one equation. Hence, the inter-cell interference is unusable in this case; it only limits the performance. When operating such a cellular network, the transmit powers might be adjusted to determine which of the cells will be most affected by the interference. There is no other cooperation between the APs; neither CSI nor transmitted/received signals are shared between cells. These assumptions were challenged by Wyner in [**Wyner1994a**] from 1994, where the uplink was studied and the benefit of jointly decoding the



(a) An *interference channel* representing how cellular networks are conventionally operated. Each receiver wants to decode the data sent from its transmitter, subject to the dashed interfering signals from simultaneous transmissions.



(b) A *multiaccess channel* representing how distributed receive antennas can cooperate to jointly decode the signals from all transmitters. The information contained in all received signals can be utilized. There are no unusable interfering signals. This describes the ideal uplink operation of a cell-free network.



(c) A *broadcast channel* representing how distributed transmit antennas can cooperate to jointly send the signals to all receivers. The signals sent from all antennas can be utilized at each receiver. There are no unusable interfering signals. This describes the ideal downlink operation of a cell-free network.

Figure 1.6: A cellular network is conventionally operated as an interference channel, which is shown in (a). To alleviate inter-cell interference, the uplink of a cell-free network is instead operated as the multiaccess channel shown in (b) and the downlink is operated as the broadcast channel shown in (c).

data from all UEs using the received signals in all cells was explored. In this way, the interference channel is turned into a *multiaccess channel*, where all the receive antennas collaborate. Even if each antenna receives a superposition of multiple signals, there is no unusable interference but the task of the receiver is to extract the information contained in all the received signals. This alternative way of operating the system is illustrated in Figure 1.6(b). In this example, the receiver has access to three observations that contain linear combinations of the three desired signals. In the absence of noise, signal detection can be viewed as solving a linear system of equations with three unknowns and three equations, which is a well-conditioned problem. Importantly, the interference is not only canceled by this approach, but the observations made at multiple receive antennas are combined to increase the SNR compared to the case where there was no interference between the transmissions [**Gesbert2010a**]. *Interference is turned from being bad to being good!*

Similarly, Shamai and Zaidel proposed a downlink co-processing framework in [**Shamai2001a**] from 2001. Using information-theoretic terminology, the cellular downlink was transformed from an interference channel to a *broadcast channel*, where all the geographically distributed transmit antennas collaborate. This case is illustrated in Figure 1.6(c). Each antenna transmits a linear combination of the downlink signals intended for the UEs in all cells, where the linear combination is designed based on the channels to limit inter-cell interference. For example, in the setup shown in Figure 1.6(c) with three geographically distributed transmitters (APs) and three distributed receivers (UEs), zero-forcing (ZF) precoding can be utilized to completely avoid interference. This is not possible in the interference channel in Figure 1.6(a), where each signal is only sent from one transmitter and no precoding can be used.

While the premise of [**Shamai2001a**], [**Wyner1994a**] was to add co-processing to an existing cellular network, the idea of building a cell-free network from the outset was pioneered by Zhou, Zhao, Xu, Wang, and Yao in [**Zhou2003a**] from 2003. Their concept was called *Distributed Wireless Communication System* and resembles the architecture described in Section 1.1 with geographically distributed antennas and processing, and a CPU that controls the system. The paper proposes that a UE should not be served by all the antennas but only by

the nearest set of distributed antennas, as illustrated by the shaded regions in Figure 1.4. This is an early step towards a user-centric assignment of network infrastructure, where each UE is served by the user-preferred set of APs instead of by a predefined set. Similar ideas appeared for soft handoff in code-division multiple access (CDMA) systems [**Viterbi1994a**], where UEs at cell edges are jointly served by all the nearest APs.

Many other researchers contributed to this topic during the 2000s and a variety of terminologies have been used to refer to systems where the APs are jointly processing the transmitted and received signals. We will provide some key examples in this paragraph, without attempting to provide an exhaustive list. Non-linear co-processing schemes were developed by Jafar, Foschini, and Goldsmith [**Jafar2004a**] with the goal of enabling new UEs to be added to a cellular network without affecting the rates of existing UEs. Cooperative downlink processing with multi-antenna APs was studied by Zhang and Dai in [**Zhang2004a**]. The concept of *Group Cell* was introduced by Zhang, Tao, Zhang, Wang, Li, and Wang in [**Zhang2005b**] to serve mobile UEs by multiple cells to enable smooth handover during mobility. Multi-cell detection features were also discussed using the group cell name [**Tao2005a**]. Coherent coordinated transmission from the APs based on linear ZF precoding and non-linear dirty paper coding was studied by Foschini, Karakayali, and Valenzuela in [**Foschini2006a**], [**Karakayali2006a**]. The term *Network MIMO* was coined by Venkatesan, Lozano, and Valenzuela in [**Venkatesan2007a**] to describe a cellular network where all the APs within the range of a UE share their received signals over a backhaul network, to turn the cellular uplink from an interference channel to a multiaccess channel. Soft handover between distributed antennas in orthogonal frequency-division multiplexing (OFDM) systems was studied by Tölli, Codreanu, and Juntti in [**Tolli2008a**]. While AP cooperation with infinite-capacity backhaul links was assumed in the above-mentioned works, implementation of joint uplink detection with limited-capacity backhaul was considered by Sanderovich, Somekh, Poor, and Shamai in [**Sanderovich2009a**], while the downlink counterpart was studied by Simeone, Somekh, Poor, and Shamai in [**Simeone2009a**].

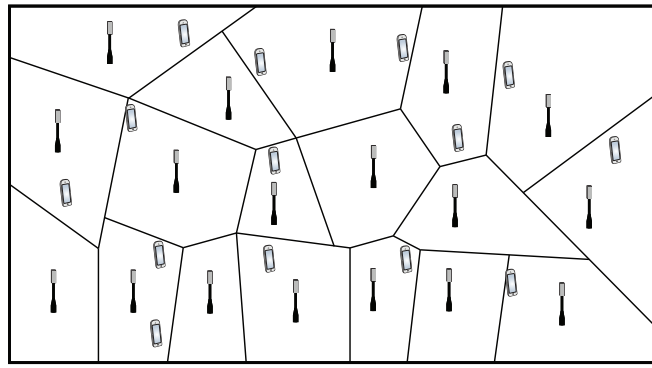
Iterative data detection methods, where the APs exchange soft information to reduce the inter-cell interference, were considered by Khattak, Rave, and Fettweis in [**Khattak2008a**]. Finally, Björnson, Zakhour, Gesbert, and Ottersten showed in [**Bjornson2010c**] that coherent joint transmission can be implemented in time-division duplex (TDD) systems without sharing CSI between the APs, at the cost of increased interference since the AP cannot cancel each others' signals at undesired receivers.

1.2.1 Towards Standardization in 4G

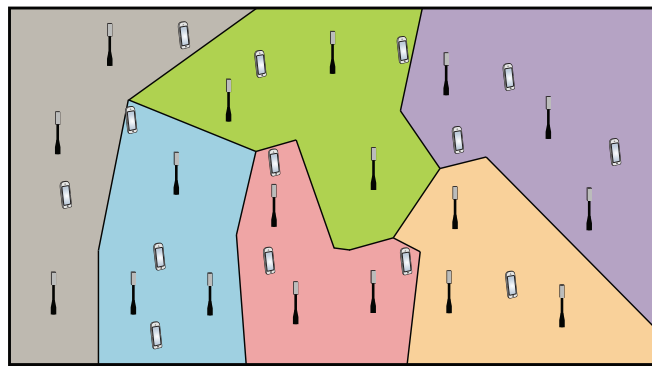
The multi-cell cooperation concepts were considered in the 4G standardization of LTE-Advanced in the late 2000s [**Parkvall2008a**], under the umbrella term of *coordinated multipoint (CoMP)* transmission/reception. The co-processing of data at multiple APs, which is the focus of this monograph, is called *joint processing (JP)* in CoMP [**Boldi2011a**]. Other CoMP options are coordinated scheduling/precoding where each cell only serves its own UEs, which fall into the category of methods that can be also implemented in conventional cellular networks. Both centralized and decentralized architectures for facilitating JP were explored in the context of CoMP. In the centralized approach, the cooperating APs are connected to a CPU (which might be co-located with an AP) and send their information to it. Hence, the APs can be also viewed as relays that facilitate communication between UEs and the CPU [**Estella2019a**]. In the decentralized approach, the cooperating APs only acquire CSI from the UEs [**Papadogiannis2009a**], but data must still be shared between APs.

Network-Centric Clustering

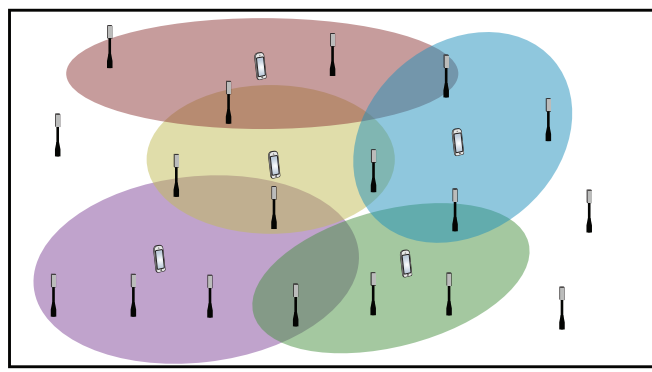
Since each UE in a conventional cellular network would only be affected by interference from its own cell and a set of neighboring cells, it is only the corresponding cluster of APs that needs to cooperate to alleviate inter-cell interference for this UE. Different ways to implement the AP clustering was explored alongside the development of LTE-Advanced [**Boldi2011a**]. The starting point for the clustering is that a cellular network already exists and needs to be improved. We will use the ex-



(a) A conventional cellular network, where each UE is only served by one AP.



(b) A network-centric implementation of CoMP in a cellular network, where the APs are divided into disjoint clusters. The UEs in a cluster are jointly served by the APs in that cluster.



(c) A user-centric implementation of CoMP in a cellular network, where each UE selects a set of preferred APs that will serve it. This is the approach taken also in cell-free networks.

Figure 1.7: Comparison between a conventional cellular network and two ways of implementing multi-cell cooperation.

ample in Figure 1.7(a) to explain the clustering approaches. The first option is *network-centric clustering* where the APs are divided into disjoint clusters [**Huang2009b**], [**Marsch2008a**], [**Zhang2009b**], each serving a disjoint set of UEs. For example, groups of three neighboring cells can be clustered into a joint region, as illustrated by the colored regions in Figure 1.7(b). Compared to the conventional cellular network in Figure 1.7(a), the cell edges within each cluster are removed, but interference will still occur between clusters. Hence, UEs that are close to a cluster edge might not benefit from the network-centric clustering. The clusters can be changed over time or frequency in an effort to make sure that most of the served UEs are in the center of a cluster and not at the edges [**Cayirci2002a**], [**Jungnickel2014a**], [**Marsch2011a**], [**Papadogiannis2009a**]. The network-centric clustering is conceptually similar to having a conventional cellular network where each cell contains a set of distributed antennas that are controlled by a single AP [**Choi2007a**]. Each cell in such a setup corresponds to one cluster in the network-centric clustering.

User-Centric Clustering

Another option is *user-centric clustering* where each UE selects a set of preferred APs [**Bjornson2011a**], [**Bjornson2013d**], [**Chen2016a**], [**Garcia2010a**], [**Kaviani2012a**], [**Xu2013a**], [**Zhang2005b**]. This is illustrated for five UEs in Figure 1.7(c), where each colored region corresponds to the set of APs selected by the corresponding UE. Note that the sets are partially overlapping between neighboring UEs, thus disjoint AP clusters cannot be created to achieve the same result. Irrespective of the UE's location, user-centric clustering will guarantee the control of interference. In this monograph, we will make use of the *dynamic cooperation clustering (DCC)* framework for user-centric clustering, which was introduced by Björnson, Jaldén, Bengtsson, and Ottersten in [**Bjornson2011a**].

If the clusters are well designed, user-centric clustering outperforms network-centric clustering since the latter is essentially a special case of the former. However, both approaches are complicated to add to an existing cellular network since the interfaces between the APs must

be standardized to enable cooperation among AP equipment from different vendors. When potential solutions were simulated in the 4G standardization body, the performance gains were often so small that the additional control signaling might remove the gains [**Boldi2011a**]. An important reason was that the algorithms were jointly designed for frequency-division duplex (FDD) and TDD systems, thus they could not exploit the particular features that only exist in one of these duplexing modes. In particular, CSI for downlink precoding had to be sent around between the APs over low-latency backhaul links to make the system work [**Fantini2016a**]. It is only in a pure TDD implementation exploiting uplink-downlink channel reciprocity that the CSI necessary for downlink precoding can be obtained at each AP without backhaul signaling [**Bjornson2010c**]. We return to this later in this section.

In Release 10 of LTE-Advanced, only a special case of network-centric clustering was supported [**Boldi2011a**]: each cluster consists of APs that are deployed on the same physical site to cover different geographical sectors. Such clustering can only limit the interference between cell sectors, but not between UEs at cell edges. Despite the lack of standardization, the major vendors of AP hardware have made proprietary implementations of CoMP with JP that can only be applied among their own APs. These solutions are often implemented using the C-RAN architecture, which was briefly introduced in Section 1.1. In this cellular context, a set of neighboring APs is connected via a low-latency fronthaul to an edge-cloud processor where the baseband processing is carried out. CoMP algorithms can be conveniently implemented in such a setup. It is not publicly known what CoMP methods are used by different vendors and how well the implementations perform. However, the pCell technology from Artemis [**Perlman2015a**] is claimed to utilize user-centric clustering.

1.2.2 Cellular Massive MIMO in 5G

Instead of focusing on CoMP, the new feature in the 5G cellular networks is Massive MIMO. This concept was introduced by Marzetta in [**Marzetta2010a**] from 2010 and essentially means that each AP *operates individually* and is equipped with an array of a very large number of active low-gain antennas that can be individually controlled

using separate radios (transceiver chains). This stands in contrast to the passive high-gain antennas traditionally used in cellular networks, which might have similar physical dimensions but only a single radio. Massive MIMO has its roots in *space-division multiple access* [**Anderson1991a**], [**Richard1996a**], [**Swales1990a**], [**Winters1987a**], which was introduced in the 1980s and 1990s to enable multiple UEs to be served by an AP at the same time and frequency. The antenna arrays enable directional transmission to each UE (and directional reception from them), thus UEs located at different locations in the same cell can be served simultaneously with little interference. This technology has later been known as multi-user MIMO.

Benefits

The characteristic feature of Massive MIMO, compared to traditional multi-user MIMO, is that each AP has many more antennas than there are active UEs in the cell. Two important propagation phenomena appear in those cases [**Larsson2014a**], [**Rusek2013a**]: *channel hardening* and *favorable propagation*. The former means that fading channels behave almost as deterministic channels if the antenna signals are processed properly to neutralize the small-scale fading. In principle, the processing makes use of the massive spatial diversity offered by having many antennas. Favorable propagation means that the channels of spatially separated UEs are nearly orthogonal in the spatial domain, since transmission and reception are very spatially directive. We will describe these phenomena in detail in Section 2.6 on p. 228. Motivated by the second phenomenon, it was initially claimed that low-complexity interference-ignoring signal processing methods, such as maximum ratio (MR) processing, are close-to-optimal when each AP is equipped with a large number of antennas. It has later been established that more advanced linear signal processing methods, such as *minimum mean-squared error (MMSE) processing*, are needed to make efficient use of Massive MIMO [**BjornsonHS17**], [**Hoydis2013a**], [**Neumann2017a**], [**Sanguinetti2019a**]. In essence, this means that interference must be actively suppressed (one cannot rely on it disappearing automatically when there are many antennas), but the loss in desired signal power is small and there is little need for non-linear methods

such as successive interference cancellation [**massivemimobook**].

A rigorous framework for analyzing the achievable data rates under imperfect CSI was developed in the Massive MIMO literature and is summarized in recent textbooks, such as [**massivemimobook**], [**Marzetta2016a**]. Many tools from this framework will be also utilized in later sections of this monograph.

Limitations

As illustrated in Figure 1.2 earlier in this chapter, Massive MIMO can increase the data rates in a cellular network compared to conventional technology, but large rate variations and inter-cell interference will still remain. Moreover, the 64-antenna panels that have been deployed in 5G cellular networks are not uniform linear arrays (ULAs), as is normally explicitly or implicitly assumed in the Massive MIMO literature [**massivemimobook**], [**Marzetta2016a**], but compact planar arrays that can be deployed in the same way as conventional antennas. Since the horizontal width of an array determines its ability to separate UEs located in different azimuth angles with respect to the array (i.e., wider arrays mean better spatial resolution), the service quality provided by planar arrays is far from what is presented in the literature [**Aslam2019a**], [**Bjornson2019d**]. In summary, Massive MIMO is a solution to some of the interference problems that are faced in conventional cellular networks. However, a cellular deployment of physically wide horizontal ULAs is practically questionable since it greatly deviates from the form factor of conventional cellular APs. Even if this practical barrier is overcome, the large variations in the distance to the served UEs will still lead to large rate variations of the kind illustrated in Figure 1.2. Hence, a different deployment architecture is required to deliver a more uniform service quality over the coverage area.

1.2.3 Cell-Free Networks Beyond 5G

The *cell-free* terminology was coined by Yang and Marzetta in [**Yang2013b**] from 2013, while the name *Cell-free Massive MIMO* first appeared in [**Ngo2015a**] by Ngo, Ashikhmin, Yang, Larsson, and Marzetta from

2015. While most of the research described earlier adds multi-cell cooperation to an existing cellular network architecture, Cell-free Massive MIMO instead follows in the footsteps of the Distributed Wireless Communication System concept from [**Zhou2003a**], where a network consisting of distributed cooperating antennas is designed from the outset. The word "massive" refers to an envisioned operating regime with many more APs than UEs [**Ngo2015a**], and is as an analogy to the conventional Massive MIMO regime in cellular networks; that is, having many more antennas at the infrastructure side than UEs to be served. Interestingly, the envisioned operating regime coincides with that of ultra-dense networks [**Hwang2013a**], [**Kamel2016a**], [**Stefanatos2014a**], but with the core difference that the APs are cooperating to form a distributed antenna array. The original motivation of Cell-free Massive MIMO was to provide an almost uniformly high service quality in a given geographical area [**Ngo2015a**], as illustrated in Figure 1.5.

Background

The cell-free architecture, shown in Figure 1.3 and Figure 1.4, was analyzed in the early works [**Nayebi2017a**], [**Ngo2017b**] with the focus on a distributed operation where the APs perform all the signal processing tasks, except for those that critically require central coordination. The system operates in TDD mode, which means that the uplink and downlink take place in the same frequency band but are separated in time. Hence, the downlink/uplink channels can be jointly estimated by sending known pilot signals from the UEs to the APs. In this way, each AP obtains local CSI regarding the channels between itself and the different UEs. In the downlink setup studied in [**Nayebi2017a**], [**Ngo2017b**], each data signal is encoded at a CPU and sent over the fronthaul to the APs, which transmit the signals using MR precoding based on the locally available CSI. Similarly, in the uplink, each AP applies MR combining locally and sends its soft data estimates over the fronthaul to the CPU, which makes the final decoding without having access to any CSI. This concept is well aligned with the cellular joint transmission framework from [**Bjornson2010c**], where the APs only make use of local CSI obtained from uplink pilots in TDD mode. Variations of this type of distributed processing

can be found in [**Bashar2019a**], [**Bjornson2019c**], [**Buzzi2017a**], [**Fan2019a**], [**Ozdogan2018a**], [**Yang2019a**], [**Zhang2018a**]. One key insight from the more recent works is that the performance can be greatly improved by using MMSE processing instead of MR [**Bjornson2019c**], which is in line with what has also been observed in the Cellular Massive MIMO literature [**BjornsonHS17**], [**Sanguinetti2019a**]. Hence, even if favorable propagation effects can be observed also in cell-free networks with many distributed antennas [**Chen2018b**], it remains important to design the signal processing schemes to actively suppress interference.

The data rates can be also improved by semi-centralized implementations, potentially, at the cost of additional fronthaul signaling. One option is to provide the CPU with statistical CSI so that it can optimize how the uplink data estimates from the APs are combined by taking their relative accuracy into account [**Adhikary2017a**], [**Bjornson2019c**], [**Nayebi2016a**], [**Ngo2018b**]. For example, an AP that is close to the UE should have more influence than an AP that is further away or that is subject to strong interference. Another option is to let the CPU take care of all the processing while the APs only act as relays [**Bashar2018a**], [**Bjornson2019c**], [**Chen2018b**], [**Nayebi2016a**], [**Riera2018a**], [**Yang2019a**]. These different options will be analyzed in detail in later sections of this monograph.

The Roots of Cell-Free Massive MIMO

The first papers on Cell-free Massive MIMO assumed all UEs are served by all APs, while the user-centric clustering from the CoMP literature was first considered in the cell-free context in [**Buzzi2017a**], [**Buzzi2017b**]. A new practical implementation of such clustering was proposed in [**Interdonato2019a**], by first dividing the APs into clusters in a network-centric fashion and then let each UE select a preferred subset of the network-centric clusters. A framework for creating the user-centric clusters in a decentralized fashion was proposed in [**Bjornson2020a**], where the scalability of the different signal processing tasks was also analyzed. Similar user-centric clustering concepts exist in the literature on ultra-dense networks [**Chen2016a**].

Many of the concepts described in the Cell-free Massive MIMO liter-

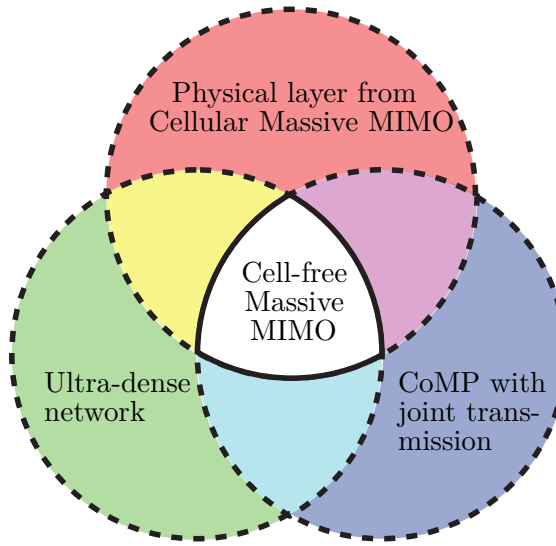


Figure 1.8: Cell-free Massive MIMO can be defined as the intersection between three technology components: The physical layer from Cellular Massive MIMO, the joint transmission concept for distributed APs in the CoMP literature, and the deployment regime of ultra-dense networks.

ature have previously (or simultaneously) appeared and been analyzed in the cellular literature; for example, in some of the papers mentioned earlier in this section. With this in mind, there are two approaches to defining Cell-free Massive MIMO. The first approach is to specify its unique characteristics. As illustrated by the Venn diagram in Figure 1.8, it can be viewed as the intersection between the physical layer from the Cellular Massive MIMO literature, the joint transmission concept for distributed APs in the CoMP literature, and the deployment regime of ultra-dense networks. This corresponds to the inner region in the diagram. In other words, we take the best aspects from three technologies, combine them into a single network, and then jointly optimize them to achieve an ultimate embodiment of a wireless network. The second approach is to view Cell-free Massive MIMO as the union of the three circles; that is, an overarching concept focused on cell-free networks but which contains conventional Massive MIMO, conventional CoMP, and conventional ultra-dense networks as three special cases.

The presentation of the technical content of this monograph will follow the first approach, thus it is that narrow definition that should be remembered when reading the term "Cell-free Massive MIMO" in later sections. We will focus on describing the foundations of Cell-free Massive MIMO, including the state-of-the-art signal processing and optimization methods. We will focus on how a user-centric viewpoint can be used to identify a scalable implementation, which are two dimensions that are not captured by the Venn diagram. We will compare the achievable performance with that of Cellular Massive MIMO and small cells, which we will extract as two special cases from our analytical formulas. The presentation is not based on a particular set of papers, but is an attempt to summarize the topic as a whole.

1.3 Three Benefits over Cellular Networks

We will end this section by showcasing three major benefits that cell-free networks have compared to conventional cellular networks. More precisely, we compare the setups illustrated in Figure 1.9. The first one is a single-cell setup with a 64-antenna Massive MIMO AP, the second one consists of 64 small cells deployed on a square grid, and the last one is a cell-free network where the same 64 AP locations are used. The comparison of these setups will be made by presenting basic mathematical expressions and simulation results, while a more in-depth analysis of cell-free networks will be provided in later sections.

1.3.1 Benefit 1: Higher SNR With Smaller Variations

The first benefit of the cell-free architecture is that it achieves a higher and more uniform SNR within the coverage area than conventional cellular networks. To explain this, we assume there is only one active UE in the network and quantify the SNR that the UE achieves in the uplink, when the UE's transmit power is p and the noise power is σ^2_{ul}. In each of the three setups in Figure 1.9, there are 64 antennas. The received power is substantially lower than the transmit power in wireless communications. For the sake of argument, we model the *channel gain* (also known as pathloss or large-scale fading coefficient)

for a propagation distance d as (in decibels)

$$\beta(d)\,[\mathrm{dB}] = -30.5 - 36.7\log_{10}\left(\frac{d}{1\,\mathrm{m}}\right). \tag{1.1}$$

The first term says that 30.5 dB of the power is lost at 1 m distance while the second term says that another 36.7 dB of power is lost for every ten-fold increase in the propagation distance. All channels are deterministic and thus known to the transmitters and receivers in this section. A more realistic channel model is provided in Section 2.5 on p. 221 and is then used in the remainder of the monograph.

Massive MIMO Setup

We first consider the single-cell Massive MIMO setup in Figure 1.9(a), where the AP is equipped with $M = 64$ antennas. This represents one cell in a cellular network. We denote by $\mathbf{g} = [g_1 \ \ldots \ g_M]^{\mathrm{T}} \in \mathbb{C}^M$ the channel response between the UE and the M antennas. The received uplink signal $\mathbf{y}^{\mathrm{MIMO}} \in \mathbb{C}^M$ at the AP is

$$\mathbf{y}^{\mathrm{MIMO}} = \mathbf{g}s + \mathbf{n} \tag{1.2}$$

where $s \in \mathbb{C}$ is the information signal with transmit power $\mathbb{E}\{|s|^2\} = p$ and $\mathbf{n} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_M, \sigma^2_{\mathrm{ul}}\mathbf{I}_M)$ is the receiver noise.

The main task for the AP is to estimate s and this can be done by applying a receive combining vector $\mathbf{v} \in \mathbb{C}^M$ to (1.2), which leads to

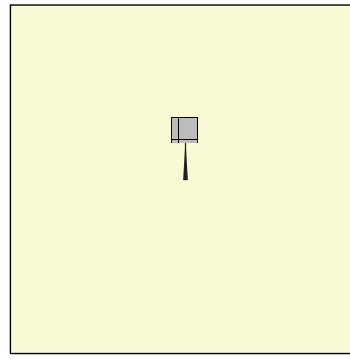
$$\hat{s}^{\mathrm{MIMO}} = \mathbf{v}^{\mathrm{H}}\mathbf{y}^{\mathrm{MIMO}} = \mathbf{v}^{\mathrm{H}}\mathbf{g}s + \mathbf{v}^{\mathrm{H}}\mathbf{n}. \tag{1.3}$$

From this expression, it is clear that the SNR is

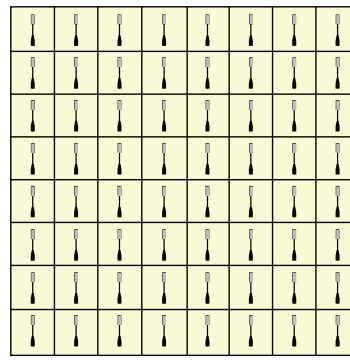
$$\frac{\mathbb{E}\{|\mathbf{v}^{\mathrm{H}}\mathbf{g}s|^2\}}{\mathbb{E}\{|\mathbf{v}^{\mathrm{H}}\mathbf{n}|^2\}} = \frac{p}{\sigma^2_{\mathrm{ul}}}\frac{|\mathbf{v}^{\mathrm{H}}\mathbf{g}|^2}{\|\mathbf{v}\|^2}. \tag{1.4}$$

The AP can select $\mathbf{v}$ based on the channel $\mathbf{g}$ to maximize the SNR. It follows from the Cauchy-Schwartz inequality that (1.4) is maximized when $\mathbf{v}$ and $\mathbf{g}$ are parallel vectors. In particular, the unit-norm MR combining vector $\mathbf{v} = \mathbf{g}/\|\mathbf{g}\|$ can be used to obtain the maximum SNR

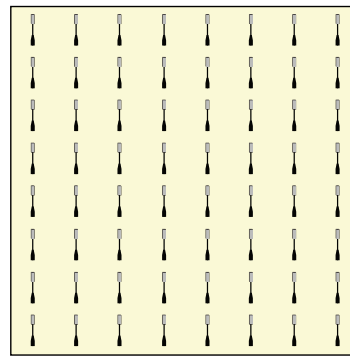
$$\mathsf{SNR}^{\mathrm{MIMO}} = \frac{p}{\sigma^2_{\mathrm{ul}}}\|\mathbf{g}\|^2. \tag{1.5}$$



(a) One cell with a 64-antenna AP in a cellular setup.



(b) Cellular setup with 64 single-antenna APs.



(c) Cell-free setup with the same AP locations as in (b).

Figure 1.9: Three basic setups are compared in Section 1.3: Two cellular networks and one cell-free network (connected to the CPU via fronthaul links, not shown for simplicity).

Since all the antennas are co-located in a big array at the AP, there is the same propagation distance d from the UE to all antennas. Hence, $|g_m|^2 = \beta(d)$ for $m = 1, \ldots, M$ using the channel gain model in (1.1). We then obtain

$$\mathsf{SNR}^{\text{MIMO}} = \frac{p}{\sigma_{\text{ul}}^2} M\beta(d) \tag{1.6}$$

which shows that, in a single-cell Massive MIMO system, the SNR is proportional to the number of antennas, M.

Cellular Setup With Small Cells

In the cellular setup in Figure 1.9(b), there are $L = 64$ geographically distributed APs. Each one has a single antenna and the UE will only be served by one of them. We let $h_l \in \mathbb{C}$ denote the channel response between the UE and AP l. In the uplink, the received signal $y_l^{\text{small-cell}} \in \mathbb{C}$ at AP l is

$$y_l^{\text{small-cell}} = h_l s + n_l \tag{1.7}$$

where $s \in \mathbb{C}$ denotes the information signal that satisfies $\mathbb{E}\{|s|^2\} = p$ and $n_l \sim \mathcal{N}_{\mathbb{C}}(0, \sigma_{\text{ul}}^2)$ is the receiver noise. The SNR at AP l is

$$\mathsf{SNR}_l^{\text{small-cell}} = \frac{\mathbb{E}\{|h_l s|^2\}}{\mathbb{E}\{|n_l|^2\}} = \frac{p}{\sigma_{\text{ul}}^2}|h_l|^2. \tag{1.8}$$

The UE needs to choose only one of the APs since this is a conventional cellular network with no cooperation among APs. The UE will naturally select the one providing the largest SNR. Hence, the SNR experienced by the UE becomes

$$\begin{aligned} \mathsf{SNR}^{\text{small-cell}} &= \max_{l \in \{1,\ldots,L\}} \mathsf{SNR}_l^{\text{small-cell}} \\ &= \frac{p}{\sigma_{\text{ul}}^2} \max_{l \in \{1,\ldots,L\}} |h_l|^2. \end{aligned} \tag{1.9}$$

If we let d_l denote the distance between the UE and AP l, then $|h_l|^2 = \beta(d_l)$ and the SNR in (1.9) can be rewritten as

$$\mathsf{SNR}^{\text{small-cell}} = \frac{p}{\sigma_{\text{ul}}^2} \max_{l \in \{1,\ldots,L\}} \beta(d_l). \tag{1.10}$$

Cell-Free Setup

In the cell-free setup in Figure 1.9(c), we have the same L APs as in the previous small-cell setup, but the APs are now cooperating to serve the UE. We can write the received signals in (1.7) jointly as

$$\mathbf{y}^{\text{cell-free}} = \mathbf{h}s + \mathbf{n} \tag{1.11}$$

where $\mathbf{h} = [h_1 \ \ldots \ h_L]^{\mathrm{T}}$ and $\mathbf{n} = [n_1 \ \ldots \ n_L]^{\mathrm{T}}$. Similar to the single-cell Massive MIMO case above, a receive combining vector $\mathbf{v} \in \mathbb{C}^L$ can be applied to (1.11) in an effort to estimate s. This leads to

$$\hat{s}^{\text{cell-free}} = \mathbf{v}^{\mathrm{H}}\mathbf{y}^{\text{cell-free}} = \mathbf{v}^{\mathrm{H}}\mathbf{h}s + \mathbf{v}^{\mathrm{H}}\mathbf{n}. \tag{1.12}$$

Since this equation has the same structure as (1.3), it follows that MR combining with $\mathbf{v} = \mathbf{h}/\|\mathbf{h}\|$ provides the maximum SNR:

$$\mathsf{SNR}^{\text{cell-free}} = \frac{p}{\sigma_{\text{ul}}^2}\|\mathbf{h}\|^2 = \frac{p}{\sigma_{\text{ul}}^2}\sum_{l=1}^{L}|h_l|^2. \tag{1.13}$$

If we compare this expression with that for the small-cell network in (1.9), we observe that the cell-free network obtains an SNR proportional to $\sum_{l=1}^{L}|h_l|^2$, while the small-cell setup only contains the largest term in that sum. Hence, the cell-free network will always obtain a larger SNR, but the difference will be small if there is one term that is much larger than the sum of the others.

If we instead compare the cell-free setup with the single-cell Massive MIMO setup, the main difference is due to the channels $\mathbf{h}$ and $\mathbf{g}$. The SNRs are proportional to $\|\mathbf{h}\|^2$ and $\|\mathbf{g}\|^2$, respectively. We cannot conclude from the mathematical expressions which of these squared norms is the largest. It will depend on the UE location. Therefore, we need to continue the comparison using simulations. Recall that d_l denotes the distance between AP l and the UE, thus we can also write (1.13) as

$$\mathsf{SNR}^{\text{cell-free}} = \frac{p}{\sigma_{\text{ul}}^2}\sum_{l=1}^{L}\beta(d_l). \tag{1.14}$$

Numerical Comparison

We will now compare the three setups in Figure 1.9 by simulation when the total coverage area is 400 m $\times$ 400 m. We will drop one UE uniformly

at random in the area and compute the uplink SNRs as described above, assuming the transmit power is $p = 10\,\text{dBm}$ and the noise power is $\sigma^2_{\text{ul}} = -96\,\text{dBm}$, which are reasonable values when the bandwidth is 10 MHz. When computing the propagation distances, we assume the APs are deployed 10 m above the UEs.

Figure 1.10 shows the cumulative distribution function (CDF) of the SNR achieved by the UE at different random locations. In the single-cell Massive MIMO case, there are 50 dB SNR variations, where the largest values are achieved when the UE is right underneath the AP and the smallest values are achieved when the UE is in the corner. The SNR variations are much smaller for the cell-free network, since the distances to the closest AP is generally much shorter than in the Massive MIMO case. Moreover, the SNR is higher at the vast majority of UE locations. If we look at the 95% likely SNR, indicated by the dashed line where the CDF value is 0.05, there is an 18 dB difference. More precisely, the cell-free network guarantees an SNR of 24.5 dB (or higher) at 95% of all UE locations, while Massive MIMO only guarantees 6.5 dB. It is only in the upper end of the CDF curves (representing the most fortunate UE locations) that Massive MIMO is the preferred option. This represents the case when the UEs are very close to the 64-antenna Massive MIMO array, while a UE can only be close to a few AP antennas at a time in the cell-free network.

As expected from the analytical expressions, the cell-free network always achieves a higher SNR than the corresponding small-cell setup. The difference is negligible in the upper end of the CDF curves, when the UE is very close to only one of the APs so there is a single dominant term in (1.14), while there is a 4 dB gap in the 95% likely SNR. Based on this example, we can conclude that distributed antennas are preferred over large co-located arrays, but the cell-free architecture only has a minor benefit compared to the cellular small-cell network having the same AP locations. To observe a more convincing practical benefit of the cell-free approach, we need to consider a setup with multiple UEs so that there is interference between the concurrent transmissions.

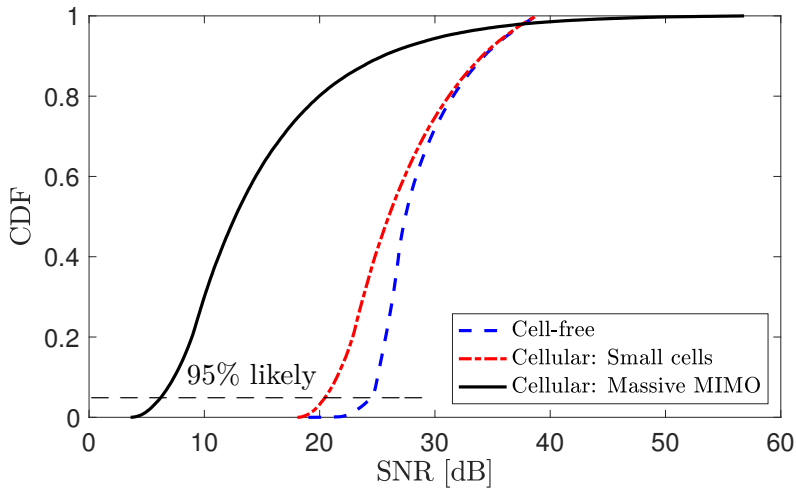


Figure 1.10: The SNR achieved by a UE in each of the setups illustrated in Figure 1.9. The UE location is selected uniformly at random in the area, which gives rise to the CDFs.

1.3.2 Benefit 2: Better Ability to Manage Interference

We will now demonstrate that cell-free networks have the ability to manage interference, which is what small-cell networks are lacking. For the sake of argument, we once again consider the uplink of the three setups shown in Figure 1.9 but now with $K = 8$ UEs. We let p denote the transmit power used by each UE, while σ^2_{ul} denotes the noise power.

Massive MIMO Setup

In the single-cell Massive MIMO setup in Figure 1.9(a), we let $\mathbf{g}_k \in \mathbb{C}^M$ denote the channel from UE k to the AP. Similar to (1.2), the received uplink signal becomes

$$\mathbf{y}^{\text{MIMO}} = \sum_{i=1}^{K} \mathbf{g}_i s_i + \mathbf{n} \tag{1.15}$$

where $s_i \in \mathbb{C}$ is the information signal transmitted by UE i (with $\mathbb{E}\{|s_i|^2\} = p$) and $\mathbf{n} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_M, \sigma^2_{\text{ul}}\mathbf{I}_M)$ is the receiver noise. The AP

applies the receive combining vector $\mathbf{v}_k \in \mathbb{C}^M$ to the received signal in (1.15) in an effort to obtain the estimate

$$\hat{s}_k^{\text{MIMO}} = \mathbf{v}_k^{\text{H}} \mathbf{y}^{\text{MIMO}} = \sum_{i=1}^{K} \mathbf{v}_k^{\text{H}} \mathbf{g}_i s_i + \mathbf{v}_k^{\text{H}} \mathbf{n} \tag{1.16}$$

of the signal s_k from UE k. The corresponding SINR is

$$\begin{aligned} \mathsf{SINR}_k^{\text{MIMO}} &= \frac{\mathbb{E}\{|\mathbf{v}_k^{\text{H}} \mathbf{g}_k s_k|^2\}}{\mathbb{E}\left\{ \left| \sum\limits_{\substack{i=1 \\ i \neq k}}^{K} \mathbf{v}_k^{\text{H}} \mathbf{g}_i s_i + \mathbf{v}_k^{\text{H}} \mathbf{n} \right|^2 \right\}} = \frac{|\mathbf{v}_k^{\text{H}} \mathbf{g}_k|^2 p}{\mathbf{v}_k^{\text{H}} \left(p \sum\limits_{\substack{i=1 \\ i \neq k}}^{K} \mathbf{g}_i \mathbf{g}_i^{\text{H}} + \sigma_{\text{ul}}^2 \mathbf{I}_M \right) \mathbf{v}_k} \\ &\leq p \mathbf{g}_k^{\text{H}} \left(p \sum_{\substack{i=1 \\ i \neq k}}^{K} \mathbf{g}_i \mathbf{g}_i^{\text{H}} + \sigma_{\text{ul}}^2 \mathbf{I}_M \right)^{-1} \mathbf{g}_k \end{aligned} \tag{1.17}$$

where the upper bound is achieved by [**massivemimobook**]

$$\mathbf{v}_k = \left(p \sum_{\substack{i=1 \\ i \neq k}}^{K} \mathbf{g}_i \mathbf{g}_i^{\text{H}} + \sigma_{\text{ul}}^2 \mathbf{I}_M \right)^{-1} \mathbf{g}_k. \tag{1.18}$$

We will provide a more detailed derivation later in this monograph. For now, the important thing is that the maximum uplink SINR of UE k is given by (1.17).

Cellular Setup With Small Cells

In the small-cell setup in Figure 1.9(b), we let $h_{kl} \in \mathbb{C}$ denote the channel response between UE k and AP l. Similar to (1.7), the received uplink signal at AP l becomes

$$y_l^{\text{small-cell}} = \sum_{i=1}^{K} h_{il} s_i + n_l \tag{1.19}$$

where $s_i \in \mathbb{C}$ denotes the information signal from UE i (with $\mathbb{E}\{|s_i|^2\} = p$) and $n_l \sim \mathcal{N}_{\mathbb{C}}(0, \sigma_{\text{ul}}^2)$ is the receiver noise. The SINR at AP l with

respect to the signal from UE k is

$$\mathsf{SINR}_{kl}^{\text{small-cell}} = \frac{\mathbb{E}\{|h_{kl}s_k|^2\}}{\mathbb{E}\left\{\left|\sum\limits_{\substack{i=1\\ i\neq k}}^{K} h_{il}s_i + n_l\right|^2\right\}} = \frac{p|h_{kl}|^2}{p\sum\limits_{\substack{i=1\\ i\neq k}}^{K}|h_{il}|^2 + \sigma_{\text{ul}}^2}. \tag{1.20}$$

Each UE selects to receive service from the AP that provides the largest SINR. Hence, the SINR of UE k is

$$\mathsf{SINR}_{k}^{\text{small-cell}} = \max_{l\in\{1,\ldots,L\}} \mathsf{SINR}_{kl}^{\text{small-cell}}. \tag{1.21}$$

The preferred AP might not be the one with the largest SNR due to the interference. It can happen that one AP serves multiple UEs.

Cell-Free Setup

In the cell-free setup in Figure 1.9(c), the L APs from the small-cell setup are cooperating in detecting the information sent from the K UEs. We can write the received signals in (1.19) jointly as

$$\mathbf{y}^{\text{cell-free}} = \sum_{i=1}^{K} \mathbf{h}_i s_i + \mathbf{n} \tag{1.22}$$

where $\mathbf{h}_i = [h_{i1} \ \ldots \ h_{iL}]^{\mathrm{T}}$ and $\mathbf{n} = [n_1 \ \ldots \ n_L]^{\mathrm{T}}$. Similar to the single-cell Massive MIMO case, a receive combining vector $\mathbf{v}_k \in \mathbb{C}^L$ is applied to (1.22) to detect the signal from UE k. This leads to the estimate

$$\hat{s}_k^{\text{cell-free}} = \mathbf{v}_k^{\mathrm{H}}\mathbf{y}^{\text{cell-free}} = \sum_{i=1}^{K} \mathbf{v}_k^{\mathrm{H}}\mathbf{h}_i s_i + \mathbf{v}_k^{\mathrm{H}}\mathbf{n} \tag{1.23}$$

of s_k. The corresponding SINR is

$$\begin{aligned}\mathsf{SINR}_{k}^{\text{cell-free}} &= \frac{\mathbb{E}\{|\mathbf{v}_k^{\mathrm{H}}\mathbf{h}_k s_k|^2\}}{\mathbb{E}\left\{\left|\sum\limits_{\substack{i=1\\ i\neq k}}^{K} \mathbf{v}_k^{\mathrm{H}}\mathbf{h}_i s_i + \mathbf{v}_k^{\mathrm{H}}\mathbf{n}\right|^2\right\}} = \frac{|\mathbf{v}_k^{\mathrm{H}}\mathbf{h}_k|^2 p}{\mathbf{v}_k^{\mathrm{H}}\left(p\sum\limits_{\substack{i=1\\ i\neq k}}^{K}\mathbf{h}_i\mathbf{h}_i^{\mathrm{H}} + \sigma_{\text{ul}}^2\mathbf{I}_M\right)\mathbf{v}_k} \\ &\leq p\mathbf{h}_k^{\mathrm{H}}\left(p\sum_{\substack{i=1\\ i\neq k}}^{K}\mathbf{h}_i\mathbf{h}_i^{\mathrm{H}} + \sigma_{\text{ul}}^2\mathbf{I}_M\right)^{-1}\mathbf{h}_k \end{aligned} \tag{1.24}$$

where the upper bound is achieved by [**massivemimobook**]

$$\mathbf{v}_k = \left(p \sum_{\substack{i=1 \\ i \neq k}}^{K} \mathbf{h}_i \mathbf{h}_i^{\mathrm{H}} + \sigma_{\mathrm{ul}}^2 \mathbf{I}_M \right)^{-1} \mathbf{h}_k. \tag{1.25}$$

Compared to the single UE case studied earlier, it is harder to utilize the SINR expressions derived in this section to deduce which setup will provide the best performance. Intuitively, the cell-free setup will provide higher SINR than the small-cell setup since we are using the optimal combining vector, while one suboptimal option is to let $\mathbf{v}_k$ contain 1 at the position representing the AP with the highest local SINR and 0 elsewhere. That suboptimal selection would lead to the same SINR as in the small-cell setup. To compare the cell-free setup with the single-cell Massive MIMO setup, we need to run simulations since the SINR expressions in (1.17) and (1.24) have a similar form, but contain channel vectors that are generated differently.

Numerical Comparison

We will now simulate the performance of this multi-user setup using the channel gain model in (1.1) and the same parameter values as in Figure 1.10. More precisely, if the propagation distance is d, then the channel is generated as $\sqrt{\beta(d)}e^{\mathrm{j}\phi}$ where ϕ is an independent random variable uniformly distributed between 0 and 2π. This variable models the random phase shift between the transmitter and receiver. This phase was omitted in the previous simulation since the result was determined only by the norms of the channels. However, it is important to include the phases when considering multi-user interference, which is also determined by the directions of the channel vectors.

Figure 1.11 shows the CDF of the SINR achieved by a randomly selected UE in a random realization of the $K = 8$ uniformly distributed UE locations. As compared to Figure 1.10, all the curves are moved to the left in Figure 1.11 due to the interference among the UEs. The Massive MIMO case is barely affected by the interference, which demonstrates that this technology has the ability to separate the UEs' channels spatially using the large array of co-located antennas. However,

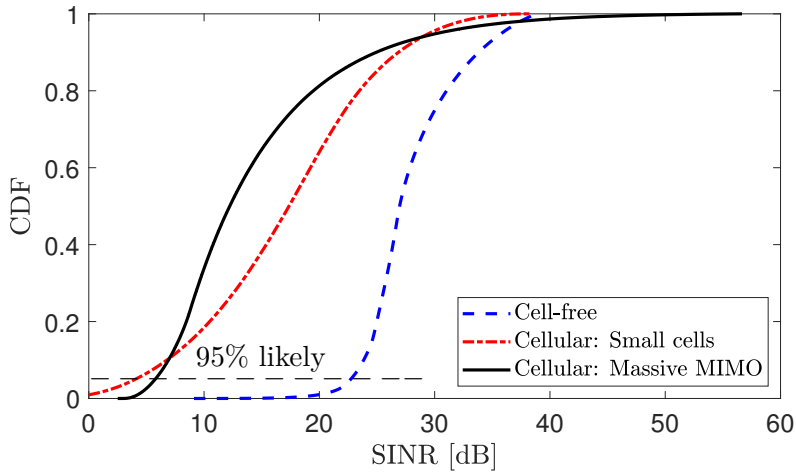


Figure 1.11: The SINR achieved by an arbitrary UE in each of the setups illustrated in Figure 1.9. There are $K = 8$ UEs that are distributed uniformly at random in the area. This gives rise to the CDF.

due to the large variations in distances to the AP, there are 50 dB variations in the SINR between different UE locations. The cell-free network is also barely affected by the interference, but one can see a tiny lower tail that corresponds to the random event that two UEs are randomly deployed at almost the same location.

The major difference from the single-UE case is that the small-cell curve is moved far to the left and the 95%-likely SINR is even lower than with Massive MIMO. The reason is that each AP only has a single antenna and thus cannot suppress inter-cell interference. The cell-free setup is greatly outperforming the small-cell setup in this multi-user setup. This is what will occur in practice since mobile networks are deployed to serve multiple UEs in the same geographical area.

1.3.3 Benefit 3: Coherent Transmission Increases the SNR

The previous two benefits were exemplified in the uplink but there are also counterparts in the downlink, which lead to similar results but for partially different reasons. One important difference is that the received power in the uplink increases with the number of receive antennas (i.e.,

a larger fraction of the transmit power is collected), thus it is always beneficial to have more antennas. Consider now a downlink scenario where we can deploy any number of antennas, but constrain the total downlink transmit power to be constant (to not change the energy consumption). We then need to determine how the power should be divided between the APs to maximize the SNR. Suppose a UE is in the vicinity of two APs but one has a substantially better channel. It might then seem logical that all the transmit power should be assigned to the AP with the better channel, but we will show that this is not the optimal strategy.

Suppose, for the sake of argument, that AP 1 has the channel response $h_1 = \sqrt{\alpha}$ to the UE, while AP 2 has the channel response $h_2 = \sqrt{\alpha/2}$. If we compare the channel gains $|h_1|^2 = \alpha$ and $|h_2|^2 = \alpha/2$, it is clear that AP 1 has the best channel. Let ρ denote the total downlink transmit power and σ_{dl}^2 denote the receiver noise power. If only AP 1 transmits to the UE, then the SNR at the receiver is

$$\frac{\rho\alpha}{\sigma_{\text{dl}}^2}. \tag{1.26}$$

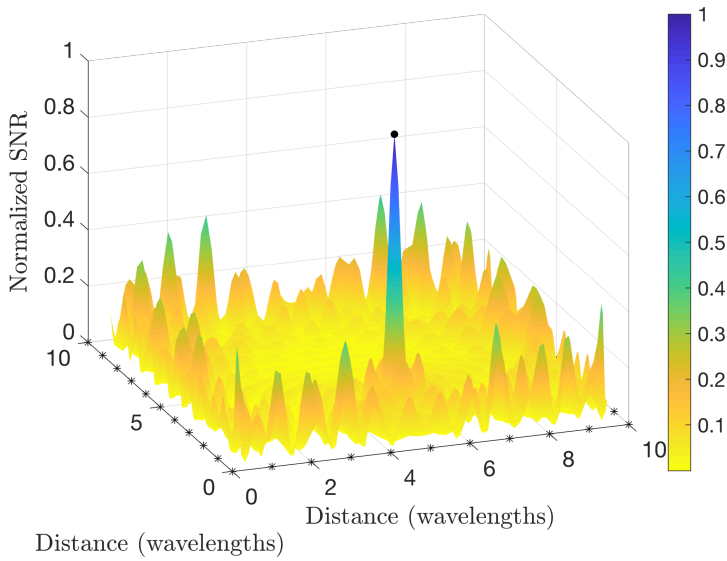
However, if AP 1 instead transmits with power $2\rho/3$ and AP 2 transmits with power $\rho/3$, which also corresponds to a total power of ρ, then the SNR is

$$\frac{1}{\sigma_{\text{dl}}^2}\left(\sqrt{\frac{2\rho}{3}}h_1 + \sqrt{\frac{\rho}{3}}h_2\right)^2 = 1.5\frac{\rho\alpha}{\sigma_{\text{dl}}^2}. \tag{1.27}$$

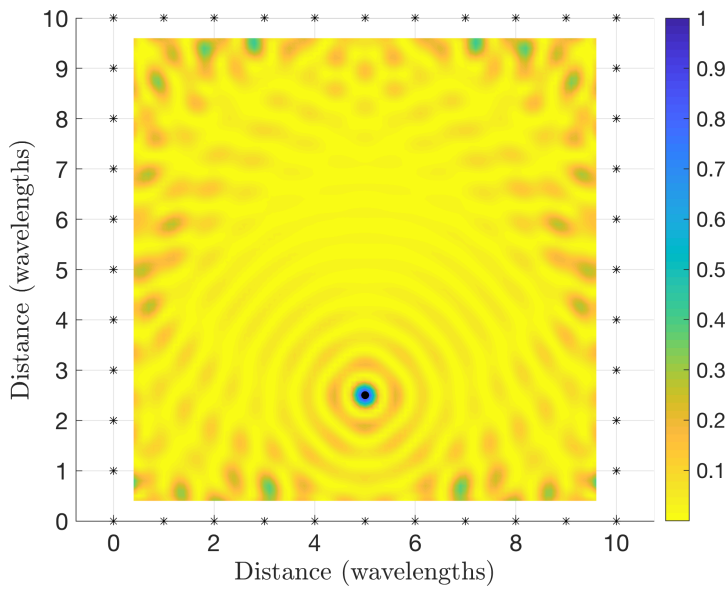
Hence, the SNR is higher when we transmit from both APs. This is a consequence of the coherent combination (i.e., constructive interference) of the signals from the two APs. The power gain is reminiscent of the beamforming gain from co-located arrays, where the transmit power is stronger in some angular directions than in other ones, but the physical interpretation is somewhat different. The coherent combination of signals that are transmitted from different geographical points does not give rise to beam patterns but rather local signal amplification in a region around the receiver. Moreover, all the antennas in a co-located array will experience (roughly) the same channel gain so it is logical that they should be jointly utilized for coherent transmission. In contrast, distributed antennas can experience very different channel gains but are anyway useful for coherent transmission.

We will now illustrate that the signal focusing obtained by a distributed array does not give rise to signal beams. Figure 1.12 shows the SNR variations when transmitting from $M = 40$ antennas that are equally spaced along the perimeter of a square. The adjacent antennas are one-wavelength-spaced and transmit with equal power. The received power decays as the square of the propagation distance (as would be the case in free space). If a narrowband signal is considered, the received signals will be phase-shifted (time-delayed) by the ratio between the propagation distance and the signal's wavelength. For the sake of argument, the distances in Figure 1.12 are therefore measured as fractions of the wavelength (each side of the square is ten wavelengths). To achieve a coherent combination at the point of the receiver, each antenna must phase-shift (time-delay) its signal before transmission to make sure that all the M signals are reaching the receiver perfectly synchronized.

In Figure 1.12(a), we show the SNRs measured at different locations when focusing all the signals into a single point. The SNR values are normalized so that they are equal to 1 at the point-of-interest (i.e., the location of the receiver) and smaller elsewhere. We observe that the SNR is much larger at that point than on all the surrounding points, where the 40 signals are not coherently combined. There are some points near the edges of the simulation area where the SNR is also strong but this is not due to a coherent combination of multiple signal components. Instead, it is because these points are close to some of the transmit antennas. Figure 1.12(b) shows the same results but from above. The figure reveals that the SNR is strong in a circular region around the point-of-interest. The diameter of this region is roughly half-a-wavelength. In summary, the signal focusing from distributed arrays will not give rise to angular beams (as in Cellular Massive MIMO) but local signal focusing around the receiver in a region that is smaller than the wavelength. When considering a three-dimensional propagation environment, the SNR will be large within a sphere around the point-of-interest with the diameter being half-a-wavelength. When transmitting multiple signals, we can focus each one at a different point and if these points are several wavelengths apart, the mutual interference will be small according to Figure 1.12.



(a) View with normalized SNR (between 0 and 1) on the vertical axis.



(b) Same as in (a) but viewed from above.

Figure 1.12: The received signal power at different locations when transmitting from one-wavelength-spaced antennas along the walls (each marked with a star). The antennas transmit with equal power and the signals are phase-shifted to achieve coherent combination at the point where the normalized SNR is 1.

1.4 Summary of the Key Points in Section 1

- Traditional wireless networks use the cellular architecture. The cellular approach was conceived for providing wide-area coverage to low-rate voice services. Each AP is surrounded by UEs at very different distances, having widely different SNRs. This architecture is badly suited for providing ubiquitous access to high-rate data services.
- The cell-free architecture turns the situation around: each UE is surrounded by APs. Each AP has relatively simple hardware and cooperates with surrounding APs to jointly serve the UEs in their area of influence. A cell-free network is user-centric if each UE is served by its nearest APs.
- The name *Cell-free Massive MIMO* signifies that it is the combination of three previously known components: The physical layer of Massive MIMO, the vision of creating ultra-dense networks with many more APs than UEs, and the coordinated multipoint methods for achieving a cell-free network. The main novelty lies in how to co-design these components to achieve a user-centric operation that is sufficiently scalable to enable large-scale deployments.
- The first key benefit of the cell-free architecture is the smaller SNR variations compared to cellular networks with a sparse deployment of APs and Massive MIMO.
- The second key benefit is the ability to manage interference by joint processing at multiple APs, which is not done in cellular networks with an equally dense AP deployment.
- The third key benefit is that coherent transmission increases the SNR. It is better to involve APs with weaker channels in the transmission than only using the AP with the best channel.

2

User-Centric Cell-Free Massive MIMO Networks

This section introduces the basic system model and concepts related to a User-centric Cell-free Massive MIMO network, which will be used in the remainder of this monograph. Section 2.1 provides a formal definition of Cell-free Massive MIMO. The user-centric operation is introduced in Section 2.2 based on the DCC framework. The modulation and communication protocol are defined in Section 2.3, along with the uplink and downlink system models. Section 2.4 defines the meaning of network scalability and reviews the system model from that perspective. A sufficient condition for a cell-free network being scalable is provided. Section 2.5 introduces the channel model that will be used for analysis and performance evaluation in later sections. Section 2.6 defines the concepts of channel hardening and favorable propagation, and analyzes them in the context of cell-free networks. The key points are summarized in Section 2.7.

2.1 Definition of Cell-Free Massive MIMO

In Section 1.2.3 on p. 184, we defined Cell-free Massive MIMO as an ultra-dense network (i.e., many more APs than UEs) where the APs are cooperating to serve the UEs by joint coherent transmission and

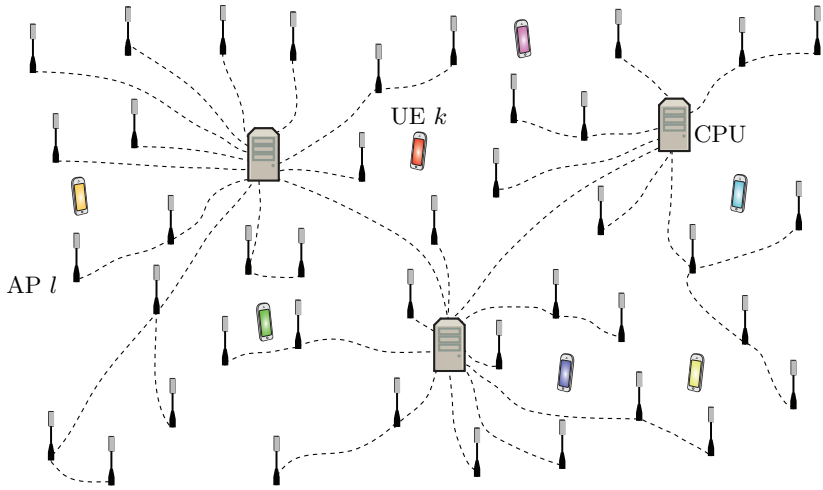


Figure 2.1: A Cell-free Massive MIMO network consists of L geographically distributed APs that jointly serve K geographically distributed UEs. Each AP is equipped with N antennas, while each UE has a single antenna.

reception, while making use of the physical layer concepts from the Cellular Massive MIMO area. We will now turn this into a technical framework and then progressively add more details to it in later sections.

We consider a network with L APs, each equipped with N antennas, that are geographically distributed over the coverage area, as shown in Figure 2.1. We let $M = NL$ denote the total number of AP antennas in the network. The APs are jointly serving K single-antenna UEs. More precisely, each UE is communicating with a subset of the APs, which is selected based on the UE's needs. The APs are connected via fronthaul links to CPUs, which facilitate the AP coordination. The cell-free architecture can also handle multi-antenna UEs, where the extra antennas are utilized to suppress interference or spatially multiplex several signals per UE. This case is not covered in this monograph because it substantially complicates the analysis and one can easily apply the presented theory to the multi-antenna case by treating each UE antenna as a separate UE. More intricate solutions can be found in [**Alonzo2019a**], [**Buzzi2017b**], [**Buzzi2020b**], [**Jin2019a**], [**Li2016b**], [**Mai2020a**].

The intended operating regime of Cell-free Massive MIMO is $L \gg K$ [**Nayebi2017a**], [**Ngo2017b**], which is the mathematical definition of an ultra-dense network. This also implies $M \gg K$; that is, the total number of AP antennas is much larger than the total number of UEs, as in conventional Cellular Massive MIMO systems [**Marzetta2010a**]. The intention is that the cell-free network will, under these circumstances, have sufficiently many spatial degrees-of-freedom so that the UEs can be separated in space by processing the transmitted and received signals. For example, whenever the APs focus the transmission towards a particular UE, the focus area will be sufficiently small so that no other UE receives high interference. Moreover, the number of antennas N per AP is intended to be small and each AP might serve more UEs than its antennas, thus AP cooperation is essential to suppress interference between the UEs. However, we stress that there is no need for a formal assumption of $M \gg K$ or $L \gg K$ in the analysis of User-centric Cell-free Massive MIMO networks. The concepts and signal processing methods presented in this monograph hold for any value of L, K, and N. In fact, we will later extract Cellular Massive MIMO and small-cell networks as two special cases of the framework presented in this monograph, which enables a fair comparison between the technologies.

2.2 User-Centric Dynamic Cooperation Clustering

The first papers on Cell-free Massive MIMO assumed all UEs were served by all APs, as mentioned in Section 1.2.3 on p. 184. This is both impractical and unnecessary in a geographically large network, where each UE is only physically close to a subset of APs. By limiting the set of APs that is allowed to transmit to the UE, the service quality will inevitably be reduced. However, if all APs that can reach the UE with a signal power that is non-negligible compared to the thermal noise power take an active part in serving that UE, then the performance loss should be negligible. This calls for a user-centric approach to the formation of the AP clusters that serve each UE. The practical benefits are that the fronthaul signaling is reduced when only a subset of the APs must receive the downlink data intended for the UE and send

their corresponding estimates of the uplink data to the CPU. Moreover, the computational complexity is reduced when each AP only needs to process signals related to a subset of the UEs.

To model which APs are serving which UEs, we will make use of the DCC framework initially proposed in [**Bjornson2011a**], [**Bjornson2013d**].[1]

Definition 2.1. Dynamic cooperation clustering (DCC) means that UE k is served only by the APs with indices in the set $\mathcal{M}_k \subset \{1, \ldots, L\}$.

The word *dynamic* refers to the fact that the sets $\mathcal{M}_k$ can be adapted to time-variant characteristics, such as UE locations, service requirements, interference situation, etc. Figure 2.2 illustrates one instance of the DCC framework with four UEs and a large number of APs. The colored regions illustrate which clusters of APs are serving which UEs. The APs within a colored region are those with indices in the set $\mathcal{M}_k$ and we select them to involve all the neighboring ones. The fact that the clusters are partially overlapping between the UEs is showcasing that this is a user-centric cell-free network; we cannot divide the APs into disjoint sets that serve disjoint subsets of the UEs. More generally, every AP is co-serving UEs with a set of other APs, which in turn are co-serving other UEs with another set of APs, and the chain continues like this until all APs have been considered (at least when many UEs are distributed over the coverage area).

When analyzing the performance of the data transmission, we assume the sets $\mathcal{M}_k$, for $k = 1, \ldots, K$, are fixed and known everywhere needed. The reason is that the dynamic variations occur over much larger time intervals than channel fading variations. More precisely, the data is transmitted in blocks that are sufficiently large to be subject to many fading realizations but only one realization of $\mathcal{M}_1, \ldots, \mathcal{M}_K$. Different ways to select these sets are later discussed in Section 4.4 on p. 286.

For notational convenience, we also define a set of diagonal matrices $\mathbf{D}_{kl} \in \mathbb{C}^{N \times N}$, for $k = 1, \ldots, K$ and $l = 1, \ldots, L$, determining which APs communicate with which UEs. More precisely, $\mathbf{D}_{kl}$ is the identity

[1] The DCC framework was introduced to enable "*unified analysis of anything from interference channels to Network MIMO*". Therefore, the user-centric cooperation clusters considered in Cell-free Massive MIMO is only one of the many instances of this framework.

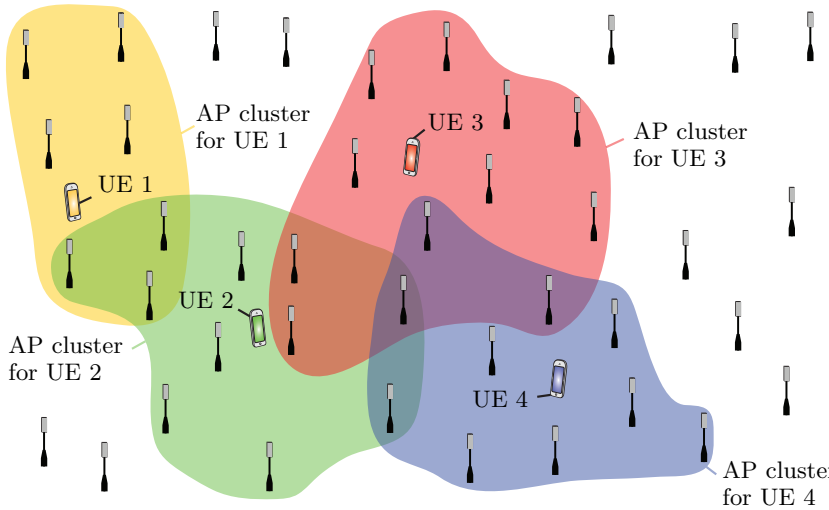


Figure 2.2: Example of dynamic cooperation clusters for four UEs in a Cell-free Massive MIMO network with a large number of APs.

matrix $\mathbf{I}_N$ if AP l is allowed to transmit to and decode signals from UE k and $\mathbf{0}_{N\times N}$ otherwise. Following the notation from Definition 2.1, we have that

$$\mathbf{D}_{kl} = \begin{cases} \mathbf{I}_N & l \in \mathcal{M}_k \\ \mathbf{0}_{N\times N} & l \notin \mathcal{M}_k. \end{cases} \tag{2.1}$$

Note that if we select $\mathcal{M}_k = \{1, \ldots, L\}$, for $k = 1, \ldots, K$ and, hence, all matrices $\mathbf{D}_{kl} = \mathbf{I}_N$, then we obtain the original form of Cell-free Massive MIMO from [**Nayebi2017a**], [**Ngo2017b**], where all APs jointly serve all UEs in the network. Another special case of the DCC framework is a small-cell network, which is obtained by selecting $\mathcal{M}_k$ to have a single element for $k = 1, \ldots, K$, so that each UE is only served by one AP.

2.3 System Models for Uplink and Downlink

In electrical engineering, a *system* is something that takes an input signal and filters it to produce an output signal. A wireless communication channel is an example of such a system: it filters the transmitted signal and outputs the received signal. In this subsection, we motivate and describe the system model for User-centric Cell-free Massive MIMO

that is used in the remainder of this monograph. We begin by describing the block-fading model and introducing the coherence block concept.

2.3.1 Block Fading and Coherence Blocks

The type of systems that are convenient to analyze is linear and time-invariant. Wireless channels are linear due to the superposition principle prescribed by Maxwell's equations. However, the channels are generally time-varying since the transmitter, receiver, and objects in the wireless propagation environment can move. Even minor movements over a distance proportional to the wavelength (i.e., a few millimeters or centimeters) will substantially change the channel. For example, the UE at the focus point in Figure 1.12 on p. 201 only needs to move a quarter-of-a-wavelength to leave the area where the SNR is high. However, if we consider a sufficiently short time interval, the channel can be approximated as constant and, therefore, the communication system is time-invariant. That interval is called *channel coherence time*. This time is determined by the speed of movement and is typically a few milliseconds in mobile networks. Since it is important for the communication system to know the channel properties, it needs to estimate them once per coherence time interval.

Within the coherence time, the channel can be described by a finite impulse response (FIR) filter where each term of the impulse response describes one distinguishable propagation path in a multipath environment with a distinct time delay and pathloss. Such a filter reacts differently to input signals with different frequencies, but the variations are rather smooth when varying the signal frequency. Hence, if we consider a sufficiently narrow frequency range, the channel can be considered constant. The *channel coherence bandwidth* defines the frequency interval in which the frequency response is approximately constant.

A *channel coherence block* is a time-frequency block with a time duration that equals the coherence time and a frequency width that equals the coherence bandwidth. The channel between two antennas is constant and frequency-flat within a coherence block, which implies that it can be described by only one scalar coefficient. The time-frequency

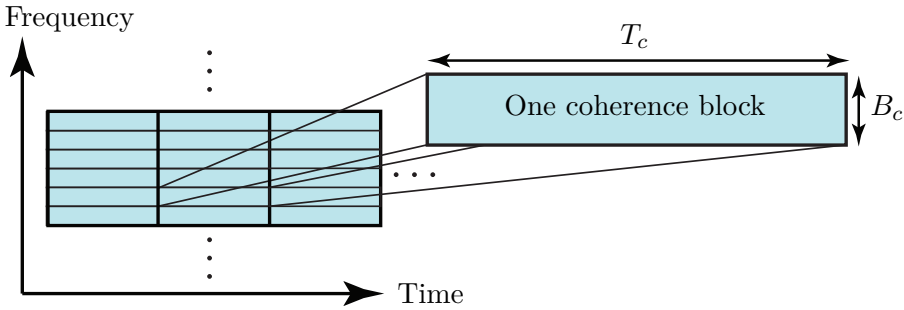


Figure 2.3: In the block-fading model, the time-frequency resources are divided into coherence blocks in which the channel is time-invariant and frequency-flat. The channel is modeled as independent between different coherence blocks.

resources of a communication system can be divided into different coherence blocks as illustrated in Figure 2.3. Note that the blocks are distributed over both time and frequency. When developing communication algorithms, it is convenient to break down the operation into blocks like this and then design, analyze, and optimize them separately. If one further assumes that the channel-describing scalar coefficient takes an independent random realization in each coherence block, then one can study one block at a time without loss of generality. This is called the *block-fading model* since the channel fading takes one independent realization in each coherence block. The block-fading model is assumed throughout this monograph.

Let T_c denote the coherence time (in seconds) and B_c denote the coherence bandwidth (in Hertz). According to the Nyquist-Shannon sampling theorem, a signal that fits into this block is uniquely described by $\tau_c = T_c B_c$ complex-valued samples. These are the parameters that can be used to convey information in a communication system. We will call them *transmission symbols* or just symbols.

In practice, the coherence time and bandwidth depend on many factors such as the channel delay spread (i.e., the time difference between the shortest and longest resolvable paths), UE mobility, and carrier frequency. The exact values are user-dependent and can be measured experimentally, but we will provide some rules-of-thumb. A common approximation of the coherence time is $T_c = \lambda/(4v)$ [**Tse2005a**], where

λ is the wavelength and v is the velocity of the UE. This is the time it takes to move a quarter-of-a-wavelength. Hence, if the carrier frequency is high, the channel changes more rapidly for a given velocity. The same happens if the UE velocity increases for a given carrier frequency. The coherence bandwidth can be approximated by $B_c = 1/(2T_d)$ where T_d is the channel delay spread; that is, the time difference between the earliest and last propagation path.

We need a common block size when operating the system but every UE has different values of T_c and B_c. A practical solution is to fix τ_c based on the worst-case scenario that the network should support [**massivemimobook**]. To give quantitative numbers, suppose the carrier frequency is 2 GHz, which gives the wavelength $\lambda = 15$ cm. We consider two examples:

- Outdoor scenario: Suppose the delay spread is up to 2.5 μs (i.e., 750 m path differences) and we want to support mobility up to $v = 37.5$ m/s $= 135$ km/h. The approximations above then become $T_c = 1$ ms and $B_c = 200$ kHz. The coherence block contains $\tau_c = 200$ transmission symbols in this scenario that supports fairly high mobility and high channel dispersion.

- Indoor scenario: Suppose the delay spread is up to 0.1 μs (or 30 m path differences) and we want to support mobility up to $v = 0.75$ m/s $= 2.7$ km/h. The approximations above then become $T_c = 50$ ms and $B_c = 5$ MHz. The coherence block contains $\tau_c = 250\,000$ transmission symbols in this scenario with low mobility and low channel dispersion.

In summary, the number of transmission symbols per coherence block ranges from hundreds (with high mobility and high channel dispersion) to hundreds of thousands of samples (with low mobility and low channel dispersion). Setups similar to the outdoor scenario with $\tau_c = 200$ are commonly studied in the literature on Cellular Massive MIMO [**massivemimobook**], [**Marzetta2016a**]. If a cell-free network is deployed to serve the same type of UEs as in conventional cellular networks, we can safely assume that $\tau_c \geq 200$. However, the coherence

blocks can be much larger in cell-free networks if we primarily target low-mobility use cases and deploy the antennas closer to the UEs, so that the delay spread shrinks.

Remark 2.1 (Relation to practical systems)**.** The main reason for studying block-fading channels, instead of a system with a practical multicarrier modulation scheme, is that we can neglect many of the technicalities that appear in practical systems. In this way, we can focus on the main concepts and develop a theory that applies to many types of systems. In a practical system where OFDM is utilized, the available bandwidth is divided into many subcarriers and each one features a scalar channel. Several subcarriers will then fit into what we have defined as a coherence block [**Marzetta2016a**]. These subcarriers will not have identical channels, but there is a known transformation between them, thus if one learns the channel coefficient at some of the subcarriers, the other ones can be obtained by interpolation [**Kashyap2016a**]. Moreover, the channel coefficients will not change abruptly every coherence block but gradually. In particular, the channel realizations will be correlated between adjacent coherence blocks, which can actually be utilized to improve the performance compared to the methods described in this monograph. This is especially important for low-mobility UEs for which the channel might be constant for a much longer time than the worst-case value of T_c that is used by the system.

2.3.2 Time-Division Duplex Protocol

The channel between AP l and UE k in an arbitrary coherence block is represented by the vector $\mathbf{h}_{kl} \in \mathbb{C}^N$. It is an N-dimensional vector where the nth element represents the complex-valued scalar channel coefficient between the UE and the nth antenna of the AP. We also define the collective channel $\mathbf{h}_k \in \mathbb{C}^M$ from all APs to UE k as

$$\mathbf{h}_k = \begin{bmatrix} \mathbf{h}_{k1} \\ \vdots \\ \mathbf{h}_{kL} \end{bmatrix}. \tag{2.2}$$

According to the block-fading model assumption, the channel vector $\mathbf{h}_{kl}$ takes one independent realization in each coherence block from

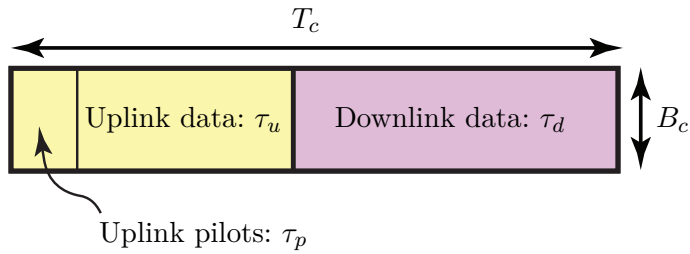


Figure 2.4: A TDD protocol is considered where each coherence block is used for both uplink and downlink transmissions.

a stationary random distribution. Due to the channel reciprocity of physical channels, the channel realization is the same in both uplink and downlink. We will define the channel distribution later in Section 2.5.

The APs must know the channel realizations, at least partially, to perform coherent processing both between the antennas on each AP and across the cooperating APs. The most efficient way to estimate the channels is to consider a TDD protocol where each coherence block is used for both uplink and downlink transmissions. It is then sufficient to transmit pilots only in the uplink. Based on the received uplink signals, each AP can then estimate the channels between itself and all the UEs. These channel estimates can be utilized for both uplink and downlink transmissions, thanks to the channel reciprocity. By applying coherent precoding in the downlink based on the channel estimates, the complex-valued M-dimensional channel vector to each UE is transformed into a scalar channel that is positive (except for minor perturbations due to estimation errors). This scalar channel can be deduced at the UE from the downlink data signals, without sending explicit pilots [**Ngo2017a**]. We will elaborate more on this later in this monograph. We refer to [**massivemimobook**] for a more detailed discussion on this and a comparison with the alternative FDD protocol, in terms of signaling overhead. While the uplink can operate identically in TDD and FDD systems, the downlink must be implemented differently in FDD. Applications of FDD protocols to Cell-free Massive MIMO can be found in [**Abdallah2020a**], [**Kim2020a**], [**Kim2018a**] and will not be covered in this monograph.

We consider the standard Cellular Massive MIMO TDD protocol, where the τ_c transmission symbols of a coherence block are used for three purposes:

1. τ_p symbols for uplink pilots;
2. τ_u symbols for uplink data;
3. τ_d symbols for downlink data.

This protocol is illustrated in Figure 2.4. The three sets of transmission symbols can be divided between the three purposes in different ways, under the constraint that $\tau_c = \tau_p + \tau_u + \tau_d$. Note that the pilots and data are transmitted at different times and that the TDD protocol is synchronized across the APs. The uplink pilots must be transmitted before the downlink data so that the downlink transmission can be precoded based on the uplink estimates. However, it is plausible to transmit uplink data before the uplink pilots, or to change the order between the uplink and downlink data transmissions, as long as it all fits into a single coherence block. The drawback with the former variation is that the latency increases since the data detection cannot be initiated until after the pilots have been received. The drawback with the latter variation is that the switch between downlink and uplink requires a guard time interval so that all UEs receive the downlink data before they switch to uplink transmission mode.

2.3.3 Uplink System Model

During the uplink data transmission, all APs will receive a superposition of the signals sent from all UEs. The received signal $\mathbf{y}_l^{\text{ul}} \in \mathbb{C}^N$ at AP l is

$$\mathbf{y}_l^{\text{ul}} = \sum_{i=1}^{K} \mathbf{h}_{il} s_i + \mathbf{n}_l \tag{2.3}$$

where $s_i \in \mathbb{C}$ is the signal transmitted from UE i with a power that we denote as $p_i = \mathbb{E}\{|s_i|^2\}$ during the uplink data transmission and as $\eta_i = |s_i|^2$ in the uplink pilot transmission. The independent additive receiver noise is $\mathbf{n}_l \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_N, \sigma_{\text{ul}}^2 \mathbf{I}_N)$. The uplink signals can either contain data

or pilots. While the channels are constant within a coherence block, the signals and noise take new realizations at every transmission symbol.

Based on the received signal in (2.3), AP l can compute an estimate $\widehat{s}_{kl}$ of the signal s_k transmitted by UE k. This is the first step towards decoding the information that the signal contains. However, the AP should only do that if $l \in \mathcal{M}_k$ or, equivalently, $\mathbf{D}_{kl} = \mathbf{I}_N$. For notational convenience, we want to represent both cases jointly by setting $\widehat{s}_{kl} = 0$ if $l \notin \mathcal{M}_k$. This is achieved by defining a receive combining vector $\mathbf{v}_{kl} \in \mathbb{C}^N$ that AP l could use if it serves UE k. We then define the *effective receive combining vector*

$$\mathbf{D}_{kl}\mathbf{v}_{kl} = \begin{cases} \mathbf{v}_{kl} & l \in \mathcal{M}_k \\ \mathbf{0}_N & l \notin \mathcal{M}_k. \end{cases} \tag{2.4}$$

Note that it is equal to the actual receive combining vector for APs that serve UE k and zero otherwise. The estimate of s_k at AP l can then be computed by taking the inner product between $\mathbf{D}_{kl}\mathbf{v}_{kl}$ and $\mathbf{y}_l^{\mathrm{ul}}$ (noting that $\mathbf{D}_{kl} = \mathbf{D}_{kl}^{\mathrm{H}}$):

$$\widehat{s}_{kl} = \mathbf{v}_{kl}^{\mathrm{H}}\mathbf{D}_{kl}\mathbf{y}_l^{\mathrm{ul}}. \tag{2.5}$$

By utilizing (2.4), the expression in (2.5) can be rewritten as

$$\widehat{s}_{kl} = \begin{cases} \mathbf{v}_{kl}^{\mathrm{H}}\mathbf{h}_{kl}s_k + \sum\limits_{\substack{i=1 \\ i\neq k}}^{K} \mathbf{v}_{kl}^{\mathrm{H}}\mathbf{h}_{il}s_i + \mathbf{v}_{kl}^{\mathrm{H}}\mathbf{n}_l & l \in \mathcal{M}_k \\ 0 & l \notin \mathcal{M}_k. \end{cases} \tag{2.6}$$

However, we will be utilizing the general expression in (2.5) in the majority of this monograph since it allows us to cover both cases in a joint manner, instead of treating them separately.

The receive combining vectors $\mathbf{v}_{kl}$, for $l \in \mathcal{M}_k$, should be selected based on the CSI available at the APs. Different designs based on various degrees of cooperation among the APs will be introduced and analyzed in detail in Section 5 on p. 294. Channel estimation is considered in Section 4 on p. 264.

2.3.4 Downlink System Model

In the downlink, we let $\mathbf{x}_l \in \mathbb{C}^N$ denote the signal transmitted by AP l. The received signal at UE k can then be written as

$$y_k^{\mathrm{dl}} = \sum_{l=1}^{L} \mathbf{h}_{kl}^{\mathrm{H}} \mathbf{x}_l + n_k \tag{2.7}$$

where $n_k \sim \mathcal{N}_{\mathbb{C}}(0, \sigma_{\mathrm{dl}}^2)$ is the receiver noise. Note that we have represented the downlink channel by $\mathbf{h}_{kl}^{\mathrm{H}}$, although there will only be a transpose and not any complex conjugate in practice. However, the additional conjugation is a standard way of simplifying the notation without changing the performance. We will therefore adopt this convention.[2]

We let $\varsigma_i \in \mathbb{C}$ denote the independent unit-power data signal intended for UE i, thus $\mathbb{E}\{|\varsigma_i|^2\} = 1$. The signal transmitted by AP l consists of a precoded superposition of the signals intended for the different UEs that this AP serves. This can be expressed as

$$\mathbf{x}_l = \sum_{i=1}^{K} \mathbf{D}_{il} \mathbf{w}_{il} \varsigma_i \tag{2.8}$$

where $\mathbf{w}_{il} \in \mathbb{C}^N$ denotes the transmit precoding vector that AP l assigns to UE i. This vector has two important impacts on the transmission: the direction $\mathbf{w}_{il}/\|\mathbf{w}_{il}\|$ determines the spatial directionality and the average squared norm $\mathbb{E}\{\|\mathbf{w}_{il}\|^2\}$ determines the average transmit power. We can define the *effective transmit precoding vector* as

$$\mathbf{D}_{il}\mathbf{w}_{il} = \begin{cases} \mathbf{w}_{il} & l \in \mathcal{M}_i \\ \mathbf{0}_N & l \notin \mathcal{M}_i \end{cases} \tag{2.9}$$

since $\mathbf{D}_{il} = \mathbf{0}_{N \times N}$ implies $\mathbf{D}_{il}\mathbf{w}_{il} = \mathbf{0}_N$. Hence, each AP only needs to select precoding vectors for the UEs that it serves.

[2] When applying the downlink results of this monograph in practical systems, we should transmit $\mathbf{x}_l^\star$ instead of $\mathbf{x}_l$, we should treat $(y_k^{\mathrm{dl}})^\star$ as the true received signal, and the true noise term will be $n_k^\star \sim \mathcal{N}_{\mathbb{C}}(0, \sigma_{\mathrm{dl}}^2)$. Hence, it is only a few conjugates that differ, while the communication performance is identical.

By substituting the transmitted signal in (2.8) into (2.7), the received downlink signal at UE k becomes

$$\begin{aligned} y_k^{\mathrm{dl}} &= \sum_{l=1}^{L} \mathbf{h}_{kl}^{\mathrm{H}} \left(\sum_{i=1}^{K} \mathbf{D}_{il} \mathbf{w}_{il} \varsigma_i \right) + n_k \\ &= \sum_{l=1}^{L} \mathbf{h}_{kl}^{\mathrm{H}} \mathbf{D}_{kl} \mathbf{w}_{kl} \varsigma_k + \sum_{\substack{i=1 \\ i \neq k}}^{K} \sum_{l=1}^{L} \mathbf{h}_{kl}^{\mathrm{H}} \mathbf{D}_{il} \mathbf{w}_{il} \varsigma_i + n_k. \end{aligned} \tag{2.10}$$

The first term in (2.10) is the desired signal, the second term is inter-user interference, and the third term is noise. The property of the effective transmit precoding vectors in (2.9) can be used to rewrite (2.10) as

$$y_k^{\mathrm{dl}} = \sum_{l \in \mathcal{M}_k} \mathbf{h}_{kl}^{\mathrm{H}} \mathbf{w}_{kl} \varsigma_k + \sum_{\substack{i=1 \\ i \neq k}}^{K} \sum_{l \in \mathcal{M}_i} \mathbf{h}_{kl}^{\mathrm{H}} \mathbf{w}_{il} \varsigma_i + n_k \tag{2.11}$$

which will not simplify but rather complicate the notation later in this monograph. We will therefore utilize the general form in (2.10) for performance analysis and keep in mind that one can always utilize (2.9) to get alternative expressions.

As in the uplink, the selection of the precoding vectors $\mathbf{w}_{kl}$ depends on the available channel estimates and the degree of cooperation among the APs. This will be thoroughly discussed in Section 6 on p. 355.

2.4 Network Scalability

Since the geographical coverage area of the network can be huge, the technology must be designed to be scalable in the sense that one can add more APs and/or UEs to the network without having to increase the capabilities of the existing ones. Conventional cellular networks achieve scalability through a divide-and-conquer approach. The coverage area is divided into cells and each cell operates autonomously. Within a cell, there is a maximum number of UEs that can be simultaneously served by spatial multiplexing (if there are more UEs, then time-frequency scheduling must be used). However, the cell operation is unaffected by the addition of new APs or UEs belonging to other cells. Therefore,

cellular technology is inherently scalable and this is also confirmed by the fact that nation-wide cellular networks exist all over the world.

The situation is more complicated in cell-free networks since all APs are somehow involved in the service of all UEs. A given AP can either be directly involved in the service of a particular UE k, in the sense of sending/receiving data to/from the UE, or indirectly involved; for example, by serving another UE together with an AP that transmits to UE k. When we deploy an AP, it will have a maximum computational capability and a maximum fronthaul capacity. These resources must remain sufficient even as the network size increases, in terms of deploying new APs, and as the network load grows, in terms of adding more UEs.

To determine if a cell-free network is scalable or not, it is instrumental to let $K \to \infty$ and see which of the following tasks remain practically implementable at each AP:

1. *Signal processing for channel estimation*;
2. *Signal processing for data reception and transmission*;
3. *Fronthaul signaling for data and CSI sharing*;
4. *Power allocation optimization.*

Based on this list, we make the following definition.

Definition 2.2 (Scalability)**.** A Cell-free Massive MIMO network is *scalable* if all the four above-listed tasks have finite computational complexity and resource requirements with respect to each AP as $K \to \infty$.

The intention with this definition is *not* that a cell-free network will serve an infinitely large number of UEs in practice, but to uncover fundamental scalability issues that become clear in the asymptotic regime. There are many algorithms that have manageable complexity for generating simulations in academic papers, but cannot be used in practice where the number of APs and UEs will be much larger than in a basic simulation. Definition 2.2 might give the impression that all the signal processing and optimization must be carried out locally at the AP, but this is not necessary. An implementation where each AP

sends the data and CSI to a CPU, which carries out those tasks, can be scalable as long as the corresponding complexity and fronthaul signaling do not grow with K.

2.4.1 Example of Unscalable Network Operation

Before explaining how to achieve scalability in a cell-free network, we will exemplify the opposite: an unscalable network operation. Suppose AP l is serving all the K UEs in the entire cell-free network, which implies that $l \in \mathcal{M}_k$, for $k = 1, \ldots, K$. The AP carries out the signal processing and optimization locally. Four potential scalability issues were mentioned in Definition 2.2 and we will go through them one after the other to explain why they are not satisfied.

Firstly, we have the complexity of the signal processing related to channel estimation. If AP l should serve K UEs, it will have to learn the channels of all these UEs. The details on channel estimation are provided in Section 4 on p. 264, but since there are K channel vectors $\mathbf{h}_{1l}, \ldots, \mathbf{h}_{Kl}$ to estimate, the computational complexity will for sure grow at least linearly with the number of UEs. Hence, the complexity approaches infinity as $K \to \infty$. Moreover, the memory size needed to store these channel estimates will be proportional to K and, thus, any finite-sized memory module will be insufficient as $K \to \infty$.

Secondly, AP l needs to create the downlink signal $\sum_{i=1}^{K} \mathbf{w}_{il}\varsigma_i$ in (2.8), where the summation implies infinite complexity as $K \to \infty$. The complexity of computing the K precoding vectors $\{\mathbf{w}_{kl} : k = 1, \ldots, K\}$ depends on the precoding scheme, but will also grow at least linearly with K since there are NK coefficients to compute, and every vector also needs to be properly scaled. The same scalability issue appears in the uplink, where the AP needs to compute $\{\mathbf{v}_{kl}^{\mathrm{H}}\mathbf{y}_l^{\mathrm{ul}} : k = 1, \ldots, K\}$ using K different combining vectors $\{\mathbf{v}_{kl} : k = 1, \ldots, K\}$.

Thirdly, AP l needs to receive the K downlink data signals $\{\varsigma_k : k = 1, \ldots, K\}$ from a CPU and must forward its K processed received signals $\{\mathbf{v}_{kl}^{\mathrm{H}}\mathbf{y}_l^{\mathrm{ul}} : k = 1, \ldots, K\}$ over the fronthaul links. The number of scalars to be sent over the fronthaul grows unboundedly as $K \to \infty$, thus the AP's fronthaul capacity will be insufficient. Moreover, the memory capacity required to temporarily store the data symbols of the K UEs at AP l will also grow unboundedly with K.

Fourthly, AP l needs to select how to allocate its transmit power between the K UEs. Any power allocation algorithm that makes use of channel knowledge related to all UEs will have a complexity that grows at least linearly with K, which is not scalable as $K \to \infty$.

This example demonstrates that the core of the scalability issue is that AP l serves all the UEs. Based on this observation, we will provide a sufficient condition for achieving a scalable implementation of cell-free networks.

2.4.2 A Sufficient Condition for Scalability

Recall that $\mathbf{D}_{il}$ is non-zero only if AP l serves UE i. Hence, the set of UEs served by AP l can be defined as

$$\mathcal{D}_l = \Big\{ i : \mathrm{tr}(\mathbf{D}_{il}) \geq 1, i \in \{1, \ldots, K\} \Big\}. \tag{2.12}$$

The following lemma presents a condition on $\mathcal{D}_l$ that partially guarantees scalability in a cell-free network.

Lemma 2.1. If the cardinality $|\mathcal{D}_l|$ remains finite as $K \to \infty$ for $l = 1, \ldots, L$, then the cell-free network satisfies the first three scalability tasks in Definition 2.2.

Proof. AP l only needs to compute the channel estimates and precoding/combining vectors for $|\mathcal{D}_l|$ UEs. This has a finite complexity as $K \to \infty$ if $|\mathcal{D}_l|$ remains finite. Moreover, AP l only needs to send/receive data related to these $|\mathcal{D}_l|$ UEs over the fronthaul links, which is a finite number as $K \to \infty$. □

The implication of Lemma 2.1 is that we need to limit the number of active UEs that each AP can serve. Ideally, we should assign APs to UEs so that every UE is served by all the surrounding APs; that is, the APs that will influence the UE's performance noticeably. At the same time, we need to make sure that $|\mathcal{D}_l|$ remains small and finite. How to determine sets $\mathcal{D}_l$ that guarantee both scalability and good service to the UEs will be explored first in Section 4.4 on p. 286 and then in further detail in Section 7.5 on p. 431. Until then, we assume that $\mathcal{D}_1, \ldots, \mathcal{D}_L$ are given.

Note that only the fourth scalability condition (regarding the complexity of the power allocation) is not guaranteed by the finite cardinality of $\mathcal{D}_l$ in Lemma 2.1. The next lemma provides the missing piece.

Lemma 2.2. Suppose every AP l selects its downlink transmit power based only on information about the $|\mathcal{D}_l|$ UEs served by that AP. Furthermore, suppose every UE is assigned a transmit power by only one of the serving APs and this AP selects that power based only on information about the $|\mathcal{D}_l|$ UEs that it serves. If Lemma 2.1 is satisfied, then the cell-free network satisfies all the scalability conditions in Definition 2.2.

Proof. AP l only needs to compute the downlink transmit power for $|\mathcal{D}_l|$ UEs using an information set that is proportional to $|\mathcal{D}_l|$, but independent of K. Furthermore, every AP will at most compute the uplink transmit powers for $|\mathcal{D}_l|$ UEs. If the condition in Lemma 2.1 is satisfied, then the fourth condition in Definition 2.2 will also be satisfied as $K \to \infty$ by following the power allocation requirements specified in this lemma. □

The implication of Lemma 2.2 is that the transmit powers should be selected in a distributed manner to achieve scalability. There exist several network-wide power allocation algorithms that jointly select the uplink or downlink powers. These can, for instance, be the solutions to linear or convex optimization problems, whose computational complexity grows polynomially in the number of optimization variables and, thus, are unscalable. Some key examples are provided in Section 7.1 on p. 394 but these are meant as benchmarks rather than as practical solutions.

Scalable power allocation algorithms should preferably be implemented separately for every AP, with no or limited interaction between the APs. It is easy to design such algorithms; for example, every UE can transmit with full power in the uplink and every AP can divide its transmit power equally between the UEs that it serves. However, it is much harder to create algorithms that operate close to the network-wide power allocation benchmarks. State-of-the-art algorithms for scalable power allocation will be reviewed in Section 7.2 on p. 410. Until then, we will assume the transmit powers are fixed and given.

Remark 2.2 (Scalability in centralized operation)**.** The sufficient condition for scalability in Lemma 2.2 states that APs make decisions related to the transmit powers. This does not mean that each decision has to be made in a local processor at an AP, but it can also be carried out at a neighboring CPU that is associated with the AP. In other words, a centralized operation where the CPUs carry out almost all the signal processing and optimization can also be scalable if designed to satisfy Definition 2.2.

2.5 Channel Modeling

A realistic performance assessment of any MIMO technology requires the use of a channel model that reflects its main characteristics. The wireless channel is deterministic in nature but is often so complicated, due to multipath propagation, that the channel variations appear as random. Hence, both deterministic and stochastic channel models can be used for analysis (cf. [**massivemimobook**]). Deterministic models can be based on ray-tracing or recorded channel measurements. The free-space line-of-sight (LoS) propagation model is another example of a deterministic model [**massivemimobook**]. The general drawback of deterministic models is that they are only valid for specific scenarios. In contrast, stochastic models are independent of a particular propagation environment and, consequently, allow for more far-reaching quantitative analysis. Nevertheless, the random distribution and its parameter values limit the scope of each stochastic model to a certain category of propagation conditions, such as urban, suburban, or rural scenarios.

A classical stochastic model for non-line-of-sight (NLoS) communications is *uncorrelated Rayleigh fading* where the channel between AP l and UE k is generated as $\mathbf{h}_{kl} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_N, \beta_{kl}\mathbf{I}_N)$. The complex Gaussian distribution accounts for the random small-scale fading caused by small-scale movements of the transmitter, receiver, or other objects in the propagation environment. More precisely, the received signal is the summation of signal copies arriving from a large number of propagation paths with seemingly random phase shifts. This gives rise to constructive and destructive interference between the paths that can be modeled by

a Gaussian distribution thanks to the central limit theorem. This fading model is called Rayleigh fading since the magnitude of each element of $\mathbf{h}_{kl}$ has a Rayleigh distribution.

When Rayleigh fading is used in a block-fading model, a new independent realization is drawn from the Gaussian distribution in each coherence block. Moreover, the variance β_{kl} describes the large-scale fading and determines the average channel quality as the UE moves around in a relatively small area. In other words, β_{kl} is determined by the geometric pathloss and shadow fading. The uncorrelated Rayleigh fading model has been (and still is) the basis of most theoretical research in multiple antenna communications since it is an analytically tractable model. However, the physical consequence of having independently and identically distributed elements in $\mathbf{h}_{kl}$ is that the AP can transmit signals in any direction and the average received power at the UE will always be the same. Measurements campaigns have repeatedly shown that this model is inadequate in practical systems since the elements of the channel vector will be statistically correlated [**Sanguinetti2019a**]. This is called *spatial correlation* and originates from two phenomena: 1) Transmissions in some spatial directions are more likely to lead to the UE than other directions; 2) The geometry of the antenna array (i.e., shape and antenna spacing) makes it more suited to transmit/receive signals in some directions than in other directions. It is only under very strict technical conditions that perfectly uncorrelated fading can be achieved [**Pizzo2020**].[3]

2.5.1 Correlated Rayleigh Fading

In this monograph, we capture the spatial correlation characteristics by making use of the relatively tractable *correlated Rayleigh fading model*. This model is suitable for NLoS channels with substantial multipath propagation so that the elements of the channel vector can be modeled as being complex Gaussian distributed. The channel between AP l and

[3]The classical example is a ULA with half-wavelength-spaced omni-directional antennas that are placed in an isotropic scattering environment [**Marzetta2016a**], where the scattering objects are uniformly distributed over all angles in three dimensions. Note that both the omni-directionality and isotropic scattering are theoretical constructs that are not practically occurring.

UE k is generated as

$$\mathbf{h}_{kl} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_N, \mathbf{R}_{kl}) \tag{2.13}$$

where $\mathbf{R}_{kl} \in \mathbb{C}^{N \times N}$ is the spatial correlation matrix between AP l and UE k. Note that this is the correlation matrix of $\mathbf{h}_{kl}$ since $\mathbb{E}\{\mathbf{h}_{kl}\mathbf{h}_{kl}^{\mathrm{H}}\} = \mathbf{R}_{kl}$. The Gaussian distribution models the small-scale fading whereas the positive semi-definite correlation matrix $\mathbf{R}_{kl}$ describes the large-scale fading, including geometric pathloss, shadowing, antenna gains, and spatial channel correlation [**massivemimobook**].

The diagonal elements of $\mathbf{R}_{kl}$ might be different but, in analogy with uncorrelated Rayleigh fading, we define the large-scale fading coefficient as the average value:

$$\beta_{kl} = \frac{1}{N}\mathrm{tr}\left(\mathbf{R}_{kl}\right). \tag{2.14}$$

This is the average channel gain between an antenna at AP l and UE k. Moreover, we assume the channel vectors of different APs are independently distributed, thus $\mathbb{E}\{\mathbf{h}_{kn}\mathbf{h}_{kl}^{\mathrm{H}}\} = \mathbf{0}_{N \times N}$ for $l \neq n$. This is a reasonable assumption whenever the APs are separated by tens of wavelengths or more, which will be an underlying assumption in this monograph. The collective channel is thus distributed as follows:

$$\mathbf{h}_k \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_M, \mathbf{R}_k) \tag{2.15}$$

where $\mathbf{R}_k = \mathrm{diag}(\mathbf{R}_{k1}, \ldots, \mathbf{R}_{kL}) \in \mathbb{C}^{M \times M}$ is the block-diagonal collective spatial correlation matrix.

The correlated Rayleigh fading model is used throughout this monograph. Furthermore, we assume the spatial correlation matrices $\mathbf{R}_{kl}$ are available wherever needed; see [**BSD16A**], [**CaireC17a**], [**NeumannJU17**], [**Sanguinetti2019a**], [**UpadhyaJU17**] for practical methods for estimating spatial correlation matrices.

2.5.2 Large-Scale Fading Model for Urban Deployments

The mathematical analysis in this monograph is applicable for arbitrary values of the large-scale fading coefficients but we will now define a model that will be used for simulations. We assume that the APs are deployed in urban environments to serve a dense population of UEs.

Moreover, the APs are deployed ten meters above the plane where the UEs are located. This matches well with the 3GPP Urban Microcell model that is defined in [**LTE2017a**]. The model is designed for the traditional 2 GHz band but is representative also for other sub-6 GHz frequency bands. We follow that model in the simulations and compute the large-scale fading coefficient (channel gain) in dB as

$$\beta_{kl}\,[\mathrm{dB}] = -30.5 - 36.7\log_{10}\left(\frac{d_{kl}}{1\,\mathrm{m}}\right) + F_{kl} \tag{2.16}$$

where d_{kl} [m] is the three-dimensional distance between AP l and UE k, taking the height difference into account. Moreover, $F_{kl} \sim \mathcal{N}(0, 4^2)$ is the shadow fading. The shadowing terms from an AP to different UEs are correlated as [**LTE2017a**]

$$\mathbb{E}\{F_{kl}F_{ij}\} = \begin{cases} 4^2 2^{-\delta_{ki}/9\,\mathrm{m}} & l = j \\ 0 & l \neq j \end{cases} \tag{2.17}$$

where δ_{ki} is the distance between UE k and UE i. The second row in (2.17) accounts for the correlation of the shadowing terms related to two different APs, which is assumed to be zero. The reason is that APs are typically distributed in the network at a distance larger than tens of meters and purposely deployed to view the deployment area from different directions than other APs. Note that it is only the simulations that rely on these specific assumptions, while all the analytical results in this monograph can be applied to any propagation environment that features Rayleigh fading.

2.5.3 Local Scattering Model for Spatial Correlation

The spatial correlation matrix depends on two main factors: the array geometry and the angular distribution of the multipath components. For small-sized APs, which are envisioned to be used in cell-free networks, it is common to utilize a ULA where the N antennas are equally spaced on a horizontal line. We will consider a ULA with half-wavelength-spacing in the simulations of this monograph. If all the multipaths arrive from the far-field of the array, the (m, ℓ)th element of a generic

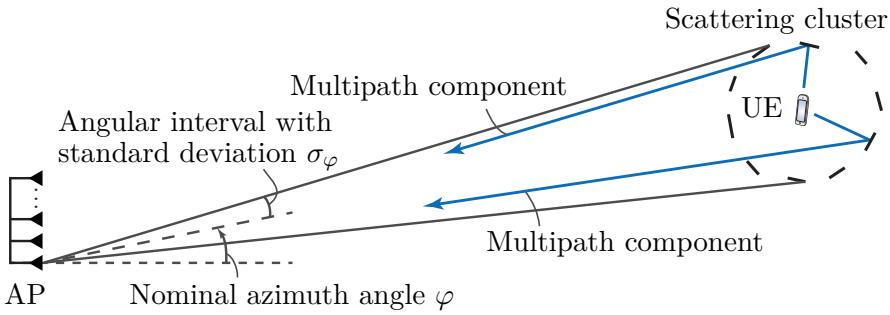


Figure 2.5: Illustration of NLoS propagation under the local scattering model, where the scattering is localized around the UE. The figure only shows the azimuth plane and two of the many multipath components are indicated. The nominal angle φ and the ASD σ_φ of the multipath components are key parameters to model the spatial correlation matrix.

spatial correlation matrix $\mathbf{R}$ can be computed as

$$[\mathbf{R}]_{m\ell} = \beta \int \int e^{\mathrm{j}\pi(m-\ell)\sin(\bar{\varphi})\cos(\bar{\theta})} f(\bar{\varphi}, \bar{\theta}) d\bar{\varphi} d\bar{\theta} \tag{2.18}$$

where β is the common large-scale fading coefficient while $\bar{\varphi}$ denotes the azimuth angle and $\bar{\theta}$ denotes the elevation angle of a multipath component, both computed with respect to the broadside of the array [**massivemimobook**]. Moreover, $f(\bar{\varphi}, \bar{\theta})$ is the joint probability density function (PDF) of $\bar{\varphi}$ and $\bar{\theta}$. Hence, the double-integral in (2.18) computes $\mathbb{E}\{e^{\mathrm{j}\pi(m-\ell)\sin(\bar{\varphi})\cos(\bar{\theta})}\}$ with respect to randomly located multipath components that are distributed according to $f(\bar{\varphi}, \bar{\theta})$.

The integrals in (2.18) can be computed numerically for any PDF. In the simulations, we will adopt the local scattering model, which is a common way to select the PDF by assuming the multipath components are symmetrically distributed around a straight line drawn between the AP and UE. The intention with this model is that there is a scattering cluster centered around the UE and the multipath components are arriving from nearby angles. The model has been utilized in numerous studies along with Gaussian [**Adachi1986a**], [**massivemimobook**], [**Trump1996a**], [**Yin2013a**], [**Zetterberg1995a**], Laplace [**jiang2015achievable**] [**Molisch2007**], [**Pedersen1997a**], and uniform [**Adhikary2013a**], [**Salz1994a**], [**Shiu2000a**], [**Yin2013a**] angular distributions. We will

consider the jointly Gaussian distribution

$$f(\bar{\varphi},\bar{\theta}) = \frac{1}{2\pi\sigma_{\varphi}\sigma_{\theta}} e^{-\frac{(\bar{\varphi}-\varphi)^2}{2\sigma_{\varphi}^2}} e^{-\frac{(\bar{\theta}-\theta)^2}{2\sigma_{\theta}^2}} \tag{2.19}$$

where φ and θ are the nominal azimuth and elevation angles, computed by drawing a straight line between the AP and UE. Note that the random variations in the azimuth and elevation angles are assumed to be independent. The standard deviations $\sigma_{\varphi} \geq 0$ and $\sigma_{\theta} \geq 0$ are called the *angular standard deviation (ASD)*. This model is illustrated from the above in Figure 2.5, where the multipath variations in the azimuth angle are visible.

Remark 2.3 (Beyond correlated Rayleigh fading)**.** The theoretical framework of this monograph applies to any scenario with correlated Rayleigh fading channels and is not limited to the local scattering model. However, not all practical channels are well modeled by Rayleigh fading since the Gaussian distribution can only be motivated when there are many propagation paths and these have pathlosses drawn from a common distribution. These conditions are not satisfied when there are few propagation paths or in LoS scenarios where the direct path is substantially stronger than all other paths. Under those circumstances, one can either consider channel models with a finite number of paths [**Femenias2019a**], [**Jin2019a**] or Rician fading where there is one strong path plus Rayleigh fading that describes all the other paths [**Alonzo2019a**], [**Demir2020b**], [**Jin2019a**], [**Ngo2018b**], [**Ozdogan2019a**], [**Zhang2020b**]. We will not cover any of these models in this monograph to keep the presentation simple, but we note that correlated Rayleigh fading can be viewed as a worst-case approximation of the aforementioned models if one uses the same spatial correlation matrices but replaces the more structured fading distribution with the Gaussian distribution [**massivemimobook**].

One way to quantify the effect of spatial channel correlation is by studying the eigenvalues of $\mathbf{R}$. Figure 2.6 shows the eigenvalues in decreasing order for an AP equipped with $N = 8$ antennas. The nominal azimuth and elevation angles are set as $\varphi = 30^{\circ}$ and $\theta = -15^{\circ}$, respectively. We consider the Gaussian local scattering model described above

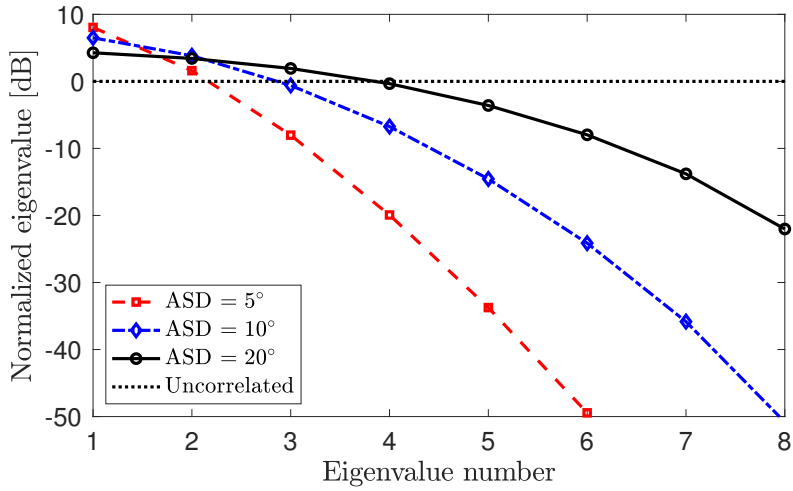


Figure 2.6: Eigenvalues of the spatial correlation matrix $\mathbf{R}$ when using the local scattering model with $N = 8$, the nominal azimuth angle $\varphi = 30°$, and the nominal elevation angle $\theta = -15°$. The correlation matrix is computed based on (2.18) using the Gaussian local scattering model in (2.19). Three different ASDs are considered: $\sigma_\varphi = \sigma_\theta \in \{5°, 10°, 20°\}$. Uncorrelated Rayleigh fading is shown as a reference case.

with three different sets of ASD values: $\sigma_\varphi = \sigma_\theta = 5°$, $\sigma_\varphi = \sigma_\theta = 10°$, and $\sigma_\varphi = \sigma_\theta = 20°$. The correlation matrices are normalized such that $\mathrm{tr}(\mathbf{R})/N = 1$. The figure also shows the reference case of uncorrelated Rayleigh fading with $\mathbf{R} = \mathbf{I}_N$ for which all eigenvalues are equal to one.

The main observation from Figure 2.6 is that the eigenvalues are widely different when the ASD is small, but the eigenvalue spread decreases with an increasing ASD. In the case of $\sigma_\varphi = \sigma_\theta = 5°$, the first four eigenvalues contribute to 99.99% of the sum of the eigenvalues. Hence, the correlation matrix is nearly rank-deficient and all channel realizations will essentially be linear combinations of the corresponding four eigenvectors. In fact, the single largest (dominant) eigenvalue contributes to 80% of the sum. Therefore, most channel realizations will point in similar directions as the dominant eigenvector. The eigenvalue variations are substantially smaller for $\sigma_\varphi = \sigma_\theta = 20°$, but there is still a 425 times difference between the largest and smallest eigenvalues. This can be compared to the extreme case of uncorrelated fading, represented

by the dotted curve, where all eigenvalues are identical.

Spatial correlation is of fundamental importance in Cellular Massive MIMO [**Sanguinetti2019a**], where the number of antennas is so large that the eigenvalue variations play a key role in mitigating interference between UEs [**BjornsonHS17**], [**Huh2012a**], [**Yin2016a**], [**Yin2013a**]. We refer to [**massivemimobook**] for more details on the basic impact that spatial correlation can have on a multi-user system. Since the number of antennas per AP is envisaged to be fairly small in Cell-free Massive MIMO, spatial correlation cannot be utilized as efficiently but we will anyway consider it throughout this monograph. In particular, we will analyze the effect of the eigenvalue spread on the channel estimation performance in Section 4 on p. 264.

2.6 Channel Hardening and Favorable Propagation

In Cellular Massive MIMO systems, two important properties appear when employing a large number of antennas at an AP: *channel hardening* and *favorable propagation*. These properties provide an insightful explanation for the performance gain of Cellular Massive MIMO and are elaborated for channels featuring correlated fading in [**massivemimobook**] and for some other types of channels in [**Marzetta2016a**]. Whether these properties appear or not depends strongly on the number of antennas and the spatial channel correlation properties.

In this section, we define and analyze channel hardening and favorable propagation in Cell-free Massive MIMO, which are rather different from the cellular case since multiple APs serve each UE. There are several different definitions in the cellular literature and we have selected those that extend most naturally to the cell-free case. We will observe that the two properties are now affected not only by the spatial correlation and the total number of antennas in the network, $M = LN$, but also by the variations in the large-scale fading coefficients among the APs and the power allocation. Recall from Section 2.5.2 that the large-scale fading is determined by the distances between the APs and UEs, as well as other macroscopic propagation effects (e.g., shadow fading). We stress that, although we dedicate a section to describe when channel hardening and favorable propagation might appear in practice,

these properties are not formally required to make use of any of the results presented in this monograph. We refer to [**Chen2018b**] for a more detailed analysis of the two properties in Cell-free Massive MIMO.

2.6.1 Channel Hardening

Each element of a channel vector takes a new independent realization in every coherence block, but these small-scale fading variations will not necessarily affect the communication performance. As can be seen in (2.10), the received downlink signal at UE k contains the desired signal component $\sum_{l=1}^{L} \mathbf{h}_{kl}^{\mathrm{H}} \mathbf{D}_{kl} \mathbf{w}_{kl} \varsigma_k$, where the *effective channel* is

$$\sum_{l=1}^{L} \mathbf{h}_{kl}^{\mathrm{H}} \mathbf{D}_{kl} \mathbf{w}_{kl} = \sum_{l \in \mathcal{M}_k} \mathbf{h}_{kl}^{\mathrm{H}} \mathbf{w}_{kl}. \tag{2.20}$$

This is the summation of the inner product between the channel vector $\mathbf{h}_{kl}$ and the precoding vector $\mathbf{D}_{kl}\mathbf{w}_{kl}$ from all the APs. The value of this effective channel can potentially remain almost constant even if the individual elements of the L channel vectors $\mathbf{h}_{k1}, \ldots, \mathbf{h}_{kL}$ are changing. More precisely, when the random realizations of the effective channel in (2.20) are close to the mean value (i.e., the variance is small), then approximately the same effective scalar channel will appear in every coherence block and we can, thus, operate the system as if we were communicating over a deterministic channel. When this happens, we say that we have achieved *channel hardening*, where the word "harden" refers to the fact that the effective channel is becoming more deterministic.

Definition and Sufficient Condition for Channel Hardening

To give a formal definition of the channel hardening property, we assume all serving APs apply MR precoding, which we define as

$$\mathbf{w}_{kl} = \sqrt{\frac{\rho_{kl}}{\mathbb{E}\{\|\mathbf{h}_{kl}\|^2\}}} \mathbf{h}_{kl} \tag{2.21}$$

with $\rho_{kl} \geq 0$ determining how much power AP l is assigning for transmission to UE k. Note that MR is the precoding method that maximizes

the received signal power at the UE. If we insert this precoding into (2.20), the effective channel becomes

$$\sum_{l \in \mathcal{M}_k} \sqrt{\frac{\rho_{kl}}{\mathbb{E}\{\|\mathbf{h}_{kl}\|^2\}}} \|\mathbf{h}_{kl}\|^2. \tag{2.22}$$

Note that the following channel hardening definition considers the combined channel from all the serving APs to a particular UE k, which is the natural extension of the definition in Cellular Massive MIMO where only a single AP serves each UE [**massivemimobook**].

Definition 2.3 (Channel hardening)**.** For a given set $\mathcal{M}_k$ of serving APs and downlink power allocation coefficients $\rho_{k1}, \ldots, \rho_{kL}$, the effective channel to UE k is said to provide asymptotic channel hardening if

$$\frac{\sum\limits_{l \in \mathcal{M}_k} \sqrt{\frac{\rho_{kl}}{\mathbb{E}\{\|\mathbf{h}_{kl}\|^2\}}} \|\mathbf{h}_{kl}\|^2}{\mathbb{E}\left\{ \sum\limits_{l \in \mathcal{M}_k} \sqrt{\frac{\rho_{kl}}{\mathbb{E}\{\|\mathbf{h}_{kl}\|^2\}}} \|\mathbf{h}_{kl}\|^2 \right\}} \to 1 \tag{2.23}$$

in the mean-squared sense as $N \to \infty$.[4]

The expectation in (2.23) is computed with respect to the independent channel realizations in different coherence blocks, while $\mathcal{M}_k$ and the power allocation coefficients are assumed to be constant.

This definition says that the effective channel in (2.22) is close to its mean value when the number of AP antennas grows large. Since the N-length channel vectors become larger as more antennas are added, the underlying assumption is that the antenna arrays are deployed according to some distribution that determines the evolution of the channel statistics (e.g., spatial correlation matrices). This is why the convergence is specified in the mean-squared sense. When considering correlated Rayleigh fading, a sufficient condition for channel hardening is as follows.

[4]Convergence in the mean-squared sense also implies convergence in probability and in distribution. The exact type of convergence is of limited importance since we are not interested in the asymptotic regime but to get a structured way to evaluate when we can obtain approximate channel hardening for a finite number of antennas.

Lemma 2.3. For correlated Rayleigh fading channels, a sufficient condition for the effective channel of UE k to provide asymptotic channel hardening is that

$$\frac{\sum\limits_{l\in\mathcal{M}_k}\rho_{kl}\frac{\mathrm{tr}(\mathbf{R}_{kl}^2)}{N\beta_{kl}}}{N\left(\sum\limits_{l\in\mathcal{M}_k}\sqrt{\rho_{kl}\beta_{kl}}\right)^2}\to 0 \quad \text{as } N\to\infty. \tag{2.24}$$

Proof. The ratio in the left-hand side of (2.23) has a mean value of one, thus we want to prove convergence to the mean value. In this case, convergence in the mean-squared sense requires that the variance of the ratio goes asymptotically to zero. The variance can be computed as

$$\mathbb{V}\left\{\frac{\sum\limits_{l\in\mathcal{M}_k}\sqrt{\frac{\rho_{kl}}{\mathbb{E}\{\|\mathbf{h}_{kl}\|^2\}}}\|\mathbf{h}_{kl}\|^2}{\mathbb{E}\left\{\sum\limits_{l\in\mathcal{M}_k}\sqrt{\frac{\rho_{kl}}{\mathbb{E}\{\|\mathbf{h}_{kl}\|^2\}}}\|\mathbf{h}_{kl}\|^2\right\}}\right\}=\frac{\sum\limits_{l\in\mathcal{M}_k}\rho_{kl}\frac{\mathrm{tr}(\mathbf{R}_{kl}^2)}{\mathrm{tr}(\mathbf{R}_{kl})}}{\left(\sum\limits_{l\in\mathcal{M}_k}\sqrt{\rho_{kl}\mathrm{tr}(\mathbf{R}_{kl})}\right)^2} \tag{2.25}$$

by utilizing the statistical independence of the channels to the different APs and by applying [**massivemimobook**] to compute the expectations. We then obtain the sufficient condition in (2.24) by utilizing the definition of β_{kl} in (2.14), which implies that $\mathrm{tr}(\mathbf{R}_{kl})=N\beta_{kl}$. □

Although the definition of channel hardening relies on asymptotic arguments, we can use it to evaluate the *degree of channel hardening* for setups with a finite number of antennas. If the expression in (2.24) is close to zero, this means that the left-hand side of (2.23) is close to one and thus channel hardening is approximately achieved.

It has been shown in the Cellular Massive MIMO literature that spatial correlation reduces the convergence rate to channel hardening [**massivemimobook**]; that is, more antennas are needed to achieve a certain degree of channel hardening. This can be also observed in (2.24) where $\mathrm{tr}(\mathbf{R}_{kl}^2)$ can be computed as the sum of the squared eigenvalues of the spatial correlation matrix $\mathbf{R}_{kl}$. Recall from Figure 2.6 that spatial correlation is represented by large eigenvalue variations. Since the sum of the eigenvalues equals $\mathrm{tr}(\mathbf{R}_{kl})=N\beta_{kl}$ by definition, the squared sum of the eigenvalues can take values between $N\beta_{kl}^2$ (when all eigenvalues

equal to β_{kl}) and $N^2\beta_{kl}^2$ (when there is only one non-zero eigenvalue that equals $N\beta_{kl}$). The latter represents an extreme case of high spatial correlation. To achieve a high degree of channel hardening, we want the expression in (2.24) to be small and the case when all eigenvalues are the same (i.e., no spatial correlation) achieves the smallest value.

Impact of the Number and Geographical Distribution of APs

In the context of Cell-free Massive MIMO, where each AP is supposed to have a relatively small number of antennas but there will be many APs instead, it is interesting to evaluate if one can compensate for having a small N by having many APs. To analyze this further, we assume that $\mathbf{R}_{kl} = \beta_{kl}\mathbf{I}_N$ so that there is no spatial correlation. The expression in (2.24) then simplifies to

$$\frac{\sum\limits_{l \in \mathcal{M}_k} \rho_{kl}\beta_{kl}}{N\left(\sum\limits_{l \in \mathcal{M}_k} \sqrt{\rho_{kl}\beta_{kl}}\right)^2}. \tag{2.26}$$

If $|\mathcal{M}_k| = 1$, or one AP has a much larger value on $\rho_{kl}\beta_{kl}$ than the other APs that serve UE k, then (2.26) becomes approximately $1/N$. If we instead consider the case when all the serving APs have the same value of $\rho_{kl}\beta_{kl}$, then (2.26) becomes $1/(N|\mathcal{M}_k|)$. Hence, it is plausible to achieve the same degree of channel hardening by having many APs with few antennas as with one AP with many antennas in some cases. If $\rho_{kl}\beta_{kl}$ is the same for all the serving APs, it is only the total number of AP antennas that determines the degree of channel hardening. However, when the APs have different values of $\rho_{kl}\beta_{kl}$, then the total number of antennas needed to achieve a certain degree of channel hardening will likely be larger than in Cellular Massive MIMO.

The geographical distribution of the APs, with respect to the UE, determines the values of the large-scale fading coefficients $\{\beta_{kl}\}$. There can naturally be tens of dB differences between the serving APs, even if they have rather similar distances to the UE. One can compensate for this by applying a power allocation algorithm where the powers $\{\rho_{kl}\}$ are selected to even out the differences. More precisely, we can reduce

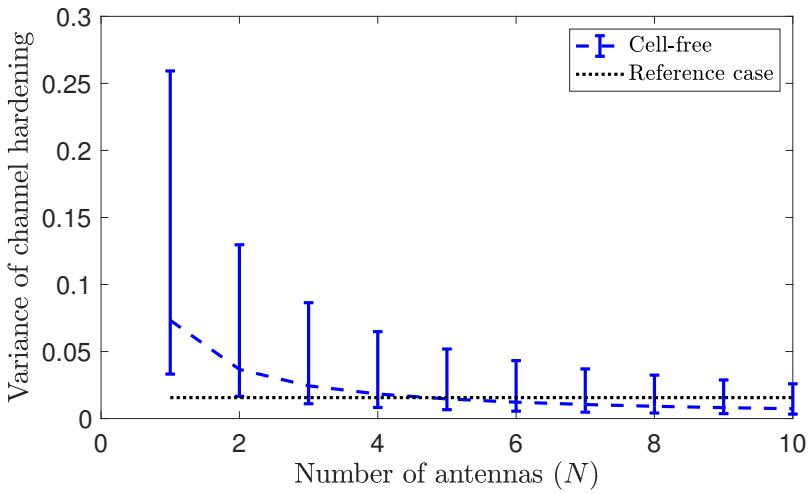


Figure 2.7: The value of (2.26), which represents the variance of the ratio between the effective channel and its mean value, is plotted for different N. When (2.26) is small, there is a high degree of channel hardening. The reference case corresponds to a 64-antenna Cellular Massive MIMO setup. The cell-free setup consists of 64 randomly distributed APs and a varying number of antennas per AP. The line shows the median value while the bars show the interval containing 90% of the random realizations.

the power from nearby APs (with larger β_{kl}) and increase the power from APs that are further away (with smaller β_{kl}) to achieve roughly the same value of $\rho_{kl}\beta_{kl}$ for all the serving APs. This will increase the degree of channel hardening but will not necessarily be desirable from an end-to-end performance perspective since we can achieve a higher total received power at the UE by allocating the same power in the opposite way (more power from the nearby APs).

Figure 2.7 exemplifies how the degree of channel hardening depends on the geographical AP distribution and the number of antennas per AP. We consider the channel to a UE in the center of 400 m $\times$ 400 m area. $L = 64$ APs are uniformly distributed in the area and transmit with equal power to the UE. We use the large-scale fading model in (2.16) and assume independent and identically distributed (i.i.d.) Rayleigh fading. The figure shows the value of the expression in (2.26) for different values of N. Since the value depends on the AP distribution, the dashed line

shows the median value while the vertical bars indicate the interval in which 90% of all realizations occur. As a reference, the dotted horizontal line shows the value of (2.26) for a single AP with 64 antennas, which represents a Cellular Massive MIMO setup that would achieve a high degree of channel hardening. If single-antenna APs are used, then the degree of channel hardening is far from that of the reference case, even if the total number of antennas is the same. The median reaches the reference case for $N = 5$ but half of the realizations give substantially larger values. It is first when the number of antennas per AP reaches 8 or 10 that the degree of channel hardening is comparable to that of Cellular Massive MIMO. Note that the total number of antennas is then an order of magnitude larger than in the reference case.

Takeaways

In summary, channel hardening is a desirable property but not mandatory for the operation of Cell-free Massive MIMO. In fact, we can expect a substantially lower degree of channel hardening in cell-free networks than in cellular networks when the number of serving antennas is the same. Channel hardening makes some capacity lower bounds (that we will present later in this monograph) closer to the true capacity and also serves as a motivation for not trying to optimize the power allocation in every coherence block but rather on average. We will return to these things later in this monograph.

Note that we defined channel hardening from a downlink precoding perspective in this section. The same argumentation can be made also in the uplink, in which case the precoding vectors are replaced by combining vectors, and the power allocation coefficients are replaced by weights that the CPU uses when fusing the signals $\hat{s}_{kl}$ from different APs. The details on this will be provided in Section 5 on p. 294.

2.6.2 Favorable Propagation

When multiple UEs are spatially multiplexed in the system, there will generally be interference between their transmissions, and the precoding/combining can be selected to strike a balance between achieving

strong desired signals and causing little interference. However, if the UEs' channels are spatially orthogonal, then the inter-user interference is automatically mitigated by MR processing so one can basically ignore it. When this happens, we say that *favorable propagation* is experienced.

If we consider the received downlink signal of UE k in (2.10), then the interference caused by the concurrent transmission to UE i is $\sum_{l=1}^{L} \mathbf{h}_{kl}^{\mathrm{H}} \mathbf{D}_{il} \mathbf{w}_{il \varsigma i}$. If the MR precoding defined in (2.21) is utilized, then the *effective interfering channel* is

$$\sum_{l=1}^{L} \sqrt{\frac{\rho_{il}}{\mathbb{E}\{\|\mathbf{h}_{il}\|^2\}}} \mathbf{h}_{kl}^{\mathrm{H}} \mathbf{D}_{il} \mathbf{h}_{il} = \sum_{l \in \mathcal{M}_i} \sqrt{\frac{\rho_{il}}{\mathbb{E}\{\|\mathbf{h}_{il}\|^2\}}} \mathbf{h}_{kl}^{\mathrm{H}} \mathbf{h}_{il}. \tag{2.27}$$

It is the magnitude of this term compared to the magnitude of the effective channel of the desired signal in (2.22) that determines how strong the interference is in relative terms. If the ratio is close to zero, then the interference will be negligible and we benefit from favorable propagation.

Definition and Sufficient Condition for Favorable Propagation

There are several different formal definitions of favorable propagation in the Cellular Massive MIMO literature but none of them is a perfect fit for the case when multiple APs are transmitting coherently. Hence, we provide the following new definition that is tailored to the cell-free scenario.

Definition 2.4 (Favorable propagation)**.** A given UE k experiences asymptotically favorable propagation with respect to UE i (with $i \neq k$) if

$$\frac{\sum\limits_{l \in \mathcal{M}_i} \sqrt{\frac{\rho_{il}}{\mathbb{E}\{\|\mathbf{h}_{il}\|^2\}}} \mathbf{h}_{kl}^{\mathrm{H}} \mathbf{h}_{il}}{\mathbb{E}\left\{ \sum\limits_{l \in \mathcal{M}_k} \sqrt{\frac{\rho_{kl}}{\mathbb{E}\{\|\mathbf{h}_{kl}\|^2\}}} \|\mathbf{h}_{kl}\|^2 \right\}} \to 0 \tag{2.28}$$

in the mean-squared sense as $N \to \infty$.

This definition says that the effective interfering channel in (2.27) (i.e., a weighted sum of the inner products between the channel vectors of the two UEs) should go asymptotically to zero, when it is normalized

by the average value of the desired effective channel in (2.22). The weights are the downlink transmit powers. Note that the numerator depends on the set $\mathcal{M}_i$ of APs that serve the interfering UE i, while the denominator depends on the set $\mathcal{M}_k$ of APs serving the considered UE k. For correlated Rayleigh fading channels, a sufficient condition for favorable propagation is obtained as follows.

Lemma 2.4. For correlated Rayleigh fading channels, a sufficient condition for UE k to experience asymptotically favorable propagation with respect to UE i is that

$$\frac{\sum\limits_{l \in \mathcal{M}_i} \rho_{il} \frac{\mathrm{tr}(\mathbf{R}_{il}\mathbf{R}_{kl})}{N\beta_{il}}}{N\left(\sum\limits_{l \in \mathcal{M}_k} \sqrt{\rho_{kl}\beta_{kl}} \right)^2} \to 0 \quad \text{as } N \to \infty. \tag{2.29}$$

Proof. The ratio in the left hand side of (2.28) has a mean value of zero, thus we want to prove convergence to the mean value. In this case, convergence in mean-squared sense requires that the variance of the ratio goes asymptotically to zero. The variance can be computed as

$$\mathbb{V}\left\{ \frac{\sum\limits_{l \in \mathcal{M}_i} \sqrt{\frac{\rho_{il}}{\mathbb{E}\{\|\mathbf{h}_{il}\|^2\}}} \mathbf{h}_{kl}^{\mathrm{H}} \mathbf{h}_{il}}{\mathbb{E}\left\{ \sum\limits_{l \in \mathcal{M}_k} \sqrt{\frac{\rho_{kl}}{\mathbb{E}\{\|\mathbf{h}_{kl}\|^2\}}} \|\mathbf{h}_{kl}\|^2 \right\}} \right\} = \frac{\sum\limits_{l \in \mathcal{M}_i} \rho_{il} \frac{\mathrm{tr}(\mathbf{R}_{il}\mathbf{R}_{kl})}{\mathrm{tr}(\mathbf{R}_{il})}}{\left(\sum\limits_{l \in \mathcal{M}_k} \sqrt{\rho_{kl}\mathrm{tr}(\mathbf{R}_{kl})} \right)^2} \tag{2.30}$$

by utilizing the statistical independence of the channels to the different APs and by applying [**massivemimobook**] to compute the expectations. We then obtain the sufficient condition in (2.29) by utilizing the definition of β_{kl} in (2.14), which implies that $\mathrm{tr}(\mathbf{R}_{kl}) = N\beta_{kl}$. □

Although the definition of favorable propagation relies on asymptotic arguments, we can use it to evaluate the degree of favorable propagation for setups with a finite number of antennas. If the expression in (2.29) is close to zero, this means that the left-hand side of (2.28) is close to zero and thus favorable propagation is approximately achieved.

The value of the expression in (2.29) depends on the spatial correlation properties in a nontrivial manner. The term $\mathrm{tr}(\mathbf{R}_{il}\mathbf{R}_{kl})$ in the

numerator takes its largest value when the spatial correlation matrices $\mathbf{R}_{il}$ and $\mathbf{R}_{kl}$ are identical up to a scaling factor, which represents the case when the two UEs have almost the same spatial directivity of their channels (statistically speaking). In contrast, the term takes its smallest value when the spatial correlation matrices have identical eigenvectors but the eigenvalues are matched together in the opposite order (i.e., large eigenvalues from one matrix are multiplied with small eigenvalues from the other matrix). This represents the case when the two UEs have very different spatial directivity. Clearly, it is the latter case that leads to the highest degree of favorable propagation.

Impact of the Geographical Distribution of APs

To instead focus on the different large-scale fading variations that the UEs experience, we can assume that $\mathbf{R}_{kl} = \beta_{kl}\mathbf{I}_N$ so that there is no spatial correlation. The expression in (2.29) then simplifies to

$$\frac{\sum\limits_{l \in \mathcal{M}_i} \rho_{il}\beta_{kl}}{N\left(\sum\limits_{l \in \mathcal{M}_k} \sqrt{\rho_{kl}\beta_{kl}}\right)^2}. \tag{2.31}$$

This expression decays as $1/N$ so more antennas lead to a higher degree of favorable propagation. Moreover, the set $\mathcal{M}_i$ of APs that serve UE i plays an important role in determining the value of (2.31). If those APs are far from UE k, all the β_{kl} terms in the numerator will likely be very small and then we can achieve favorable propagation even with $N = 1$. However, if $\mathcal{M}_i = \mathcal{M}_k$, the numerator and denominator might be of comparable size and then we might need many antennas per AP to achieve favorable propagation. The power allocation also plays a key role; to make the term $\rho_{il}\beta_{kl}$ in the numerator small, we can either have a small value of β_{kl} and/or use a low transmit power ρ_{il} on that AP.

Figure 2.8 exemplifies how the degree of favorable propagation depends on the geographical AP distribution, the distance between the UEs, and the number of antennas per AP. We consider the same setup as in Figure 2.7 but add an interfering UE that is either 10 m or 100 m to the east of the desired UE. We assume each UE is served by the 8

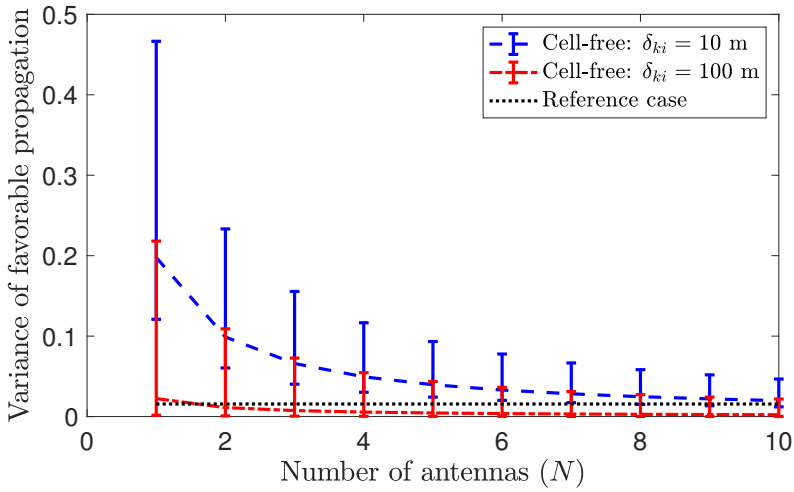


Figure 2.8: The value of (2.31), which represents the variance of the ratio between the effective interfering channel and the mean of the desired effective channel, is plotted for different N. When (2.31) is small for a selected UE pair, there is a high degree of favorable propagation between them. The reference case corresponds to a 64-antenna Cellular Massive MIMO setup. The cell-free setup consists of 64 randomly distributed APs and a varying number of antennas per AP. The UEs are either 10 or 100 m apart, and are each served by the 8 APs providing the best channels. The line shows the median value while the bars show the interval containing 90% of the random realizations.

APs that provide the eight largest large-scale fading coefficients and these APs transmit with equal power. The figure shows the value of the expression in (2.31) for different N. Since the value depends on the AP distribution, the dashed line shows the median value while the vertical bars indicate the interval in which 90% of all realizations occur. As a reference, the dotted horizontal line shows the value of (2.31) when a single AP with 64 antennas serves both UEs, which represents a Cellular Massive MIMO setup. We notice that the distance between the UEs strongly determines the degree of favorable propagation. When the UEs are close together, we need around $N = 10$ antennas per AP before favorable propagation is approximately achieved. When the UEs are 100 m apart, most realizations of the AP locations lead to favorable propagation even for $N = 1$, but there are still large variations. We can

expect that, in practice, some UEs that are far apart might still not achieve favorable propagation with respect to each other.

Takeaways

In summary, favorable propagation is a desirable property that makes inter-user interference disappear when using MR precoding (which does not actively suppress any interference). However, favorable propagation is not a necessary property. We have observed that UEs that are served by partially the same APs might not experience it with respect to each other. In those cases, the interference can instead be suppressed by using precoding methods that are actively suppressing interference. This is covered in detail in Section 6 on p. 355.

Note that we defined favorable propagation from a downlink precoding perspective in this section. The same argumentation can be made also in the uplink, as previously discussed in relation to channel hardening. When there is a lack of favorable propagation in the uplink, we can use receive combining to deal with the interference instead; see Section 5 on p. 294. Although there are clear mathematical similarities between channel hardening and favorable propagation, the properties are different. In particular, a UE might exhibit channel hardening but not favorable propagation or vice versa. It is also likely that some pairs of UEs will experience favorable propagation with respect to each other, while other pairs will not.

2.7 Summary of the Key Points in Section 2

- A Cell-free Massive MIMO network consists of L APs, each equipped with N antennas, that are arbitrarily distributed over the coverage area. The APs may jointly serve K single-antenna UEs in each channel coherence block. More precisely, each UE is communicating with a subset of the APs, which is selected based on the UE's needs.
- By having a total number of $M = LN$ AP antennas much larger than K, the cell-free network typically has sufficiently many spatial degrees-of-freedom to separate UEs in space by linearly processing the transmitted and received signals.
- Since a cell-free network may have a large geographical coverage area, it must be designed to be scalable in the sense that the computational capability and fronthaul capacity of existing APs must remain sufficient as new APs are deployed and as more UEs are being served. A cell-free network in which each AP serves all the UEs is not scalable.
- Although the number of co-located antennas per AP is envisaged to be fairly small, the spatial correlation must be anyway considered in the analysis since the channel fading is spatially correlated in practice.
- When employing a large number of co-located antennas at an AP, two important properties appear: channel hardening and favorable propagation. In cell-free networks, the antennas at multiple APs contribute to these properties. However, a lower degree of channel hardening is expected than in cellular networks because of the fairly small number of antennas per AP. Cell-free networks are expected to provide a high degree of favorable propagation between UEs that are relatively far apart, but not among those that are closely spaced.

3

Theoretical Foundations

This section describes some basic results from estimation, information, and optimization theory which will be utilized later in this monograph. Section 3.1 describes the foundations of estimating Gaussian distributed random variables, which might be the channels in a cell-free network. Section 3.2 defines the channel capacity and describes standard methods for computing achievable lower bounds, which are also known as achievable spectral efficiencies (SEs). In Section 3.3, we consider the maximization of a ratio known as a generalized Rayleigh quotient. In Section 3.4, two different utility maximization algorithms are provided, which are of particular interest for the optimization of UE performance in a cell-free network. The key points are summarized in Section 3.5.

3.1 Estimation Theory for Gaussian Variables

The estimation of channel coefficients is an important aspect of any coherent communication system. In this section, we provide the theoretical foundation for estimating the realization of a Gaussian random variable from an observation that is corrupted by independent additive Gaussian noise. We only present the main concepts and results that will be used in later chapters and refer to textbooks such as [**massivemimobook**], [**Kay1993a**] for further details and explanations.

Definition 3.1 (Minimum mean-squared error estimator)**.** Consider a random variable $\mathbf{x} \in \mathbb{C}^N$ with support in $\mathcal{X}$ and let $\hat{\mathbf{x}}(\mathbf{y})$ denote an arbitrary estimator of $\mathbf{x}$ based on the observation $\mathbf{y} \in \mathbb{C}^M$. The particular choice of $\hat{\mathbf{x}}(\mathbf{y}) : \mathbb{C}^M \to \mathbb{C}^N$ that minimizes the mean-squared error (MSE)

$$\mathbb{E}\left\{\|\mathbf{x} - \hat{\mathbf{x}}(\mathbf{y})\|^2\right\} \tag{3.1}$$

is called the MMSE estimator of $\mathbf{x}$. It can be computed as

$$\hat{\mathbf{x}}_{\text{MMSE}}(\mathbf{y}) = \mathbb{E}\{\mathbf{x}|\mathbf{y}\} = \int_{\mathcal{X}} \mathbf{x} f(\mathbf{x}|\mathbf{y}) d\mathbf{x} \tag{3.2}$$

where $f(\mathbf{x}|\mathbf{y})$ is the conditional PDF of $\mathbf{x}$ given the observation $\mathbf{y}$.

The MMSE estimator of a complex Gaussian random variable from an observation that is corrupted by independent additive complex Gaussian noise (and interference) can be computed in closed form.

Lemma 3.1. Consider estimation of the N-dimensional vector $\mathbf{x} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_N, \mathbf{R})$, with a positive semi-definite correlation matrix $\mathbf{R}$, from the observation $\mathbf{y} = \mathbf{x}q + \mathbf{n} \in \mathbb{C}^N$. The pilot signal $q \in \mathbb{C}$ is known and $\mathbf{n} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_N, \mathbf{S})$ is an independent noise/interference vector with a positive definite correlation matrix.

The MMSE estimator of $\mathbf{x}$ is

$$\hat{\mathbf{x}}_{\text{MMSE}}(\mathbf{y}) = q^\star \mathbf{R} \left(|q|^2 \mathbf{R} + \mathbf{S}\right)^{-1} \mathbf{y}. \tag{3.3}$$

The estimation error correlation matrix is

$$\mathbf{C}_{\text{MMSE}} = \mathbf{R} - |q|^2 \mathbf{R} \left(|q|^2 \mathbf{R} + \mathbf{S}\right)^{-1} \mathbf{R} \tag{3.4}$$

and the MSE is

$$\mathsf{MSE} = \text{tr}\left(\mathbf{R} - |q|^2 \mathbf{R} \left(|q|^2 \mathbf{R} + \mathbf{S}\right)^{-1} \mathbf{R}\right). \tag{3.5}$$

Proof. The detailed proof can be found in [**massivemimobook**]. □

For brevity, we will denote the MMSE estimate as $\hat{\mathbf{x}}_{\text{MMSE}}$, without explicitly specifying what observation it was based on. A consequence of Lemma 3.1 is that the MMSE estimate is distributed as

$$\hat{\mathbf{x}}_{\text{MMSE}} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_N, \mathbf{R} - \mathbf{C}_{\text{MMSE}}). \tag{3.6}$$

Moreover, the estimation error $\tilde{\mathbf{x}} = \mathbf{x} - \hat{\mathbf{x}}$ is distributed as

$$\tilde{\mathbf{x}}_{\text{MMSE}} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_N, \mathbf{C}_{\text{MMSE}}). \tag{3.7}$$

These random variables are independent, which is a useful property that will be utilized in the analysis in future sections.

3.2 Capacity Bounds and Spectral Efficiency

The SE is the performance metric that will be used throughout this monograph. This is a measure of the average amount of information that can be correctly transferred per complex-valued sample, when a very large block of data is transmitted. This is the natural metric in broadband applications, where the demand for data is large.

Definition 3.2 (Spectral efficiency)**.** The achievable SE of an encoding/decoding scheme is the average number of bits of information, per complex-valued sample, that it can transmit reliably over the channel under consideration.

When communicating over a bandwidth of B Hz, the Nyquist-Shannon sampling theorem specifies that the communication signal is fully determined by B complex-valued samples per second [**Shannon1949b**] (or $2B$ real-valued samples per second). These are the samples that the data is encoded into in communications and the SE describes the amount of data that can be transferred per such complex sample. Since there are B samples per second, the unit of the SE is either bit per complex sample or bit per second per Hertz, which is abbreviated as bit/s/Hz. We will use the latter unit in the remainder of this monograph. A related metric is the *information rate* [bit/s], which is defined as the product of the SE and the bandwidth B.

The channel between a given transmitter and receiver supports many different SEs (depending on the chosen encoding/decoding scheme), but the largest achievable SE is of key importance when designing communication systems. The maximum SE is determined by the so-called *channel capacity*. We refer to the original paper [**Shannon1948a**] by Shannon and textbooks on information theory, such as [**Cover1991a**], for the general and formal definition. In wireless communications, we

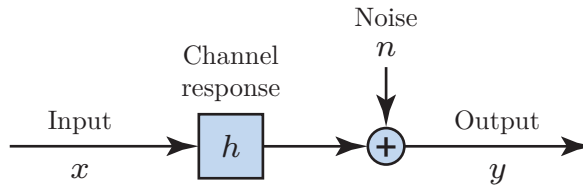


Figure 3.1: A discrete memoryless channel with input x and output $y = hx + n$, where h is the channel response and n is independent Gaussian noise.

are particularly interested in channels where the received signal is a superposition of a scaled version of the desired signal and additive Gaussian noise. These channels are commonly referred to as discrete memoryless additive white Gaussian noise (AWGN) channels, since each discrete input signal leads to a discrete output signal that is independent of previous and future inputs. We will provide exact capacity expressions and capacity lower bounds for two distinctly different cases: deterministic and random channels. Although the channels are modeled as random in this monograph, both cases will be utilized.

3.2.1 Capacity Bounds for Deterministic Channels

We begin with the canonical case when the channel is deterministic and the communication is only disturbed by Gaussian distributed noise. A block diagram of this setup is provided in Figure 3.1.

Lemma 3.2. Consider a discrete memoryless AWGN channel with input $x \in \mathbb{C}$ and output $y \in \mathbb{C}$ given by

$$y = hx + n \tag{3.8}$$

where $n \sim \mathcal{N}_{\mathbb{C}}(0, \sigma^2)$ is independent noise. The input distribution is power-limited as $\mathbb{E}\{|x|^2\} \leq p$ and the channel response $h \in \mathbb{C}$ is deterministic and known at the output.

In this case, any SE smaller or equal to the channel capacity

$$C = \log_2\left(1 + \frac{p|h|^2}{\sigma^2}\right) \tag{3.9}$$

is achievable. The capacity is achieved by selecting $x \sim \mathcal{N}_{\mathbb{C}}(0, p)$.

Proof. The detailed proof can be found in [**Shannon1949b**]. □

The practical meaning of the channel capacity is that for any SE smaller or equal to the capacity, there exists an encoding/decoding scheme such that an arbitrarily low error probability can be achieved. More precisely, if we consider an information sequence with N scalar inputs to the AWGN channel, then the probability of error goes to zero as $N \to \infty$. This means that an infinite decoding delay is required to achieve the capacity. However, in practice, one can operate very close to the capacity whenever blocks of thousands of bits are transmitted [**Bjornson2016b**], which is typically the case in broadband applications.

The channel considered in Lemma 3.2 is called a single-input single-output (SISO) channel because one input signal is sent and results in one output signal. However, we will demonstrate in later sections how the same result can be utilized in situations with multiple antennas. The capacity expression in (3.9) has a form that is typical for communications: the base-two logarithm of one plus the SNR defined as

$$\mathsf{SNR} = \frac{\overbrace{p|h|^2}^{\text{Received signal power}}}{\underbrace{\sigma^2}_{\text{Noise power}}}. \tag{3.10}$$

In the setups considered in this monograph, the communication is also disturbed by interference from other concurrent transmissions. In this case, the exact capacity is generally unknown, but convenient lower bounds can be obtained [**Biglieri1998a**]. The following lemma provides a lower bound on the capacity that will be used repeatedly in this monograph.

Lemma 3.3. Consider a discrete memoryless interference channel with input $x \in \mathbb{C}$ and output $y \in \mathbb{C}$ given by

$$y = hx + v + n \tag{3.11}$$

where $n \sim \mathcal{N}_{\mathbb{C}}(0, \sigma^2)$ is independent noise and $v \in \mathbb{C}$ is random interference with an arbitrary distribution that has zero mean, variance p_v, and is uncorrelated with the input (i.e., $\mathbb{E}\{x^\star v\} = 0$).

The input distribution is power-limited as $\mathbb{E}\{|x|^2\} \leq p$ and the channel response $h \in \mathbb{C}$ is deterministic and known at the output. The channel capacity C is then lower bounded as

$$C \geq \log_2\left(1 + \frac{p|h|^2}{p_\upsilon + \sigma^2}\right) \tag{3.12}$$

where the bound is achieved using the input distribution $x \sim \mathcal{N}_{\mathbb{C}}(0, p)$.

Proof. The detailed proof can be found in [**massivemimobook**]. □

The lower bound on the capacity in (3.12) is achieved by encoding/decoding the signal as if the interference υ is independent Gaussian noise, because this is the worst case from a communication perspective [**Hassibi2003a**]. There are no approximations involved but the achievable SE is generally smaller than the capacity, particularly when there are very strong interfering signals. However, treating non-Gaussian interference as independent Gaussian noise in the encoding/decoding is practically convenient and provably optimal in the low-interference regime [**Annapureddy2009a**], [**Annapureddy2011a**], [**Motahari2009a**], [**Shang2011b**], [**Shang2009b**]. When we use the SE expression from Lemma 3.3 in later sections, the random interference term υ will not only include interference from other concurrent transmissions but also unknown random variations in the desired channel.

The SE expression in (3.12) has a similar form as the capacity expression in Lemma 3.2. It is the base-two logarithm of one plus the expression

$$\mathsf{SINR} = \frac{\overbrace{p|h|^2}^{\text{Received signal power}}}{\underbrace{p_\upsilon}_{\text{Interference power}} + \underbrace{\sigma^2}_{\text{Noise power}}} \tag{3.13}$$

which is an SINR. We will refer to it as an *effective SINR*, which means that the lower bound in (3.12) is effectively the same as the capacity of an AWGN channel with an SNR equal to SINR in (3.13). This implies that the SE in (3.12) can be practically achieved using channel codes designed for AWGN channels.

3.2.2 Capacity Bounds for Random Channels

We consider randomly fading channels in this monograph. If the channel is a random variable that takes a new independent realization after a finite block of complex samples (e.g., a coherence block), then another capacity concept can be defined: the *ergodic capacity*. The transmission then spans asymptotically many realizations of the random variable that describes the channel and the word "ergodic" identifies that all the statistical properties of the channel are deducible from a single sequence of channel realizations. We begin with the canonical case when the communication is only disturbed by Gaussian distributed noise.

Lemma 3.4. Consider a discrete memoryless channel with input $x \in \mathbb{C}$ and output $y \in \mathbb{C}$ given by

$$y = hx + n \tag{3.14}$$

where $n \sim \mathcal{N}_{\mathbb{C}}(0, \sigma^2)$ is independent noise and the input distribution is power-limited as $\mathbb{E}\{|x|^2\} \leq p$. The channel h is a realization of a random variable $\mathbb{H}$ that is independent of the signal and noise, and the realization $\mathbb{H} = h$ is known at the receiver.

The ergodic channel capacity is

$$C = \mathbb{E}\left\{\log_2\left(1 + \frac{p|h|^2}{\sigma^2}\right)\right\} \tag{3.15}$$

where the expectation is with respect to h. The capacity is achieved by selecting the input distribution $x \sim \mathcal{N}_{\mathbb{C}}(0, p)$.

Proof. The detailed proof can be found in [**massivemimobook**]. □

The ergodic capacity of the block-fading channel is similar to the capacity in (3.9), but the key difference is that there is an expectation in front of the logarithm in (3.15). In this case, $\mathsf{SNR} = \frac{p|h|^2}{\sigma^2}$ is called the *instantaneous SNR*. We can also define the average SNR as

$$\mathsf{SNR} = \frac{p\mathbb{E}\{|h|^2\}}{\sigma^2} \tag{3.16}$$

where the expectation is computed with respect to the channel realizations.

As mentioned above, the communication will be disturbed by interference from other concurrent transmissions in this monograph. In this case, the exact ergodic capacity is generally unknown, but a convenient lower bound can be obtained by once again treating the interference as noise [**Biglieri1998a**], [**Medard2000a**]. The following lemma provides a lower bound on the ergodic capacity that will be used repeatedly in this monograph.

Lemma 3.5. Consider a discrete memoryless interference channel with input $x \in \mathbb{C}$ and output $y \in \mathbb{C}$ given by

$$y = hx + v + n \tag{3.17}$$

where $n \sim \mathcal{N}_{\mathbb{C}}(0, \sigma^2)$ is independent noise, the channel response $h \in \mathbb{C}$ is known at the output, and $v \in \mathbb{C}$ is random interference with an arbitrary distribution satisfying the properties listed below. The input is power-limited as $\mathbb{E}\{|x|^2\} \leq p$.

Suppose $h \in \mathbb{C}$ is a realization of the random variable $\mathbb{H}$ and that $\mathbb{U}$ is a random variable with realization u that affects the interference variance. The realizations of these random variables are known at the output. If the noise n is conditionally independent of v given h and u, the interference v has conditional zero mean (i.e., $\mathbb{E}\{v|h,u\} = 0$) and conditional variance denoted by $p_v(h,u) = \mathbb{E}\{|v|^2|h,u\}$, and the interference is conditionally uncorrelated with the input (i.e., $\mathbb{E}\{x^\star v|h,u\} = 0$), then the ergodic channel capacity C is lower bounded as

$$C \geq \mathbb{E}\left\{\log_2\left(1 + \frac{p|h|^2}{p_v(h,u) + \sigma^2}\right)\right\} \tag{3.18}$$

where the expectation is taken with respect to h and u, and the bound is achieved using the input distribution $x \sim \mathcal{N}_{\mathbb{C}}(0, p)$.

Proof. The detailed proof can be found in [**massivemimobook**]. □

Note that in Lemma 3.5, we used the shorthand notation $\mathbb{E}\{v|h,u\}$ for the conditional expectation $\mathbb{E}\{v|\mathbb{H} = h, \mathbb{U} = u\}$. For notational convenience, we omit the random variables in similar expressions in the remainder of this monograph and only write out the realizations.

The lower bound on the ergodic capacity in (3.18) is an achievable SE and will often be called that in this monograph. Except for the list of technical conditions that must be satisfied, the SE in (3.18) is essentially the same as for the case of deterministic channels in (3.12), but with an expectation in front of the logarithm. The expression $\frac{p|h|^2}{p_\upsilon(h,u)+\sigma^2}$ in (3.18) will be called the *effective instantaneous SINR*, because it takes the same role as the instantaneous SNR in the ergodic capacity expression in (3.15). In other words, the capacity lower bound is equivalent to the ergodic capacity of a fading AWGN channel where the instantaneous SNR is $\frac{p|h|^2}{p_\upsilon(h,u)+\sigma^2}$. This implies that the SE in (3.18) can be practically achieved using channel codes designed for fading AWGN channels.

When we use the SE expression result in Lemma 3.5 in later sections, the random interference term υ will not only include interference from other concurrent transmissions but also channel uncertainty due to estimation errors regarding the desired channel. In those cases, h represents an estimate of the desired channel and u represents the estimates of the interfering channels.

3.3 Maximization of Rayleigh Quotients

A commonly appearing problem formulation in multi-antenna communications is that of maximizing a power ratio between the desired signal and interference plus noise. The optimization variable is a vector determining how the signal observations made at the different receive antennas are weighted together. For example, consider the received signal $\mathbf{y} \in \mathbb{C}^N$ at N antennas. It is modeled as

$$\mathbf{y} = \mathbf{h}x + \mathbf{n} \tag{3.19}$$

where $\mathbf{h} \in \mathbb{C}^N$ is a deterministic channel vector, x is the random information signal with power $\mathbb{E}\{|x|^2\} = 1$, and $\mathbf{n} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_N, \mathbf{I}_N)$ is independent noise. If we apply a receive combining vector $\mathbf{v}$ to (3.19), we get

$$\mathbf{v}^{\mathrm{H}}\mathbf{y} = \mathbf{v}^{\mathrm{H}}\mathbf{h}x + \mathbf{v}^{\mathrm{H}}\mathbf{n} \tag{3.20}$$

and the SNR becomes

$$\frac{\mathbb{E}\{|\mathbf{v}^{\mathrm{H}}\mathbf{h}x|^2\}}{\mathbb{E}\{|\mathbf{v}^{\mathrm{H}}\mathbf{n}|^2\}} = \frac{|\mathbf{v}^{\mathrm{H}}\mathbf{h}|^2}{\mathbf{v}^{\mathrm{H}}\mathbf{v}}. \tag{3.21}$$

The type of expression in the right-hand side of (3.21) is known as a *Rayleigh quotient*[1] and can be maximized with respect to $\mathbf{v}$ as follows.

Lemma 3.6. For a given channel vector $\mathbf{h} \in \mathbb{C}^N$, it holds that

$$\max_{\mathbf{v}\in\mathbb{C}^N} \frac{|\mathbf{v}^{\mathrm{H}}\mathbf{h}|^2}{\mathbf{v}^{\mathrm{H}}\mathbf{v}} = \|\mathbf{h}\|^2 \tag{3.22}$$

where the maximum is attained by $\mathbf{v} = \mathbf{h}$.

Proof. The Cauchy-Schwarz inequality for two vectors $\mathbf{v}$ and $\mathbf{h}$ says that

$$|\mathbf{v}^{\mathrm{H}}\mathbf{h}|^2 \leq \|\mathbf{v}\|^2 \|\mathbf{h}\|^2 \tag{3.23}$$

with equality if and only if $\mathbf{v}$ and $\mathbf{h}$ are linearly dependent. It follows that $\frac{|\mathbf{v}^{\mathrm{H}}\mathbf{h}|^2}{\|\mathbf{v}\|^2} \leq \|\mathbf{h}\|^2$ where the maximum is attained for $\mathbf{v} = \mathbf{h}$ (among other solutions where the vectors are linearly dependent). □

The result in Lemma 3.6 is commonly used in signal processing and the optimal vector $\mathbf{v}$ is known as MR combining or matched filtering. This method maximizes the SNR, however, it is a generalization that will be particularly useful in this monograph. Let us generalize (3.19) by replacing the i.i.d. noise term $\mathbf{n}$ with the colored noise term $\mathbf{n}' \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_N, \mathbf{B})$ having a positive definite covariance matrix $\mathbf{B} \in \mathbb{C}^{N\times N}$. The new received signal is

$$\mathbf{y} = \mathbf{h}x + \mathbf{n}' \tag{3.24}$$

and by applying a receive combining vector $\mathbf{v}$ to (3.24), the SNR becomes

$$\frac{\mathbb{E}\{|\mathbf{v}^{\mathrm{H}}\mathbf{h}x|^2\}}{\mathbb{E}\{|\mathbf{v}^{\mathrm{H}}\mathbf{n}'|^2\}} = \frac{|\mathbf{v}^{\mathrm{H}}\mathbf{h}|^2}{\mathbf{v}^{\mathrm{H}}\mathbf{B}\mathbf{v}}. \tag{3.25}$$

This type of expression is known as a *generalized Rayleigh quotient* and can be maximized with respect to $\mathbf{v}$ as follows.

Lemma 3.7. For a given channel vector $\mathbf{h} \in \mathbb{C}^N$ and Hermitian positive definite matrix $\mathbf{B} \in \mathbb{C}^{N\times N}$, it holds that

$$\max_{\mathbf{v}\in\mathbb{C}^N} \frac{|\mathbf{v}^{\mathrm{H}}\mathbf{h}|^2}{\mathbf{v}^{\mathrm{H}}\mathbf{B}\mathbf{v}} = \mathbf{h}^{\mathrm{H}}\mathbf{B}^{-1}\mathbf{h} \tag{3.26}$$

[1]The expression in (3.21) is also known as the Rayleigh–Ritz ratio. Note that the only connection between Rayleigh fading and Rayleigh quotient is that they have been named after the same researcher.

where the maximum is attained by $\mathbf{v} = \mathbf{B}^{-1}\mathbf{h}$ or $\mathbf{v} = (\mathbf{h}\mathbf{h}^{\mathrm{H}} + \mathbf{B})^{-1}\mathbf{h}$.

Proof. The matrix square root $\mathbf{B}^{\frac{1}{2}}$ of $\mathbf{B}$ exists since $\mathbf{B}$ is positive definite, thus we can define the new variable $\bar{\mathbf{v}} = \mathbf{B}^{\frac{1}{2}}\mathbf{v}$ and establish that

$$\frac{|\mathbf{v}^{\mathrm{H}}\mathbf{h}|^2}{\mathbf{v}^{\mathrm{H}}\mathbf{B}\mathbf{v}} = \frac{|\bar{\mathbf{v}}^{\mathrm{H}}\mathbf{B}^{-\frac{1}{2}}\mathbf{h}|^2}{\bar{\mathbf{v}}^{\mathrm{H}}\bar{\mathbf{v}}}. \tag{3.27}$$

The expression at the right-hand side is a Rayleigh quotient of the kind in Lemma 3.6 and, thus, maximized by $\bar{\mathbf{v}} = \mathbf{B}^{-\frac{1}{2}}\mathbf{h}$. Hence, the solution to (3.26) is $\mathbf{v} = \mathbf{B}^{-1}\mathbf{h}$ and the maximum value follows accordingly. Since any vector that is parallel to $\mathbf{B}^{-1}\mathbf{h}$ also achieves the maximum value, we can also use $\mathbf{v} = (\mathbf{h}\mathbf{h}^{\mathrm{H}} + \mathbf{B})^{-1}\mathbf{h}$ since

$$(\mathbf{h}\mathbf{h}^{\mathrm{H}} + \mathbf{B})^{-1}\mathbf{h} = \frac{1}{1 + \mathbf{h}^{\mathrm{H}}\mathbf{B}^{-1}\mathbf{h}}\mathbf{B}^{-1}\mathbf{h} \tag{3.28}$$

according to the matrix-inversion lemma (Lemma B.1 on p. 445). □

This lemma shows how to maximize a generalized Rayleigh quotient and what the maximum value is. The solution to (3.26) is not unique but $\mathbf{v} = c\mathbf{B}^{-1}\mathbf{h}$ for any $c \neq 0$ will also maximize the expression. This lemma will be utilized several times in Section 5 on p. 294 to optimize different kinds of receiver processing in the uplink.

3.4 Optimization Algorithms for Utility Maximization

A utility function is a metric of system performance and can be formulated in different ways. In this section, we provide state-of-the-art optimization algorithms for solving two main types of utility maximization problems. These algorithms will be utilized in Section 7 on p. 393 to solve several resource allocation problems.

The SE is used in this monograph to measure the communication performance of each UE. When designing the network operation, there are K different SEs to take into consideration and these are mutually conflicting due to interference. For example, we can improve the SE of one UE by reducing the transmit powers assigned to other UEs, but with the side-effect that the SEs achieved by the other UEs are then reduced. To identify a good tradeoff between the UEs' individual

performance, a scalar-valued utility function can be defined to take all the SEs into consideration. This is called scalarization in the field of multi-objective optimization [**Bjornson2014c**]. We will consider two utilities: max-min fairness and sum SE.

Since we have not provided any explicit SE expressions for User-centric Cell-free Massive MIMO systems yet, the presentation in this section is based on a generic SISO scenario. Suppose the received signal used for decoding the signal of UE k can be expressed as

$$y_k = \mathrm{DS}_k(\mathbf{p})\, s_k + \sum_{i=1}^{K} \mathrm{I}_{ki}(\mathbf{p})\, s_i + n_k \tag{3.29}$$

where $s_k \in \mathbb{C}$ denotes the normalized, independent random data signal of UE k with $\mathbb{E}\{|s_k|^2\} = 1$ and $\mathbf{p} = [p_1 \ \ldots \ p_K]^{\mathrm{T}}$ is the set of transmit power coefficients for all the K UEs.[2] The transmit powers are non-negative: $p_k \geq 0$, for $k = 1, \ldots, K$. We write this as $\mathbf{p} \geq \mathbf{0}_K$.

The received signal in (3.29) contains three parts. The term $\mathrm{DS}_k(\mathbf{p})s_k$ is the desired part, containing the data signal s_k multiplied with a deterministic amplitude $\mathrm{DS}_k(\mathbf{p})$ that is a function of $\mathbf{p}$ (it usually only depends on p_k). The interference term $\sum_{i=1}^{K} \mathrm{I}_{ki}(\mathbf{p})s_i$ contains the interfering data signal s_i of UE i, for $i = 1, \ldots, K$, and the term $\mathrm{I}_{ki}(\mathbf{p})$ is random and determines the strength of the interference caused to UE k by UE i. This term is also a function of $\mathbf{p}$. Note that we have included a self-interference term $\mathrm{I}_{kk}(\mathbf{p})s_k$ with a random amplitude $\mathrm{I}_{kk}(\mathbf{p})$, which we further assume to have zero mean. This term can model the uncertainty that the receiver has regarding the channel that the desired signal propagated over, which will be utilized and further explained in later sections. Finally, $n_k \sim \mathcal{N}_{\mathbb{C}}(0, \sigma^2)$ is the independent additive noise.

We obtain the following result by using Lemma 3.3.

Lemma 3.8. An achievable SE of the channel for UE k in (3.29) is

$$\mathrm{SE}_k(\mathbf{p}) = \log_2(1 + \mathrm{SINR}_k(\mathbf{p})) \tag{3.30}$$

[2] As we see in Section 6 on p. 355, there can be more than one transmit power coefficient per UE in the downlink. Although here we assume the length of the vector $\mathbf{p}$ is K, we can simply change it for the mentioned downlink power allocation optimization.

where the effective SINR is

$$\mathrm{SINR}_k(\mathbf{p}) = \frac{|\mathrm{DS}_k(\mathbf{p})|^2}{\sum_{i=1}^{K} \mathbb{E}\left\{|\mathrm{I}_{ki}(\mathbf{p})|^2\right\} + \sigma^2}. \tag{3.31}$$

Proof. We obtain this result from Lemma 3.3 by letting $y = y_k$ be the received signal in the lemma, $x = s_k$ is the desired signal, $h = \mathrm{DS}_k(\mathbf{p})$ is the channel response, and $n = n_k$ is the noise. The interference term $\upsilon = \sum_{i=1}^{K} \mathrm{I}_{ki}(\mathbf{p}) s_i$ is random with zero mean and variance $p_\upsilon = \sum_{i=1}^{K} \mathbb{E}\{|\mathrm{I}_{ki}(\mathbf{p})|^2\}$ since the data signals have zero mean, unit variance, and are independent. The condition $\mathbb{E}\{x^\star \upsilon\} = 0$ is satisfied since $\mathrm{I}_{kk}(\mathbf{p})$ was assumed to have zero mean. □

With the notation provided by Lemma 3.8, the individual performance of the UEs are represented by their K SEs $\{\mathrm{SE}_k(\mathbf{p}) : k = 1, \ldots, K\}$, which are K different performance metrics. For each choice of the transmit power vector $\mathbf{p}$, the K UEs will achieve a certain combination of SEs that can be gathered in a K-dimensional vector $[\mathrm{SE}_1(\mathbf{p}) \ \ldots \ \mathrm{SE}_K(\mathbf{p})]^{\mathrm{T}}$. By considering the set of all such vectors that can be obtained with feasible transmit power vectors (i.e., those satisfying a set of power constraints that exist in the system), we can generate a K-dimensional region that represents all the possible ways of operating the system. This region is called an SE region and is exemplified in Figure 3.2 for $K = 2$. The shaded region contains all the feasible operating points. We need to select one of these points and there is no truly optimal way to do it. All the points in the interior of the region are strictly suboptimal since there are points at the outer boundary where all the UEs achieve higher SEs. On the other hand, each point on the outer boundary represents one possible tradeoff between the performance achieved by the different UEs. When comparing two points on the outer boundary, one UE will prefer the first point and another UE will prefer the second point. This conflict cannot be resolved in a fully objective manner but the network operator must inject a subjective opinion of what kind of operating point is preferred by the network as a whole.

As mentioned at the beginning of this section, we will adopt the scalarization approach to identify a tradeoff between the K metrics

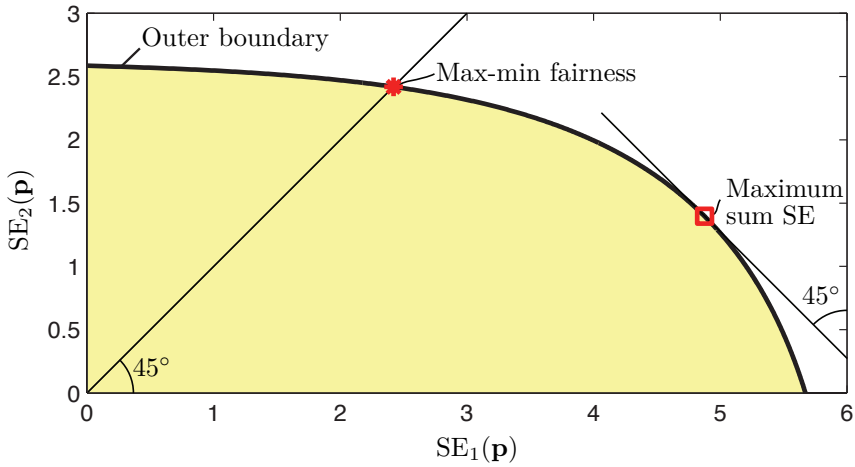


Figure 3.2: Example of the SE region (shaded) containing all the points $(\mathrm{SE}_1(\mathbf{p}), \mathrm{SE}_2(\mathbf{p}))$ that can be obtained by different feasible selections of the power coefficient vector $\mathbf{p}$. The points at the outer boundary that give max-min fairness and maximum sum SE are indicated.

[**Bjornson2014c**]. This means that we combine the K SE metrics into a single scalar utility function representing the network-wide performance. We consider the two most common utility functions from the literature: max-min fairness and sum SE. These utilities represent two extremes in how to balance between transmitting many bits and obtaining user fairness. The max-min fairness utility requires every UE to get the same SE, which means that the optimal operating point in Figure 3.2 lies on a line from the origin with 45° slope. As indicated in the figure, the max-min fairness solution is the point where this line intersects the outer boundary. The maximum sum SE is generally obtained at another operating point, as illustrated in Figure 3.2. It lies on a line with equation $\mathrm{SE}_1(\mathbf{p}) + \mathrm{SE}_2(\mathbf{p}) = c$, where c is the maximum achievable sum SE. While this point maximizes the number of bits that are transmitted in total, the downside is that the bits are unequally distributed among the UEs. In this example, UE 2 gets a smaller SE than at the max-min fairness point, while UE 1 gets a larger SE. In the remainder of this section, we will describe algorithms for finding these operating points in networks with arbitrarily many UEs.

3.4.1 Max-Min SE Fairness

The aim of max-min SE fairness is to maximize the lowest SE among all the UEs in the network, which is known as max-min fairness. Since the SE of UE k in (3.30) is an increasing function of the effective SINR, $\mathrm{SINR}_k(\mathbf{p})$, maximizing the minimum SE is the same as maximizing the minimum effective SINR among all the UEs. We assume there are R linear constraints on the transmit power vector:

$$\mathbf{a}_r^{\mathrm{T}}\mathbf{p} \leq p_{\max}, \quad r = 1, \ldots, R, \tag{3.32}$$

where the fixed vector $\mathbf{a}_r \in \mathbb{R}^K_{\geq 0}$ specifies the weighting coefficient for each UE's power coefficient and $p_{\max}$ is the maximum allowed power,[3] the max-min fairness problem can be formulated as

$$\begin{aligned} \underset{\mathbf{p}\geq \mathbf{0}_K}{\text{maximize}} \quad & \min_{k\in\{1,\ldots,K\}} \mathrm{SINR}_k(\mathbf{p}) \\ \text{subject to} \quad & \mathbf{a}_r^{\mathrm{T}}\mathbf{p} \leq p_{\max}, \quad r = 1, \ldots, R. \end{aligned} \tag{3.33}$$

To solve this problem, we first introduce the auxiliary variable t that represents the lowest SINR among all the UEs. We can then obtain the following optimization problem that is equivalent to (3.33) in terms of having the same maximum utility value and optimal value of $\mathbf{p}$:

$$\begin{aligned} \underset{\mathbf{p}\geq \mathbf{0}_K, t\geq 0}{\text{maximize}} \quad & t \\ \text{subject to} \quad & \mathrm{SINR}_k(\mathbf{p}) \geq t, \quad k = 1, \ldots, K \\ & \mathbf{a}_r^{\mathrm{T}}\mathbf{p} \leq p_{\max}, \quad r = 1, \ldots, R. \end{aligned} \tag{3.34}$$

The equivalence of the two problems can be verified by noting that t is at most equal to the lowest SINR among the UEs and it should be equal to that value to maximize the objective function. The reformulation in (3.34) is known as the epigraph form of (3.33) [**Boyd2004a**].

An optimal solution to the problem in (3.34) can be obtained by a bisection search over t. To see this, let t^{opt} denote the optimal objective

[3] This model can be utilized for both downlink and uplink power allocations. In the downlink, per-AP power constraints are considered, which leads to L power constraints (one per AP). In this case, the non-zero elements of $\mathbf{a}_l$ are multiplied with the power coefficients for the UEs that are served by AP l. On the other hand, in the uplink, there are K individual power constraints (one per UE). Hence, only the kth element of the vector $\mathbf{a}_k$ is non-zero while all other elements are zero.

value of the problem in (3.34). If we select $t > t^{\text{opt}}$ in (3.34), the problem will obviously be infeasible. On the other hand, if we select any $t < t^{\text{opt}}$, the problem in (3.34) is feasible. Hence, we can apply a bisection search over t to find t^{opt} [**Boyd2004a**]. At each iteration, we consider a value $t = t^{\text{candidate}}$ and need to solve the following feasibility problem:

$$\begin{aligned} &\text{find} \quad \mathbf{p} \\ &\text{subject to} \quad \text{SINR}_k(\mathbf{p}) \geq t^{\text{candidate}}, \quad k = 1, \ldots, K \\ &\qquad\qquad \mathbf{a}_r^{\text{T}}\mathbf{p} \leq p_{\max}, \quad r = 1, \ldots, R \\ &\qquad\qquad \mathbf{p} \geq \mathbf{0}_K. \end{aligned} \tag{3.35}$$

The goal of solving the feasibility problem in (3.35) is to find any solution $\mathbf{p}$ that satisfies the constraints. The numerical precision that is needed to find the value of $t^{\text{candidate}}$ that corresponds to the global optimum can be extremely high, particularly if some variables in $\mathbf{p}$ have a much smaller impact on the SINR of the UE with the worst channel conditions than the other variables. One can prove by contradiction that all UEs should get identical SINRs at the global optimum but this is only one of the many feasible solutions to the feasibility problem in (3.35) for any $t^{\text{candidate}} < t^{\text{opt}}$. We have noticed experimentally that this is a real issue when optimizing large cell-free networks, where some of the serving APs have much weaker channels than other serving APs. To improve the convergence rate, we will therefore replace the feasibility problem with a problem having the same constraints but where the total power $\sum_{k=1}^{K} p_k$ is minimized as well. This will encourage numerical solvers to identify the feasible point where all UEs get the same SINR, while retaining optimality. By utilizing this property, we obtain Algorithm 3.1 where the revised subproblem is given in (3.36).

Instead of solving a sequence of optimization problems, as in Algorithm 3.1, a fixed-point algorithm can be implemented when the SINR expression satisfies certain additional conditions stated in the next lemma from [**Hong2014**].

Algorithm 3.1 Bisection search algorithm for solving the max-min fairness problem in (3.34).

Initialization: Set the solution accuracy $\epsilon > 0$
Set the initial lower and upper bounds for the max-min SINR:
$t^{\text{lower}} \leftarrow 0$
$t^{\text{upper}} \leftarrow \min\limits_{k\in\{1,\ldots,K\}} \max\limits_{\mathbf{p}\succeq\mathbf{0}_K} \text{SINR}_k(\mathbf{p})$
Initialize solution variables: $\mathbf{p}^{\text{opt}} = \mathbf{0}_K$, $t^{\text{opt}} = 0$
while $t^{\text{upper}} - t^{\text{lower}} > \epsilon$ **do**
    $t^{\text{candidate}} \leftarrow \frac{t^{\text{lower}} + t^{\text{upper}}}{2}$
    Solve the following problem for fixed $t = t^{\text{candidate}}$:

$$\begin{aligned} \underset{\mathbf{p}\succeq\mathbf{0}_K}{\text{minimize}} \quad & \sum_{k=1}^{K} p_k \\ \text{subject to} \quad & \text{SINR}_k(\mathbf{p}) \geq t^{\text{candidate}}, \quad k = 1,\ldots,K \\ & \mathbf{a}_r^{\mathrm{T}}\mathbf{p} \leq p_{\max}, \quad r = 1,\ldots,R \\ & \mathbf{p} \succeq \mathbf{0}_K \end{aligned} \tag{3.36}$$

    if (3.36) is feasible **then**
        $t^{\text{lower}} \leftarrow t^{\text{candidate}}$
        $\mathbf{p}^{\text{opt}} \leftarrow \mathbf{p}$, which is the solution to (3.36)
    else
        $t^{\text{upper}} \leftarrow t^{\text{candidate}}$
    end if
end while
Output: $\mathbf{p}^{\text{opt}}$, $t^{\text{opt}} = \min\limits_{k\in\{1,\ldots,K\}} \text{SINR}_k(\mathbf{p}^{\text{opt}})$

Lemma 3.9. Suppose $\{\mathrm{SINR}_k(\mathbf{p}) : k = 1, \ldots, K\}$ satisfy the following:

1. $\mathrm{SINR}_k(\mathbf{p}) > 0$ if $\mathbf{p} > \mathbf{0}_K$ and $\mathrm{SINR}_k(\mathbf{p}) = 0$ if and only if $p_k = 0$, $\forall k$;
2. $\mathrm{SINR}_k(\mathbf{p})$ is strictly increasing with respect to p_k and is strictly decreasing with respect to p_i, for $i \neq k$, when $p_k > 0$, $\forall k$;
3. For $\lambda > 1$ and $p_k > 0$, $\mathrm{SINR}_k(\lambda\mathbf{p}) > \mathrm{SINR}_k(\mathbf{p})$, $\forall k$.

The optimal solution to the optimization problem in (3.34) then satisfies $t^{\mathrm{opt}} > 0$ and $\mathbf{p}^{\mathrm{opt}} > \mathbf{0}_K$. Moreover, $\mathrm{SINR}_k(\mathbf{p}^{\mathrm{opt}}) = t^{\mathrm{opt}}$, for all k and $\mathbf{a}_r^{\mathrm{T}}\mathbf{p}^{\mathrm{opt}} = p_{\max}$, for at least one r.

Define $\mathbf{T}(\mathbf{p}) = \left[p_1/\mathrm{SINR}_1(\mathbf{p}) \ \ldots \ p_K/\mathrm{SINR}_K(\mathbf{p})\right]^{\mathrm{T}}$ and let $\mathcal{U}$ denote the set of feasible $\mathbf{p}$ that satisfies this last property:

$$\mathcal{U} = \{\mathbf{p} \geq \mathbf{0}_K : \mathbf{a}_r^{\mathrm{T}}\mathbf{p} \leq p_{\max}, \quad r = 1, \ldots, R, \\ \text{with at least one equality for some } r\}. \tag{3.37}$$

The power vector $\mathbf{p}$ computed by Algorithm 3.2 converges to the optimal solution to (3.34) if the following additional conditions are satisfied:

- There exist numbers $a > 0$, $b > 0$, and a vector $\mathbf{e} > \mathbf{0}_K$ such that $a\mathbf{e} \leq \mathbf{T}(\mathbf{p}) \leq b\mathbf{e}$, for all $\mathbf{p} \in \mathcal{U}$;
- For any $\mathbf{p}_1, \mathbf{p}_2 \in \mathcal{U}$ and $0 \leq \lambda \leq 1$: $\lambda\mathbf{p}_1 \leq \mathbf{p}_2 \implies \lambda\mathbf{T}(\mathbf{p}_1) \leq \mathbf{T}(\mathbf{p}_2)$;
- For any $\mathbf{p}_1, \mathbf{p}_2 \in \mathcal{U}$ and $0 \leq \lambda < 1$: $\lambda\mathbf{p}_1 \leq \mathbf{p}_2$ and $\lambda\mathbf{p}_1 \neq \mathbf{p}_2 \implies \lambda\mathbf{T}(\mathbf{p}_1) < \mathbf{T}(\mathbf{p}_2)$.

Whenever the conditions stated in Lemma 3.9 are satisfied, the fixed-point algorithm in Algorithm 3.2 should be utilized to solve the max-min fairness problem since it fully utilizes the SINR structure.

3.4.2 Sum SE Maximization

A potential drawback of the max-min SE fairness problem is that all the emphasis is put on the UEs with the worst channel conditions. When having a large network where every UE only causes interference to a

Algorithm 3.2 Fixed-point algorithm for solving the max-min fairness problem in (3.34).

1: **Initialization:** Set arbitrary $\mathbf{p} > \mathbf{0}_K$ and the solution accuracy $\epsilon > 0$
2: **while** $\max\limits_{k\in\{1,\ldots,K\}} \mathrm{SINR}_k(\mathbf{p}) - \min\limits_{k\in\{1,\ldots,K\}} \mathrm{SINR}_k(\mathbf{p}) > \epsilon$ **do**
3: $\quad p_k \leftarrow \frac{p_k}{\mathrm{SINR}_k(\mathbf{p})}$, $k = 1, \ldots, K$
4: $\quad \mathbf{p} \leftarrow \frac{p_{\max}}{\max\limits_{r\in\{1,\ldots,R\}} \mathbf{a}_r^{\mathrm{T}}\mathbf{p}}\mathbf{p}$
5: **end while**
6: **Output:** $\mathbf{p}$, $t = \min\limits_{k\in\{1,\ldots,K\}} \mathrm{SINR}_k(\mathbf{p})$

small subset of neighboring UEs, it is likely that many UEs can achieve substantially larger SEs while barely affecting the UEs having the worst conditions. In these situations, it might be better to maximize the sum SE, which represents the total number of bits that are transmitted without considering how the bits are assigned between the UEs. The sum SE maximization problem can be expressed as

$$\begin{aligned} \underset{\mathbf{p}\geq\mathbf{0}_K}{\text{maximize}} \quad & \sum_{k=1}^{K} \log_2\left(1 + \mathrm{SINR}_k(\mathbf{p})\right) \\ \text{subject to} \quad & \mathbf{a}_r^{\mathrm{T}}\mathbf{p} \leq p_{\max}, \quad r = 1, \ldots, R. \end{aligned} \tag{3.38}$$

Note that the above problem is usually not convex and, hence, it is hard to obtain the optimal solution [**Luo2008a**]. There exist global optimization methods [**Bjornson2013d**], [**Weeraddana2012a**] that find the optimum but their computational complexity is unsuitable for real-time applications. A pragmatic solution is to instead settle for a local optimum. A common approach for finding a local optimum to sum SE maximization problems is the *weighted MMSE* method [**Shi2011**], which results in iterative algorithms.

To obtain the weighted MMSE reformulation of the sum SE maximization problem at hand, we recall the SISO channel for UE k in (3.29) where the received signal is y_k. The receiver can compute an estimate $\widehat{s}_k = u_k^\star y_k$ of the desired signal s_k using a scalar combining coefficient $u_k \in \mathbb{C}$ that can amplify it and rotate the phase. The corresponding

MSE is given by

$$e_k(\mathbf{p}, u_k) = \mathbb{E}\left\{|\widehat{s}_k - s_k|^2\right\}$$
$$= |u_k|^2 \left(|\mathrm{DS}_k(\mathbf{p})|^2 + \sum_{i=1}^{K} \mathbb{E}\left\{|\mathrm{I}_{ki}(\mathbf{p})|^2\right\} + \sigma^2\right) - 2\Re\left(u_k^\star \mathrm{DS}_k(\mathbf{p})\right) + 1 \tag{3.39}$$

and it is a convex function of u_k. One can show that the value of u_k that minimizes the MSE for UE k in (3.39) for a given $\mathbf{p}$ is

$$u_k(\mathbf{p}) = \frac{\mathrm{DS}_k(\mathbf{p})}{|\mathrm{DS}_k(\mathbf{p})|^2 + \sum_{i=1}^{K} \mathbb{E}\left\{|\mathrm{I}_{ki}(\mathbf{p})|^2\right\} + \sigma^2}. \tag{3.40}$$

After plugging $u_k(\mathbf{p})$ into (3.39), it follows that the corresponding e_k is equal to $1/(1 + \mathrm{SINR}_k(\mathbf{p}))$.

The main idea of the weighted MMSE method is to introduce the auxiliary weight $d_k \geq 0$ for the MSE e_k and attempt to solve the following optimization problem:

$$\begin{aligned} \underset{\mathbf{p} \geq \mathbf{0}_K, \{u_k,\ d_k \geq 0:\ k=1,\ldots,K\}}{\text{minimize}} \quad & \sum_{k=1}^{K} \Big(d_k e_k(\mathbf{p}, u_k) - \ln(d_k)\Big) \\ \text{subject to} \quad & \mathbf{a}_r^{\mathrm{T}} \mathbf{p} \leq p_{\max}, \quad r = 1, \ldots, R. \end{aligned} \tag{3.41}$$

The problem is equivalent to the original sum SE maximization problem in (3.38) in the sense that they have the same global optimal solution. The equivalence of the two problems follows from the fact that the optimal d_k in (3.41) is $1/e_k$, which is equal to $1 + \mathrm{SINR}_k$. Hence, we obtain the original sum SE maximization problem in (3.38) with the same constraints. The benefit of the reformulation in (3.41) is that we have the following result that is adapted from [**Shi2011**].

Lemma 3.10. The block coordinate descent algorithm given in Algorithm 3.3 converges to a local optimum (stationary point) of the problem in (3.41) by alternating between optimizing three blocks of variables: $\{u_k : k = 1, \ldots, K\}$, $\{d_k : k = 1, \ldots, K\}$, and $\mathbf{p}$. The obtained $\mathbf{p}$ is also a local optimum to the problem in (3.38).

Algorithm 3.3 Block coordinate descent algorithm for solving the sum SE maximization problem in (3.41).

1: **Initialization:** Set the solution accuracy $\epsilon > 0$
2: Set an arbitrary feasible $\mathbf{p}$
3: **while** the objective function is either improved more than ϵ or not improved at all **do**
4: $\quad u_k \leftarrow \dfrac{\mathrm{DS}_k(\mathbf{p})}{|\mathrm{DS}_k(\mathbf{p})|^2+\sum\limits_{i=1}^{K}\mathbb{E}\{|\mathrm{I}_{ki}(\mathbf{p})|^2\}+\sigma^2}, \quad k=1,\ldots,K$
5: $\quad d_k \leftarrow 1/e_k\left(\mathbf{p},u_k\right), \quad k=1,\ldots,K$
6: $\quad$ Solve the following problem for the current values of u_k and d_k:

$$\begin{aligned} \underset{\mathbf{p}\geq\mathbf{0}_K}{\text{minimize}} \quad & \sum_{k=1}^{K} d_k e_k\left(\mathbf{p},u_k\right) \\ \text{subject to} \quad & \mathbf{a}_r^{\mathrm{T}}\mathbf{p} \leq p_{\max}, \quad r=1,\ldots,R \end{aligned} \tag{3.42}$$

7: $\quad$ Update $\mathbf{p}$ by the obtained solution to (3.42)
8: **end while**
9: **Output:** $\mathbf{p}$

When implementing Algorithm 3.3, the subproblem in (3.42) is required to be solved optimally with respect to $\mathbf{p}$ (when the other variables are kept constant), otherwise, the block coordinate descent algorithm will not converge to a local optimum. Fortunately, the problem in (3.42) is convex for the power allocation problems we consider in Section 7 on p. 393 and the optimal solution can, thus, be obtained by any of the standard algorithms from convex optimization theory [**Boyd2004a**]. We elaborate further on this in Remark 7.2 on p. 403.

3.5 Summary of the Key Points in Section 3

- The realization of a complex Gaussian distributed random variable can be estimated using the MMSE estimator presented in Lemma 3.1. It will be used in later sections for channel estimation.
- The channel capacity is a suitable performance metric in broadband applications and will be utilized throughout this monograph. The exact capacity is hard to compute in situations with interference or limited channel knowledge, but there are lower bounds that can be utilized and achieved in practice. Lemma 3.3 presents a lower bound for deterministic channels and Lemma 3.5 presents a lower bound for fading channels. Both will be utilized in later sections.
- A lower bound on the capacity is called an achievable SE if there is a known way to achieve it. The capacity lower bounds presented in this section will be called SEs in later sections and are achievable using channel codes designed for AWGN channels.
- A commonly appearing problem in multi-antenna communications is that of maximizing a power ratio between the desired signal and interference plus noise. This problem can be formalized as a generalized Rayleigh quotient. Lemma 3.7 shows how to maximize it and what its maximum value is.
- The K SEs of the UEs can be combined into a single scalar utility function that can be maximized, for example, with respect to the transmit power coefficients. Two examples of utility functions are max-min fairness and sum SE. The corresponding utility maximization problems are solved by Algorithm 3.1 and Algorithm 3.3, respectively.

4

Channel Estimation

This section describes how channel estimation is carried out in User-centric Cell-free Massive MIMO systems based on the transmission of uplink pilots. The system model during pilot transmission is defined in Section 4.1. The MMSE channel estimation scheme is presented in Section 4.2. The impact of the cell-free architecture, interference from pilot-sharing UEs, and spatial correlation in the estimation process is evaluated in Section 4.3. A basic algorithm for pilot assignment and dynamic cooperation cluster formation is described in Section 4.4. The key points are summarized in Section 4.5.

4.1 Uplink Pilot Transmission

We consider the cell-free network defined in Section 2 on p. 203 where L APs, each equipped with N antennas, are arbitrarily distributed over the coverage area. The APs are connected to CPUs via fronthaul links. These can be used for sharing the CSI and channel statistics needed to decode the uplink data signals and precode the downlink data.

UE k is served only by the APs with indices in the set $\mathcal{M}_k \subset \{1, \ldots, L\}$, which is assumed fixed and known wherever needed. To perform coherent transmission processing (both between the antennas

at each AP and across the cooperating APs), knowledge of the channel responses from the UEs to the serving APs is required. It is particularly important for AP l to have estimates of the channel vector $\mathbf{h}_{kl}$ from UE k if $l \in \mathcal{M}_k$. Since the channels are assumed to be constant throughout one coherence block and change independently from one block to another, following the block fading model described in Section 2.3.1 on p. 208, they need to be estimated once per each coherence block. According to the TDD protocol defined in Section 2.3.2 on p. 211, τ_p samples are reserved for uplink pilot signaling in each coherence block. Therefore, each UE can transmit a pilot sequence that spans these τ_p samples and each AP can use the received signals to estimate the channel.

Ideally, we would like every UE to use a pilot sequence that is orthogonal to the pilots of all other UEs so there is no interference between the transmissions. However, since the pilots are τ_p-dimensional vectors, we can only find a set of at most τ_p mutually orthogonal sequences (i.e., a set of vectors that forms an orthogonal basis that spans $\mathbb{C}^{\tau_p}$). The finite length of the coherence blocks imposes the constraint $\tau_p \leq \tau_c$ that makes it impossible to assign mutually orthogonal pilots to all UEs in practical networks with $K \gg \tau_c$. Hence, we instead assume the network utilizes a set of τ_p mutually orthogonal pilot sequences $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_{\tau_p} \in \mathbb{C}^{\tau_p}$ that are designed to satisfy $\|\boldsymbol{\phi}_t\|^2 = \tau_p$, for $t = 1, \ldots, \tau_p$. One option is to use the columns of $\sqrt{\tau_p}\mathbf{I}_{\tau_p}$ as pilot sequences. We refer to [**massivemimobook**] for further examples of how to select pilot sequences. The only important aspect in the context of this monograph is that

$$\boldsymbol{\phi}_{t_1}^{\mathrm{H}} \boldsymbol{\phi}_{t_2} = \begin{cases} \tau_p & t_1 = t_2 \\ 0 & t_1 \neq t_2. \end{cases} \tag{4.1}$$

These τ_p pilots are assigned to the UEs in a deterministic way, for example, when they get access to the network. We return to the pilot assignment problem in Section 4.4.

More than one UE might be assigned to each pilot sequence. We denote the index of the pilot assigned to UE k as $t_k \in \{1, \ldots, \tau_p\}$ and define

$$\mathcal{P}_k = \{i : t_i = t_k, i = 1, \ldots, K\} \subset \{1, \ldots, K\} \tag{4.2}$$

as the set of UEs that use the same pilot as UE k, including itself. The

elements of $\boldsymbol{\phi}_{t_k}$ are scaled by the square-root of the uplink transmit power of UE k, which we denote as $\eta_k \geq 0$ in the pilot transmission phase. The elements of $\sqrt{\eta_k}\boldsymbol{\phi}_{t_k}$ are transmitted as the signal s_k in (2.3) over τ_p consecutive samples. The received signal at AP l during the entire pilot transmission is denoted as $\mathbf{Y}_l^{\text{pilot}} \in \mathbb{C}^{N \times \tau_p}$ and is given by

$$\mathbf{Y}_l^{\text{pilot}} = \sum_{i=1}^{K} \sqrt{\eta_i}\mathbf{h}_{il}\boldsymbol{\phi}_{t_i}^{\mathrm{T}} + \mathbf{N}_l \tag{4.3}$$

where $\mathbf{N}_l \in \mathbb{C}^{N \times \tau_p}$ is the receiver noise with i.i.d. elements distributed as $\mathcal{N}_{\mathbb{C}}(0, \sigma_{\text{ul}}^2)$. The received uplink signal $\mathbf{Y}_l^{\text{pilot}}$ is the observation that AP l can use to estimate the channels $\{\mathbf{h}_{kl} : l \in \mathcal{M}_k\}$ to all the UEs that it serves. The estimation can either be carried out directly at AP l or be delegated to the CPU. In the latter case, AP l acts as a relay and sends the received pilot signal $\mathbf{Y}_l^{\text{pilot}}$ to the CPU via the fronthaul link. The CPU can in principle compute all the channel estimates $\{\widehat{\mathbf{h}}_{kl} : k = 1, \ldots, K, l \in \mathcal{M}_k\}$ using the received pilot signals $\{\mathbf{Y}_l^{\text{pilot}} : l \in \mathcal{M}_k\}$. However, since the channel vectors are independent, there is no loss of optimality if the channel estimates are computed separately at each AP l in the set $\mathcal{M}_k$. The estimation results presented in this section apply to both cases.

4.2 MMSE Channel Estimation

Suppose we want to estimate $\mathbf{h}_{kl}$ (either at AP l or at the CPU), based on the received pilot signal $\mathbf{Y}_l^{\text{pilot}}$ in (4.3). The first step is to remove the interference from UEs using orthogonal pilots by multiplying the received signal $\mathbf{Y}_l^{\text{pilot}}$ with the normalized conjugate of the associated pilot $\boldsymbol{\phi}_{t_k}$. This yields $\mathbf{y}_{t_k l}^{\text{pilot}} = \mathbf{Y}_l^{\text{pilot}}\boldsymbol{\phi}_{t_k}^{\star}/\sqrt{\tau_p} \in \mathbb{C}^N$, given by

$$\begin{aligned} \mathbf{y}_{t_k l}^{\text{pilot}} &= \sum_{i=1}^{K} \frac{\sqrt{\eta_i}}{\sqrt{\tau_p}}\mathbf{h}_{il}\boldsymbol{\phi}_{t_i}^{\mathrm{T}}\boldsymbol{\phi}_{t_k}^{\star} + \frac{1}{\sqrt{\tau_p}}\mathbf{N}_l\boldsymbol{\phi}_{t_k}^{\star} \\ &= \underbrace{\sqrt{\eta_k\tau_p}\mathbf{h}_{kl}}_{\text{Desired part}} + \underbrace{\sum_{i \in \mathcal{P}_k \setminus \{k\}} \sqrt{\eta_i\tau_p}\mathbf{h}_{il}}_{\text{Interference}} + \underbrace{\mathbf{n}_{t_k l}}_{\text{Noise}} \end{aligned} \tag{4.4}$$

where the first term contains the desired channel $\mathbf{h}_{kl}$ scaled by $\sqrt{\eta_k\tau_p}$, which is the square root of the total pilot power $\|\sqrt{\eta_k}\boldsymbol{\phi}_{t_k}\|^2 = \eta_k\tau_p$.

The second term is the interference generated by the pilot-sharing UEs and the third term is the noise $\mathbf{n}_{t_k l} = \frac{1}{\sqrt{\tau_p}} \mathbf{N}_l \boldsymbol{\phi}_{t_k}^{\star} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_N, \sigma_{\text{ul}}^2 \mathbf{I}_N)$. The noise term has independent elements with variance σ_{ul}^2 since $\mathbf{N}_l$ has i.i.d. $\mathcal{N}_{\mathbb{C}}(0, \sigma_{\text{ul}}^2)$-elements and $\boldsymbol{\phi}_{t_k}^{\star}/\sqrt{\tau_p}$ is a vector with unit norm.

Note that $\mathbf{y}_{t_k l}^{\text{pilot}}$ is sufficient statistics for the estimation of $\mathbf{h}_{kl}$ since no information about $\mathbf{h}_{kl}$ or the pilot-sharing UEs' channels is lost by multiplying $\mathbf{Y}_l^{\text{pilot}}$ with $\boldsymbol{\phi}_{t_k}^{\star}/\sqrt{\tau_p}$. This is a positive consequence of using mutually orthogonal pilots; if a larger set of partially overlapping pilots would be used instead, then many more UEs are *partially* sharing the same pilot dimensions and this would make the computation of the MMSE estimator more computationally involved. In our case, the resulting signal $\mathbf{y}_{t_k l}^{\text{pilot}}$ in (4.4) matches with the MMSE estimation structure in Lemma 3.1 on p. 242, leading to the following result.

Corollary 4.1. The MMSE estimate of $\mathbf{h}_{kl}$ based on $\mathbf{y}_{t_k l}^{\text{pilot}}$ is

$$\widehat{\mathbf{h}}_{kl} = \sqrt{\eta_k \tau_p} \mathbf{R}_{kl} \boldsymbol{\Psi}_{t_k l}^{-1} \mathbf{y}_{t_k l}^{\text{pilot}} \tag{4.5}$$

where

$$\boldsymbol{\Psi}_{t_k l} = \mathbb{E}\left\{ \mathbf{y}_{t_k l}^{\text{pilot}} (\mathbf{y}_{t_k l}^{\text{pilot}})^{\text{H}} \right\} = \sum_{i \in \mathcal{P}_k} \eta_i \tau_p \mathbf{R}_{il} + \sigma_{\text{ul}}^2 \mathbf{I}_N \tag{4.6}$$

is the correlation matrix of the received signal in (4.4). The estimate $\widehat{\mathbf{h}}_{kl}$ and estimation error $\widetilde{\mathbf{h}}_{kl} = \mathbf{h}_{kl} - \widehat{\mathbf{h}}_{kl}$ are independent random variables distributed as

$$\widehat{\mathbf{h}}_{kl} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}_N, \eta_k \tau_p \mathbf{R}_{kl} \boldsymbol{\Psi}_{t_k l}^{-1} \mathbf{R}_{kl} \right) \tag{4.7}$$

$$\widetilde{\mathbf{h}}_{kl} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_N, \mathbf{C}_{kl}) \tag{4.8}$$

with the error correlation matrix

$$\mathbf{C}_{kl} = \mathbb{E}\{\widetilde{\mathbf{h}}_{kl} \widetilde{\mathbf{h}}_{kl}^{\text{H}}\} = \mathbf{R}_{kl} - \eta_k \tau_p \mathbf{R}_{kl} \boldsymbol{\Psi}_{t_k l}^{-1} \mathbf{R}_{kl}. \tag{4.9}$$

Proof. We consider the received signal in (4.4) and define $q = \sqrt{\eta_k \tau_p}$, $\mathbf{n} = \sum_{i \in \mathcal{P}_k \setminus \{k\}} \sqrt{\eta_i \tau_p} \mathbf{h}_{il} + \mathbf{n}_{t_k l}$, $\mathbf{R} = \mathbf{R}_{kl}$ and $\mathbf{S} = \sum_{i \in \mathcal{P}_k \setminus \{k\}} \eta_i \tau_p \mathbf{R}_{il} + \sigma_{\text{ul}}^2 \mathbf{I}_N$. The MMSE estimate and its properties then follow directly from Lemma 3.1 on p. 242. □

The MMSE estimator has this name because it minimizes the MSE $\mathbb{E}\{\|\mathbf{h}_{kl} - \widehat{\mathbf{h}}_{kl}\|^2\} = \mathbb{E}\{\|\widetilde{\mathbf{h}}_{kl}\|^2\}$ among all possible estimators. The MMSE

estimator in (4.5) is linear in the sense that $\widehat{\mathbf{h}}_{kl}$ is computed by multiplying the processed received signal $\mathbf{y}_{t_k l}^{\text{pilot}}$ with matrices. It is therefore sometimes called the linear MMSE estimator. However, we prefer to use the pure MMSE notion to make it clear that one cannot further reduce the MSE by using a non-linear estimator. Note that

$$\eta_k \tau_p \mathbf{R}_{kl} \boldsymbol{\Psi}_{t_k l}^{-1} \mathbf{R}_{kl} = \mathbf{R}_{kl} - \mathbf{C}_{kl} \tag{4.10}$$

so the correlation matrix of the channel estimate $\widehat{\mathbf{h}}_{kl}$ can also be expressed as $\mathbf{R}_{kl} - \mathbf{C}_{kl}$. The distribution of $\widehat{\mathbf{h}}_{kl}$ in (4.7) allows us to generate random realizations of the channel estimate without having to compute the received signal $\mathbf{y}_{t_k l}^{\text{pilot}}$ as an intermediate step. It is convenient for both mathematical analysis and numerical simulations.

In practice, the computation of $\widehat{\mathbf{h}}_{kl}$ in (4.5) requires knowledge of two matrices:

1. The spatial correlation matrix $\mathbf{R}_{kl} = \mathbb{E}\{\mathbf{h}_{kl}\mathbf{h}_{kl}^{\text{H}}\}$ of $\mathbf{h}_{kl}$;
2. The sum $\boldsymbol{\Psi}_{t_k l} = \sum_{i \in \mathcal{P}_k} \eta_i \tau_p \mathbf{R}_{il} + \sigma_{\text{ul}}^2 \mathbf{I}_N$ of the correlation matrices of pilot-sharing UEs and noise.

Both matrices depend on the statistics of the channels and thus are constant. Therefore, we can assume that the matrix $\sqrt{\eta_k \tau_p} \mathbf{R}_{kl} \boldsymbol{\Psi}_{t_k l}^{-1}$ can be precomputed. This matrix can be described by $N^2/2$ complex scalars, which can be exchanged via the fronthaul links to make $\sqrt{\eta_k \tau_p} \mathbf{R}_{kl} \boldsymbol{\Psi}_{t_k l}^{-1}$ available wherever needed (e.g., at AP l or the CPU). The required fronthaul signaling is negligible since the channel statistics are deterministic and thus constant throughout the entire data transmission.[1] If the matrix is precomputed, computing the MMSE estimate in a given coherence block entails first computing $\mathbf{y}_{t_k l}^{\text{pilot}}$ and then multiplying it with $\sqrt{\eta_k \tau_p} \mathbf{R}_{kl} \boldsymbol{\Psi}_{t_k l}^{-1}$ of each UE served by AP l. The first operation requires $N\tau_p$ complex multiplications per pilot sequence while the second needs N^2 complex multiplications per UE [**massivemimobook**]. In summary, the computational complexity for channel estimation at AP l

[1] In practice, the channel statistics will change due to UE mobility or new scheduling decisions, but measurements suggest roughly two orders of magnitude slower variations compared to the channel vectors.

is

$$|\mathcal{D}_l|(N\tau_p + N^2) \quad \text{complex multiplications} \tag{4.11}$$

per coherence block, if every UE that it serves uses a different pilot sequence (which is preferable). This value is scalable as $K \to \infty$ if $|\mathcal{D}_l|$ remains finite, which is in line with the sufficient condition for scalability presented in Lemma 2.1 on p. 219.

Remark 4.1 (Alternative estimation methods)**.** The MMSE estimator is the optimal choice when having full statistical knowledge, as assumed in this monograph (see Section 2.5.1 on p. 222). Note that each spatial correlation matrix $\mathbf{R}_{kl}$ contains N^2 elements and the computational complexity in (4.11) is proportional to N^2. There are alternative channel estimators in the literature that provide larger MSEs, since the MMSE estimator is optimal, but can be useful to limit the computational complexity or deal with incomplete statistical knowledge. This can be of interest in Cellular Massive MIMO systems, where N is large, since the complexity or overhead required to learn all the matrix elements can be substantial. However, none of these issues are pressing in Cell-free Massive MIMO, where N is small, and will thus not be covered in this monograph. We refer to [**massivemimobook**] for details on alternative estimators and to [**BSD16A**], [**CaireC17a**], [**NeumannJU17**], [**Sanguinetti2019a**], [**UpadhyaJU17**] for methods to estimate spatial correlation matrices in practice.

4.2.1 Normalized MSE

The value of the MSE depends on the average channel gain β_{kl} (among other factors) and a strong channel might have larger errors in absolute terms than a weaker channel. However, it is the relative size of the error that matters, not the absolute size. Hence, the estimation accuracy is quantified by considering the normalized MSE (NMSE). When it comes to the channel between AP l and UE k, the NMSE is defined as

$$\mathsf{NMSE}_{kl} = \frac{\mathbb{E}\{\|\mathbf{h}_{kl} - \widehat{\mathbf{h}}_{kl}\|^2\}}{\mathbb{E}\{\|\mathbf{h}_{kl}\|^2\}} = \frac{\mathrm{tr}(\mathbf{C}_{kl})}{\mathrm{tr}(\mathbf{R}_{kl})} = 1 - \frac{\eta_k \tau_p \mathrm{tr}(\mathbf{R}_{kl} \boldsymbol{\Psi}_{t_k l}^{-1} \mathbf{R}_{kl})}{\mathrm{tr}(\mathbf{R}_{kl})} \tag{4.12}$$

and measures the relative estimation error variance per antenna. The NMSE is 0 when perfect estimation is achieved while it is 1 if $\eta_k = 0$ so that we are only receiving noise and interference. In general, any reasonable estimator will provide an NMSE between 0 and 1, where smaller values are preferable.

In the case of single-antenna APs (i.e., $N = 1$), the spatial correlation matrix $\mathbf{R}_{kl}$ reduces to the average channel gain β_{kl} and (4.12) simplifies to

$$\mathsf{NMSE}_{kl} = 1 - \frac{\eta_k \tau_p \beta_{kl}}{\eta_k \tau_p \beta_{kl} + \underbrace{\sum_{i \in \mathcal{P}_k \setminus \{k\}} \eta_i \tau_p \beta_{il}}_{\text{Interference from pilot-sharing UEs}} + \sigma_{\text{ul}}^2} \tag{4.13}$$

which can be rewritten as

$$\mathsf{NMSE}_{kl} = 1 - \frac{\mathsf{SNR}_{kl}^{\text{pilot}}}{\mathsf{SNR}_{kl}^{\text{pilot}} + \sum_{i \in \mathcal{P}_k \setminus \{k\}} \mathsf{SNR}_{il}^{\text{pilot}} + 1} \tag{4.14}$$

where

$$\mathsf{SNR}_{kl}^{\text{pilot}} = \frac{\eta_k \tau_p \beta_{kl}}{\sigma_{\text{ul}}^2} \tag{4.15}$$

is the *effective* SNR of the pilot transmission from UE k to AP l. The word "effective" implies that the pilot processing gain τ_p is included in the SNR expression, which is achieved since the receiver processing in (4.4) coherently captures all the transmitted pilot energy without increasing the noise. If the pilot sequences are $\tau_p = 10$ samples long, then the effective SNR is 10 dB larger than the nominal SNR at a single sample. This gain is highly desirable for achieving good estimation quality also for UEs with limited transmit power and/or weak channel conditions.

Since it is the total pilot energy $\eta_k \tau_p$ that determines $\mathsf{SNR}_{kl}^{\text{pilot}}$, in practice, one can choose between spreading it over the τ_p samples of the pilot sequence or concentrating it into only one of them. The former solution keeps the peak-to-average-power ratio low but is sensitive to hardware imperfections; for example, a UE might only be able to approximately generate its pilot sequence, leading to only approximate

orthogonality between the sequences that are meant to be orthogonal. This problem does not occur in the latter case since UEs with orthogonal pilots are transmitting at different samples of the coherence block. On the other hand, the concentration of pilot energy into specific samples corresponds to boosting their power, which is not possible in all implementations since it can increase the peak-to-average-power ratio. Generally speaking, the system can improve the effective SNR of the pilot transmission by either increasing the pilot length or boosting the pilot power on specific samples.

4.2.2 Pilot Contamination

From (4.13), it follows that the interference generated by the pilot-sharing UEs increases the NMSE, thereby reducing the channel estimation quality. In the case of multiple-antenna APs (i.e., $N > 1$), the NMSE in (4.12) is exactly the same in (4.13) if uncorrelated Rayleigh fading is assumed (i.e., $\mathbf{R}_{kl} = \beta_{kl}\mathbf{I}_N$). This means that there is neither an advantage nor a disadvantage in using multiple antennas at the APs in the absence of spatial correlation. With spatially correlated channels, the NMSE in (4.12) has a more complicated structure. It is clearly increased by the interference from the pilot-sharing UEs, which enters into $\boldsymbol{\Psi}_{t_{kl}}$ in (4.6). Unlike (4.13), (4.12) depends not only on the average channel gains but on the full spatial correlation matrices $\mathbf{R}_{il}$, for $i \in \mathcal{P}_k$. Intuitively, the NMSE should be better when the interfering UEs' channels have very different spatial correlation properties than the desired UE. For example, if we know a priori that the channels of two UEs contain multipath components distributed over widely different sets of angular directions, then their corresponding spatial correlation matrices should have different dominant eigenspaces and this should somehow be beneficial during channel estimation. This intuition will be confirmed later on in Section 4.3.3.

The "pilot interference" generated by the pilot-sharing UEs is known as *pilot contamination* in the Cellular Massive MIMO literature and behaves differently from the receiver noise; it does not only reduce the estimation quality, but also makes the channel estimates of pilot-sharing UEs correlated. To see this, let us compare the MMSE estimate $\widehat{\mathbf{h}}_{kl}$ in

(4.5) of UE k with the estimate

$$\widehat{\mathbf{h}}_{il} = \sqrt{\eta_i \tau_p} \mathbf{R}_{il} \boldsymbol{\Psi}_{t_i l}^{-1} \mathbf{y}_{t_i l}^{\text{pilot}} \tag{4.16}$$

of another UE $i \in \mathcal{P}_k \setminus \{k\}$ using the same pilot. In this case, we have that $\boldsymbol{\Psi}_{t_k l} = \boldsymbol{\Psi}_{t_i l}$ and $\mathbf{y}_{t_k l}^{\text{pilot}} = \mathbf{y}_{t_i l}^{\text{pilot}}$ such that (4.16) can be rewritten as

$$\widehat{\mathbf{h}}_{il} = \sqrt{\eta_i \tau_p} \mathbf{R}_{il} \boldsymbol{\Psi}_{t_k l}^{-1} \mathbf{y}_{t_k l}^{\text{pilot}}. \tag{4.17}$$

By comparing (4.17) with (4.5), we notice that only the scalar and the first matrix differ. If $\mathbf{R}_{kl}$ is invertible, we can write the relation as

$$\widehat{\mathbf{h}}_{il} = \sqrt{\frac{\eta_i}{\eta_k}} \mathbf{R}_{il} \mathbf{R}_{kl}^{-1} \widehat{\mathbf{h}}_{kl}, \quad i \in \mathcal{P}_k. \tag{4.18}$$

This implies that the two estimates are correlated, with the cross-correlation matrix given by

$$\mathbb{E}\left\{ \widehat{\mathbf{h}}_{kl} \widehat{\mathbf{h}}_{il}^{\text{H}} \right\} = \sqrt{\eta_k \eta_i} \tau_p \mathbf{R}_{kl} \boldsymbol{\Psi}_{t_k l}^{-1} \mathbf{R}_{il}, \quad i \in \mathcal{P}_k. \tag{4.19}$$

This happens although the true channels were assumed statistically independent (i.e., $\mathbb{E}\{\mathbf{h}_{kl} \mathbf{h}_{il}^{\text{H}}\} = \mathbf{0}_{N \times N}$, for $i \neq k$). Observe that if there is no spatial correlation so that $\mathbf{R}_{kl} = \beta_{kl} \mathbf{I}_N$ and $\mathbf{R}_{il} = \beta_{il} \mathbf{I}_N$, then the channel estimate $\widehat{\mathbf{h}}_{il}$ in (4.17) for any $i \in \mathcal{P}_k \setminus \{k\}$ will be the same as the channel estimate $\widehat{\mathbf{h}}_{kl}$ except for a scaling factor. In this case, the channel estimates are fully correlated.

In conclusion, UEs that transmit the same pilot sequence contaminate each others' channel estimates, in the same way as in Cellular Massive MIMO systems [**Marzetta2010a**], [**Sanguinetti2019a**] and other types of cellular systems [**Niemela2004a**], [**Sapienza2001a**]. The contamination not only reduces the estimation quality (i.e., increases the NMSE) but also makes the channel estimates statistically dependent [**Sanguinetti2019a**]. This has an important impact beyond channel estimation, since the contamination makes it harder to mitigate interference between UEs that use the same pilot in uplink and downlink. This will be investigated in Sections 5 and 6 on p. 294 and p. 355, respectively.

4.2.3 MMSE Estimation of the Collective Channel

So far, we have considered the estimation of a channel vector $\mathbf{h}_{kl}$ between a generic AP l and UE k. To perform coherent processing of

the signals at multiple APs, it is necessary to have knowledge of the collective channel $\mathbf{h}_k \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_M, \mathbf{R}_k)$ defined in (2.15), where $\mathbf{R}_k = \text{diag}(\mathbf{R}_{k1}, \ldots, \mathbf{R}_{kL}) \in \mathbb{C}^{M \times M}$ is the block-diagonal collective spatial correlation matrix. However, the estimate $\widehat{\mathbf{h}}_k = [\widehat{\mathbf{h}}_{k1}^{\mathrm{T}} \ldots \widehat{\mathbf{h}}_{kL}^{\mathrm{T}}]^{\mathrm{T}}$ of $\mathbf{h}_k$ from all APs to UE k can be only partially computed since only the APs in $\mathcal{M}_k \subset \{1, \ldots, L\}$ are computing estimates of the channels and/or send their pilot signals to the CPU. This means that only the following partial channel estimate is known in a scalable cell-free network:

$$\mathbf{D}_k \widehat{\mathbf{h}}_k \triangleq \begin{bmatrix} \mathbf{D}_{k1} \widehat{\mathbf{h}}_{k1} \\ \vdots \\ \mathbf{D}_{kL} \widehat{\mathbf{h}}_{kL} \end{bmatrix} \sim \mathcal{N}_{\mathbb{C}} \left(\mathbf{0}_M, \eta_k \tau_p \mathbf{D}_k \mathbf{R}_k \boldsymbol{\Psi}_{t_k}^{-1} \mathbf{R}_k \mathbf{D}_k \right) \tag{4.20}$$

where $\mathbf{D}_k = \text{diag}(\mathbf{D}_{k1}, \ldots, \mathbf{D}_{kL})$ is a block-diagonal matrix with $\mathbf{D}_{kl}$ being defined in (2.1) as $\mathbf{I}_N$ for APs that serve UE k and zero otherwise. Moreover, $\boldsymbol{\Psi}_{t_k}^{-1} = \text{diag}(\boldsymbol{\Psi}_{t_k 1}^{-1}, \ldots, \boldsymbol{\Psi}_{t_k L}^{-1})$ contains the inverses of the received pilot signal correlation matrices defined in (4.6). The estimation error is $\mathbf{D}_k \widetilde{\mathbf{h}}_k = \mathbf{D}_k \mathbf{h}_k - \mathbf{D}_k \widehat{\mathbf{h}}_k \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_M, \mathbf{D}_k \mathbf{C}_k)$ with $\mathbf{C}_k = \text{diag}(\mathbf{C}_{k1}, \ldots, \mathbf{C}_{kL})$. Note that $\mathbf{D}_k \mathbf{C}_k = \mathbf{D}_k \mathbf{C}_k \mathbf{D}_k$ so we use the former expression for brevity.

The NMSE of $\mathbf{D}_k \widehat{\mathbf{h}}_k$ can be computed as

$$\begin{aligned} \text{NMSE}_k &= \frac{\mathbb{E}\{\|\mathbf{D}_k \mathbf{h}_k - \mathbf{D}_k \widehat{\mathbf{h}}_k\|^2\}}{\mathbb{E}\{\|\mathbf{D}_k \mathbf{h}_k\|^2\}} = \frac{\text{tr}(\mathbf{D}_k \mathbf{C}_k)}{\text{tr}(\mathbf{D}_k \mathbf{R}_k)} \overset{(a)}{=} \frac{\sum_{l=1}^{L} \text{tr}(\mathbf{D}_{kl} \mathbf{C}_{kl})}{\sum_{l=1}^{L} \text{tr}(\mathbf{D}_{kl} \mathbf{R}_{kl})} \\ &= 1 - \frac{\eta_k \tau_p \sum_{l \in \mathcal{M}_k} \text{tr}(\mathbf{R}_{kl} \boldsymbol{\Psi}_{t_k l}^{-1} \mathbf{R}_{kl})}{\sum_{l \in \mathcal{M}_k} \text{tr}(\mathbf{R}_{kl})} \end{aligned} \tag{4.21}$$

where (a) follows from the block-diagonal structure of both $\mathbf{C}_k$ and $\mathbf{R}_k$. The NMSE of the collective channel in (4.21) has a similar form as the NMSE of an individual AP-to-UE channel in (4.12). Note that (4.21) is not the sum or average of the individual NMSEs, but contains a summation of MSEs in the numerator normalized by a summation of channel gains.

In the case of single-antenna APs or uncorrelated Rayleigh fading, (4.21) can be simplified as

$$\begin{aligned}\mathsf{NMSE}_k &= 1 - \frac{\eta_k \tau_p \sum\limits_{l \in \mathcal{M}_k} \frac{\beta_{kl}^2}{\eta_k \tau_p \beta_{kl} + \sum\limits_{i \in \mathcal{P}_k \setminus \{k\}} \eta_i \tau_p \beta_{il} + \sigma_{\mathrm{ul}}^2}}{\sum\limits_{l \in \mathcal{M}_k} \beta_{kl}} \\ &= 1 - \frac{\sum\limits_{l \in \mathcal{M}_k} \frac{(\mathsf{SNR}_{kl}^{\mathrm{pilot}})^2}{\mathsf{SNR}_{kl}^{\mathrm{pilot}} + \sum\limits_{i \in \mathcal{P}_k \setminus \{k\}} \mathsf{SNR}_{il}^{\mathrm{pilot}} + 1}}{\sum\limits_{l \in \mathcal{M}_k} \mathsf{SNR}_{kl}^{\mathrm{pilot}}}. \end{aligned} \tag{4.22}$$

In any case, the presence of pilot-sharing UEs will increase the NMSE and create a correlation between their respective channel estimates.

4.3 Impact of Architecture, Contamination, & Spatial Correlation

We will exemplify how the estimation accuracy of the MMSE estimator is affected by the cell-free architecture, pilot contamination, and spatial correlation. The numerical examples will uncover some of the basic properties that determine the NMSE and its impact on communication performance.

4.3.1 Impact of the Cell-Free Architecture

To gain insights into the basic impact of channel estimation errors in cell-free networks, we begin by considering single-antenna APs and the estimation of the collective channel of an arbitrary UE k that has a unique pilot sequence: $\mathcal{P}_k = \{k\}$. In this case, the NMSE in (4.22) reduces to

$$\mathsf{NMSE}_k = 1 - \frac{\sum\limits_{l \in \mathcal{M}_k} \frac{\mathsf{SNR}_{kl}^{\mathrm{pilot}}}{1 + \frac{1}{\mathsf{SNR}_{kl}^{\mathrm{pilot}}}}}{\sum\limits_{l \in \mathcal{M}_k} \mathsf{SNR}_{kl}^{\mathrm{pilot}}}. \tag{4.23}$$

If one AP has a much larger value of its effective SNR than the other APs in $\mathcal{M}_k$, then (4.23) can be approximated as

$$\mathsf{NMSE}_k \approx 1 - \frac{1}{1 + \frac{1}{\mathsf{SNR}_{kl_{\max}}^{\mathrm{pilot}}}} = \frac{1}{\mathsf{SNR}_{kl_{\max}}^{\mathrm{pilot}} + 1} \tag{4.24}$$

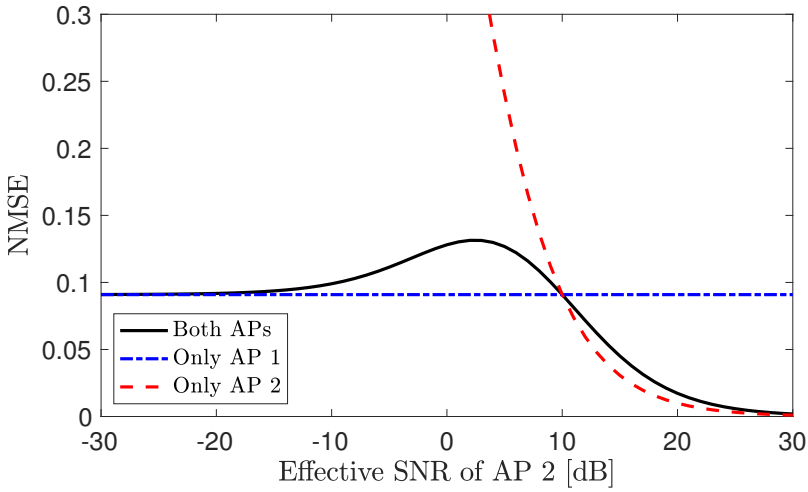


Figure 4.1: NMSE when two APs may serve UE k. We assume that $\mathsf{SNR}_{k1}^{\text{pilot}}$ is fixed to 10 dB while $\mathsf{SNR}_{k2}^{\text{pilot}}$ varies from -30 dB to 30 dB. The cases in which only AP 1 or 2 is serving UE k is reported for comparison.

with $l_{\max} = \arg\max_{l \in \mathcal{M}_k} \mathsf{SNR}_{kl}^{\text{pilot}}$. This approximation is exact if $|\mathcal{M}_k| = 1$ or when all the serving APs have the same value of $\mathsf{SNR}_{kl}^{\text{pilot}}$. However, when the APs have different values of $\mathsf{SNR}_{kl}^{\text{pilot}}$, then the exact NMSE in (4.23) will be larger than the approximation in (4.24).

This fact is exemplified in Figure 4.1 where (4.23) is plotted for the case in which UE k is served by $|\mathcal{M}_k| = 2$ APs, or just one of them. We assume that $\mathsf{SNR}_{k1}^{\text{pilot}}$ is fixed to 10 dB while we let $\mathsf{SNR}_{k2}^{\text{pilot}}$ vary from -30 dB to 30 dB. There is one NMSE curve for the case when both APs serve the UE and two curves for the case when only one of the APs serves the UE. We observe that the NMSE with both APs is higher than the NMSE with only AP 1 serving the UE if $\mathsf{SNR}_{k2}^{\text{pilot}} < \mathsf{SNR}_{k1}^{\text{pilot}}$. In contrast, a lower NMSE is obtained for $\mathsf{SNR}_{k2}^{\text{pilot}} > \mathsf{SNR}_{k1}^{\text{pilot}}$. This shows that having more than one serving AP is not necessarily improving the estimation accuracy of the collective channel.

To gain further insights into this, we revisit the simulation scenario from Figure 1.10 on p. 194, where a cell-free network was compared with a corresponding small-cell setup and a single-cell Massive MIMO setup.

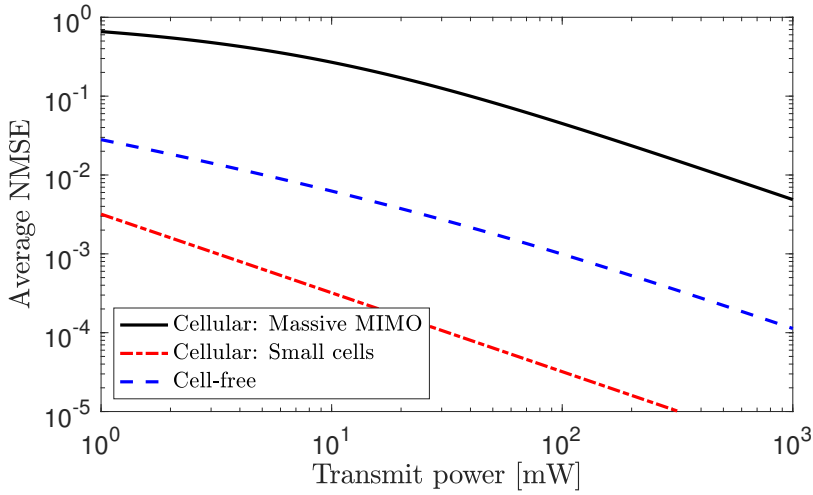


Figure 4.2: Average NMSE for a single UE in the three different setups illustrated in Figure 1.9 on p. 190.

In the cell-free case, we consider $L = 64$ single-antenna APs that are deployed on a square grid in a coverage area of $400\,\text{m} \times 400\,\text{m}$ as shown in Figure 1.9(c) on p. 190. The bandwidth is 10 MHz and the noise power is $\sigma^2_{\text{ul}} = -96\,\text{dBm}$. The channels are Rayleigh fading, the channel gains are modelled as in (1.1), and the propagation distances are computed assuming that the APs are deployed 10 m above the UE. Comparisons are made with a single-cell setup with a 64-antenna Massive MIMO AP (with uncorrelated Rayleigh fading) and with a cellular network consisting of 64 small cells deployed at the same 64 AP locations.

Figure 4.2 shows the average of the NMSE in (4.23) for different uniformly distributed UE positions. The pilot transmit power η_k varies from 1 mW to 1 W and the pilot length is $\tau_p = 10$. When comparing the three setups, we notice that it is preferable to have many single-antenna APs rather than a single AP with a large co-located array (as in conventional Cellular Massive MIMO). However, the small-cell network achieves better average estimation accuracy than the cell-free architecture since the AP with the lowest NMSE is always chosen in the small-cell network. This is in agreement with the results of Figure 4.1

where the lowest NMSE is achieved when only using the AP with the best channel. However, this does not mean that the cell-free architecture will achieve a lower communication performance, but only that the varying estimation quality must be taken into account when the cell-free network is combining the received signals from multiple APs.

The SE gain of using a cell-free architecture will be quantified in later sections of this monograph, but we will give an indication of the performance gain by considering the SNR that is achieved in the data detection. Suppose only UE k is active in the network, then the local estimate in (2.5) of the data signal s_k can be expressed as

$$\widehat{s}_{kl} = \underbrace{\mathbf{v}_{kl}^{\mathrm{H}} \widehat{\mathbf{h}}_{kl} s_k}_{\text{Signal over estimated channel}} + \underbrace{\mathbf{v}_{kl}^{\mathrm{H}} \widetilde{\mathbf{h}}_{kl} s_k}_{\text{Signal over unknown channel}} + \underbrace{\mathbf{v}_{kl}^{\mathrm{H}} \mathbf{n}_l}_{\text{Noise}}.$$

This is the data estimate at AP l. All the local estimates of the serving APs are sent to a CPU where the final estimate of s_k is obtained as

$$\begin{aligned} \widehat{s}_k &= \sum_{l \in \mathcal{M}_k} \widehat{s}_{kl} \\ &= \sum_{l \in \mathcal{M}_k} \mathbf{v}_{kl}^{\mathrm{H}} \widehat{\mathbf{h}}_{kl} s_k + \sum_{l \in \mathcal{M}_k} \mathbf{v}_{kl}^{\mathrm{H}} \widetilde{\mathbf{h}}_{kl} s_k + \sum_{l \in \mathcal{M}_k} \mathbf{v}_{kl}^{\mathrm{H}} \mathbf{n}_l. \end{aligned} \tag{4.25}$$

For simplicity, we assume the APs apply MR combining with $\mathbf{v}_{kl} = \widehat{\mathbf{h}}_{kl}$ and compute the resulting SNR as

$$\begin{aligned} \mathsf{SNR}_k^{\text{data}} &= \frac{\mathbb{E}\left\{ \left| \sum_{l \in \mathcal{M}_k} \mathbf{v}_{kl}^{\mathrm{H}} \widehat{\mathbf{h}}_{kl} s_k \right|^2 \right\}}{\mathbb{E}\left\{ \left| \sum_{l \in \mathcal{M}_k} \mathbf{v}_{kl}^{\mathrm{H}} \mathbf{n}_l \right|^2 \right\}} = \frac{p_k \mathbb{E}\left\{ \left| \sum_{l \in \mathcal{M}_k} \| \widehat{\mathbf{h}}_{kl} \|^2 \right|^2 \right\}}{\sigma_{\text{ul}}^2 \sum_{l \in \mathcal{M}_k} \mathbb{E}\left\{ \| \widehat{\mathbf{h}}_{kl} \|^2 \right\}} \\ &= \frac{\sum_{l \in \mathcal{M}_k} p_k \mathrm{tr}\left((\mathbf{R}_{kl} - \mathbf{C}_{kl})^2 \right)}{\sum_{l \in \mathcal{M}_k} \sigma_{\text{ul}}^2 \mathrm{tr}(\mathbf{R}_{kl} - \mathbf{C}_{kl})} + \sum_{l \in \mathcal{M}_k} \frac{p_k \mathrm{tr}(\mathbf{R}_{kl} - \mathbf{C}_{kl})}{\sigma_{\text{ul}}^2} \end{aligned} \tag{4.26}$$

where the expectations with respect to the channel estimates are computed using Lemma B.5 on p. 446. We stress that this is only one of the many ways to compute SNRs in the presence of estimation errors.

Figure 4.3 shows the distribution of the SNR value in (4.26) achieved at different random UE locations. We consider the same setups as in

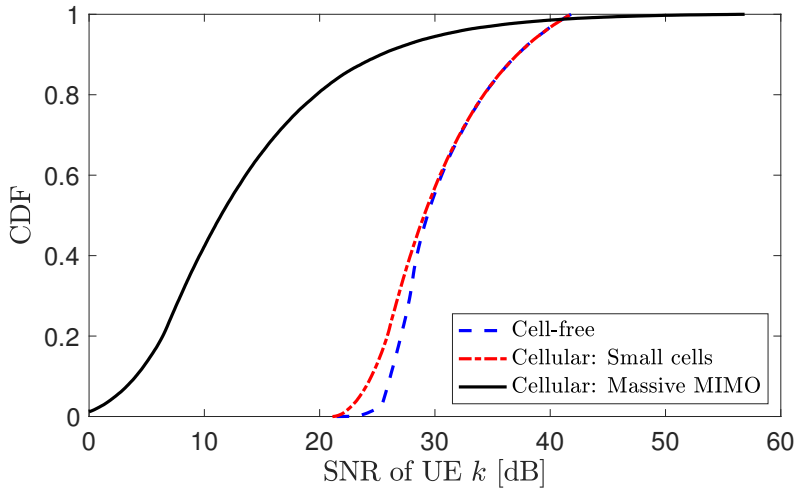


Figure 4.3: CDF of the SNR (4.26) achieved in the uplink by the UE with MR combining and MMSE estimation in the same three setups as in Figure 4.2.

Figure 4.2 and let $p_k = \eta_k = 10\,\text{mW}$. The results show that the cell-free network achieves a higher SNR in the data detection than the corresponding small-cell setup (and Cellular Massive MIMO). This happens despite the lower average estimation quality of the channel estimates, as reported in Figure 4.2. The reason is that the cell-free network achieves a beamforming gain by coherently processing the signals at multiple APs. An even larger benefit will be achieved in a setup with multiple UEs, where the cell-free system excels at mitigating interference, as first illustrated in Section 1.3.2 on p. 194. This will be thoroughly analyzed in later sections.

4.3.2 Impact of Pilot Contamination

We now shift focus to the impact of pilot contamination. We concentrate on an arbitrary single-antenna AP l and assume that it wants to estimate the channel of UE 1, while UE 2 uses the same pilot. From (4.14), we obtain

$$\mathsf{NMSE}_{1l} = 1 - \frac{\mathsf{SNR}_{1l}^{\text{pilot}}}{\mathsf{SNR}_{1l}^{\text{pilot}} + \mathsf{SNR}_{2l}^{\text{pilot}} + 1}. \tag{4.27}$$

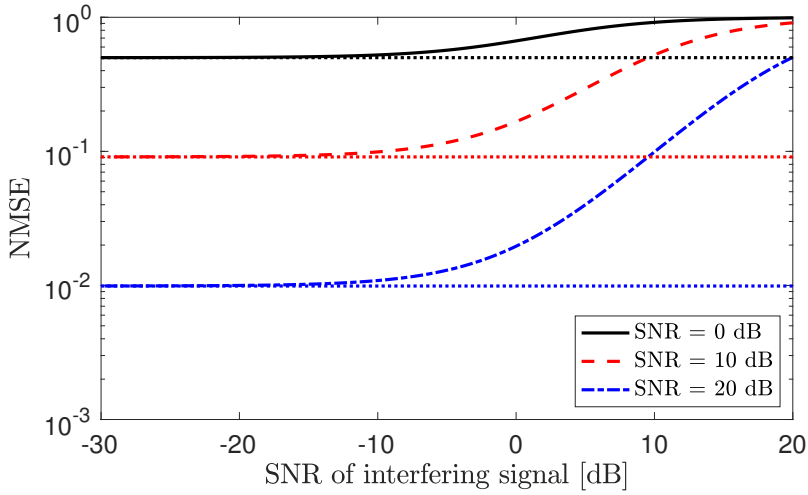


Figure 4.4: NMSE as a function of the SNR $\mathsf{SNR}_{2l}^{\text{pilot}}$ of the interfering pilot signal. The effective SNR of the desired UE is $\mathsf{SNR}_{1l}^{\text{pilot}} = 0, 10$, or 20 dB. The dotted lines correspond to the NMSE in the absence of pilot contamination.

The NMSE increment caused by pilot contamination depends on how much the interference is affecting the term $\mathsf{SNR}_{2l}^{\text{pilot}} + 1$ in the denominator of (4.27), which accounts for interference and noise. The contamination is small when the value is close to one. Figure 4.4 shows the NMSE of the desired channel estimate for $\mathsf{SNR}_{1l}^{\text{pilot}} \in \{0, 10, 20\}$ dB and varying values of $\mathsf{SNR}_{2l}^{\text{pilot}}$ from -30 to 20 dB. The dotted lines correspond to the case without pilot contamination (i.e., $\mathsf{SNR}_{2l}^{\text{pilot}} = 0$) and are reported as references. The NMSE increases when the interfering signal becomes stronger, particularly when the desired UE has a strong channel to the AP. However, the pilot contamination effect is negligible whenever the interfering signal is 10 dB weaker than the noise. We can thus conclude that each UE is only sensitive to pilot contamination from UEs that are relatively close to its serving APs. Another implication is that the UEs that an AP serves should preferably have different pilot sequences to limit pilot contamination [**Bjornson2020a**], [**Sabbagh2018a**]. This is aligned with the Cellular Massive MIMO literature where orthogonal pilots are normally utilized within every cell

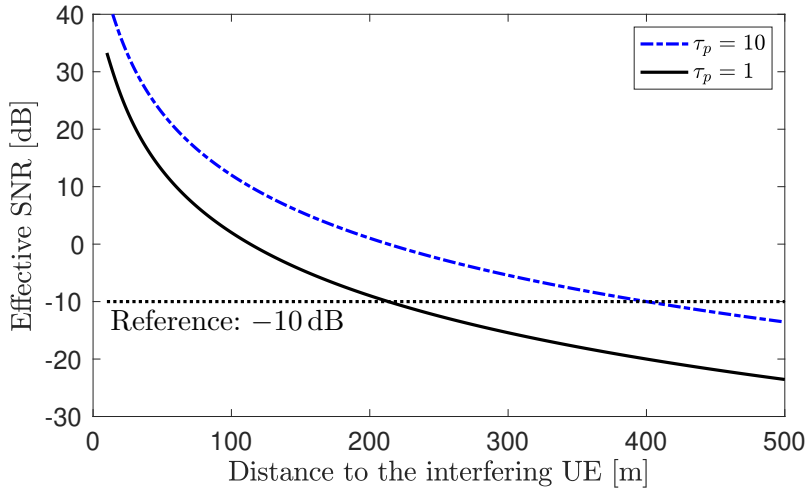


Figure 4.5: The effective SNR (4.15) of an interfering pilot signal depends on the distance between the AP and the interfering UE. In this simulation, the same propagation model as in Figure 4.2 is used. Pilot contamination has a negligible impact on the NMSE when the SNR is below the reference curve of −10 dB.

but reused in different cells [**massivemimobook**], [**Marzetta2016a**].

The distance between the AP and the interfering UE is one of the main factors determining the SNR of the interfering signal. Figure 4.5 shows the effective SNR (4.15) as a function of the propagation distance using the same parameter values as in Figure 4.3. The SNR reduces with the distance but can be above the −10 dB reference curve for several hundred meters. Hence, pilot contamination is a non-negligible issue in practice. However, it is a gradual effect so the largest impact occurs when the interfering UE is rather close to the AP. The effective SNR increases by 10 dB when increasing the pilot length from $\tau_p = 1$ to $\tau_p = 10$, which affects both the desired signal and contaminating signal. Even if the signal-to-contamination ratio remains the same, the propagation distances for which pilot contamination is a problem will grow since it is determined by comparing the contamination with the noise. On the other hand, each pilot can be reused less frequently among the UEs when τ_p is increased. Hence, we can expect the use of longer pilots to be beneficial when it comes to limiting the pilot contamination,

as long as we assign the pilots properly to the UEs. We return to this in Section 4.4. Importantly, the curves in Figure 4.5 will move up or down depending on the transmit power, bandwidth, and propagation model so it is not the exact numbers that matter but the general behavior.

4.3.3 Impact of Spatial Correlation

We now quantify the impact of spatial correlation among AP antennas on the channel estimation accuracy and pilot contamination.

Impact of Spatial Correlation on Estimation Accuracy

Consider the spatially correlated channel $\mathbf{h} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_N, \mathbf{R})$ between an arbitrary AP and an arbitrary UE, where the UE and AP indices are dropped for simplicity. We assume the UE uses a unique pilot sequence, thus the NMSE in (4.12) can be simplified as

$$\mathsf{NMSE} = 1 - \frac{\eta\tau_p \mathrm{tr}\left(\mathbf{R}\left(\eta\tau_p\mathbf{R} + \sigma_{\mathrm{ul}}^2\mathbf{I}_N\right)^{-1}\mathbf{R}\right)}{\mathrm{tr}(\mathbf{R})}. \tag{4.28}$$

Let $\mathbf{R} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{H}}$ denote the eigenvalue decomposition of the spatial correlation matrix, where the columns of the unitary matrix $\mathbf{U} \in \mathbb{C}^{N\times N}$ contain the eigenvectors (also called eigendirections) and the diagonal matrix $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_N)$ contains the corresponding non-negative eigenvalues where $\sum_{n=1}^{N}\lambda_n = \mathrm{tr}(\mathbf{R}) = N\beta$. By using the eigenvalue decomposition, the NMSE in (4.28) can be rewritten as

$$\begin{aligned}
\mathsf{NMSE} &= 1 - \frac{\eta\tau_p \mathrm{tr}\left(\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{H}}\left(\eta\tau_p\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{H}} + \sigma_{\mathrm{ul}}^2\mathbf{I}_N\right)^{-1}\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{H}}\right)}{N\beta} \\
&= 1 - \frac{\eta\tau_p}{N\beta}\mathrm{tr}\left(\boldsymbol{\Lambda}\left(\eta\tau_p\boldsymbol{\Lambda} + \sigma_{\mathrm{ul}}^2\mathbf{I}_N\right)^{-1}\boldsymbol{\Lambda}\right) \\
&= 1 - \frac{\eta\tau_p}{N\beta}\sum_{n=1}^{N}\frac{\lambda_n^2}{\eta\tau_p\lambda_n + \sigma_{\mathrm{ul}}^2} \\
&= 1 - \frac{1}{N\beta}\sum_{n=1}^{N}\frac{\mathsf{SNR}^{\mathrm{pilot}}\lambda_n^2}{\mathsf{SNR}^{\mathrm{pilot}}\lambda_n + \beta}
\end{aligned} \tag{4.29}$$

where the second equality follows from moving the matrices $\mathbf{U}$ and $\mathbf{U}^{\mathrm{H}}$ into the inverse, applying the cyclic shift property of the trace

(see the second identity in Lemma B.2 on p. 445), and utilizing that $\mathbf{U}\mathbf{U}^{\mathrm{H}} = \mathbf{U}^{\mathrm{H}}\mathbf{U} = \mathbf{I}_N$. An important observation from (4.29) is that the NMSE depends on the eigenvalues but not on the eigenvectors. This is a direct consequence of the fact that the estimation error correlation matrix $\mathbf{C} = \mathbf{R} - \eta\tau_p\mathbf{R}\left(\eta\tau_p\mathbf{R} + \sigma_{\mathrm{ul}}^2\mathbf{I}_N\right)^{-1}\mathbf{R}$ has the same eigenvectors as the spatial correlation matrix $\mathbf{R}$ [**massivemimobook**].

We cannot change the spatial correlation matrices of a practical channel, but it is anyway important to understand what impact the correlation has. To this end, we now search for the eigenvalue distributions that maximize and minimize the NMSE under the constraint that $\sum_{n=1}^{N}\lambda_n = N\beta$. We will make use of the following result.

Lemma 4.2. The function $\mathsf{NMSE} : \mathbb{R}_{\geq 0}^{N} \to \mathbb{R}_{\geq 0}$ of the eigenvalue vector $\boldsymbol{\lambda} = [\lambda_1 \ \ldots \ \lambda_N]^{\mathrm{T}}$ given in (4.29) is strictly concave on its domain.

Proof. The proof is available in Appendix C.1.1 on p. 448. □

A consequence of the strict concavity stated in Lemma 4.2 is that the maximum value of the NMSE in (4.29) under the average channel gain constraint $\sum_{n=1}^{N}\lambda_n = N\beta$ is unique and achieved when all eigenvalues are equal: $\lambda_n = \beta$, for $n = 1, \ldots, N$. This shows that, among all the spatial correlation matrices $\mathbf{R}$ that satisfy $\mathrm{tr}(\mathbf{R}) = N\beta$, the uncorrelated fading model with $\mathbf{R} = \beta\mathbf{I}_N$ provides the largest NMSE.

To identify the spatial correlation that instead minimizes the NMSE, we need the following lemma.

Lemma 4.3. Suppose the eigenvalues of $\mathbf{R}$ are sorted in decaying order:

$$\lambda_1 \geq \lambda_2 \geq \ldots \lambda_r > \lambda_{r+1} = \ldots = \lambda_N = 0 \tag{4.30}$$

where $r \leq N$ is the rank. Let $\boldsymbol{\lambda} = [\lambda_1 \ \ldots \ \lambda_{r-2} \ \lambda_{r-1} \ \lambda_r \ 0 \ \ldots \ 0]^{\mathrm{T}}$ denote the corresponding eigenvalue vector. Then, $\mathsf{NMSE}\left(\boldsymbol{\lambda}\right) > \mathsf{NMSE}\left(\boldsymbol{\lambda}'\right)$ where $\boldsymbol{\lambda}' = [\lambda_1 \ \ldots \ \lambda_{r-2} \ \lambda_{r-1} + \lambda_r \ 0 \ \ldots \ 0]^{\mathrm{T}}$.

Proof. The proof is available in Appendix C.1.2 on p. 448. □

This lemma provides a procedure for reducing the NMSE by modifying the two smallest non-zero eigenvalues of the correlation matrix. More precisely, they are replaced with one eigenvalue being the sum of

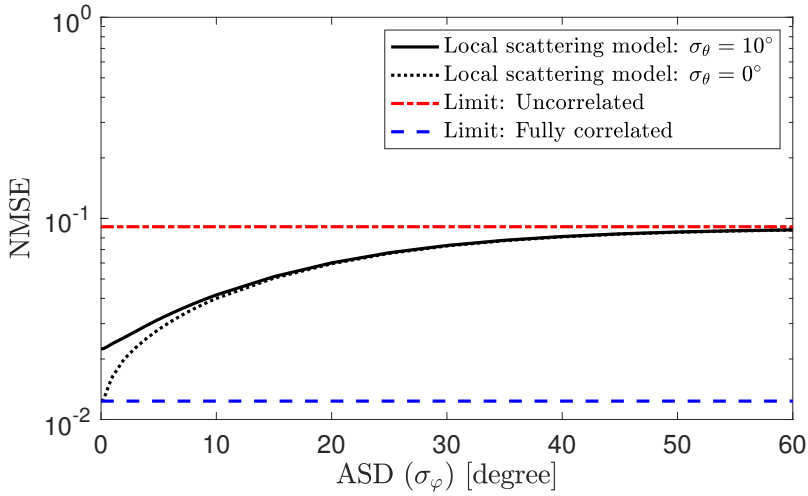


Figure 4.6: NMSE in the estimation of an arbitrary UE's spatially correlated channel, as a function of the azimuth ASD σ_φ. The local scattering model (2.18) with Gaussian angular distribution is used. The results are averaged over different nominal azimuth angles and the elevation angle is $\theta = -15°$. The effective SNR (4.15) is 10 dB and the AP is equipped with $N = 8$ antennas.

two smallest non-zero eigenvalues and one eigenvalue being identically zero. If we repeat this procedure multiple times, we will progressively reduce the NMSE and eventually reach the case when $\lambda_1 = N\beta$, and $\lambda_n = 0$, for $n \geq 2$, which provides the smallest possible NMSE among all spatial correlation matrices $\mathbf{R}$ that satisfy $\sum_{n=1}^{N} \lambda_n = \mathrm{tr}(\mathbf{R}) = N\beta$. This corresponds to the case where $\mathbf{R} = \lambda_1 \mathbf{u}_1 \mathbf{u}_1^{\mathrm{H}}$ is a rank-one matrix and thus represents the strongest type of spatial correlation. Any rank-one correlation matrix with the same $\lambda_1 = N\beta$ achieves the same estimation accuracy. The conclusion is that the higher the spatial correlation is, the easier it is to estimate the channel since there is more structure to be utilized by the estimator.

The above properties are illustrated numerically in Figure 4.6, where the NMSE is shown for the local scattering model in (2.18), as a function of the azimuth ASD σ_φ. The results are averaged over different nominal azimuth angles, while the elevation angle is fixed at $\theta = -15°$ and the elevation ASD is either $\sigma_\theta = 0°$ or $\sigma_\theta = 10°$. The effective SNR (4.15)

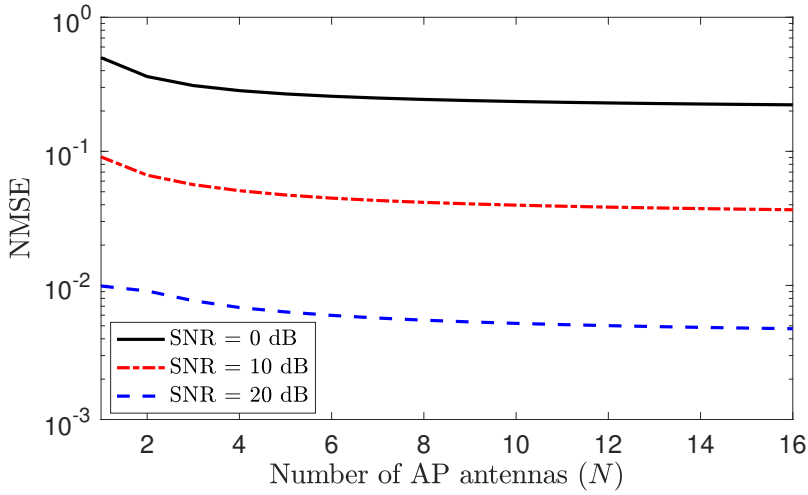


Figure 4.7: NMSE in the estimation of an arbitrary spatially correlated channel, as a function of the number of antennas N at the AP. The effective SNR (4.15) is 0, 10, or 20 dB and the spatial correlation is modeled as in Figure 4.6 with $\sigma_\varphi = \sigma_\theta = 10°$.

is 10 dB and there are $N = 8$ antennas at the AP. Figure 4.6 shows that the NMSE is smaller when the ASD reduces (i.e., with higher spatial correlation). The NMSE for uncorrelated channels is shown in Figure 4.6 as a reference. For strongly spatially correlated channels, the NMSE can be one order of magnitude smaller than in the uncorrelated case, but this benefit is basically lost when $\sigma_\varphi \geq 40°$. The ASD in the elevation dimension only affects the NMSE when the ASD in the azimuth dimension is small. The lower bound for fully correlated channels is attained in the extreme case of $\sigma_\varphi = \sigma_\theta = 0°$.

The impact of the number of AP antennas is studied in Figure 4.7 using the same correlation model as in Figure 4.6 but with the ASDs $\sigma_\varphi = \sigma_\theta = 10°$. The NMSE is shown as a function of the number of antennas for effective SNRs of 0, 10, and 20 dB. The NMSE reduces as the number of AP antennas increases since the channel realizations observed at adjacent antennas are strongly correlated, which is utilized by the MMSE estimator to improve the estimation quality. However, only marginal gains are observed for $N \geq 8$, irrespective of the effective SNR. This happens when the ULA has become so wide that the outmost

antennas experience almost uncorrelated channel realizations.

Impact of Spatial Correlation on Pilot Contamination

Next, we will analyze the impact of spatial correlation on pilot contamination. Assume UE 1 and UE 2 use the same pilot sequence and consider an arbitrary AP whose index is omitted. The NMSE of UE 1 in (4.12) takes the form

$$\mathsf{NMSE}_1 = 1 - \frac{\eta_1 \tau_p \mathrm{tr}\left(\mathbf{R}_1 \left(\eta_1 \tau_p \mathbf{R}_1 + \eta_2 \tau_p \mathbf{R}_2 + \sigma_{\mathrm{ul}}^2 \mathbf{I}_N\right)^{-1} \mathbf{R}_1\right)}{\mathrm{tr}(\mathbf{R}_1)}. \quad (4.31)$$

Unlike the NMSE in (4.27) for single-antenna APs, (4.31) depends not only on the effective SNRs but also on the full spatial correlation matrices $\mathbf{R}_1$ and $\mathbf{R}_2$ of the UEs. We now observe that (4.31) can be lower bounded by Lemma B.4 on p. 446 as

$$\mathsf{NMSE}_1 \geq 1 - \frac{\eta_1 \tau_p \mathrm{tr}\left(\mathbf{R}_1 \left(\eta_1 \tau_p \mathbf{R}_1 + \sigma_{\mathrm{ul}}^2 \mathbf{I}_N\right)^{-1} \mathbf{R}_1\right)}{\mathrm{tr}(\mathbf{R}_1)} \quad (4.32)$$

where the equality occurs if and only if the correlation matrices are spatially orthogonal $\mathbf{R}_1 \mathbf{R}_2 = \mathbf{0}_{N \times N}$. This condition means that the eigenspaces are non-overlapping. If $\mathbf{R}_1 \neq \mathbf{0}_{N \times N}$ or $\mathbf{R}_2 \neq \mathbf{0}_{N \times N}$, the condition can only be satisfied if both $\mathbf{R}_1$ and $\mathbf{R}_2$ are rank-deficient.

In the case of spatially orthogonal correlation matrices (i.e., $\mathbf{R}_1 \mathbf{R}_2 = \mathbf{0}_{N \times N}$), the NMSE is completely unaffected by the interfering UE. Therefore, in theory, it is possible to completely avoid pilot contamination between the UEs. While the orthogonality condition is unlikely to hold in practice, a good rule-of-thumb is to assign pilots to the UEs such that $\mathrm{tr}\left(\mathbf{R}_1 \mathbf{R}_2\right)$ is small.

Figure 4.8 shows the NMSE of the estimate of the desired channel with $N = 4$ antennas. The spatial correlation is modeled by the local scattering model with elevation angles $\theta_1 = \theta_2 = -15°$ and ASDs $\sigma_\varphi = 10°$, $\sigma_\theta = 0°$. The azimuth angle of the desired UE is $\varphi_1 = 30°$, while the azimuth angle φ_2 of the interfering UE is varied between $-60°$ and $60°$. The effective SNR of the desired UE is 10 dB and the interfering signal either has the same SNR or is 20 dB weaker. When the UE angles are well-separated, the NMSE is below 0.1, irrespective of how strong the interfering pilot signal is. This shows that pilot contamination has

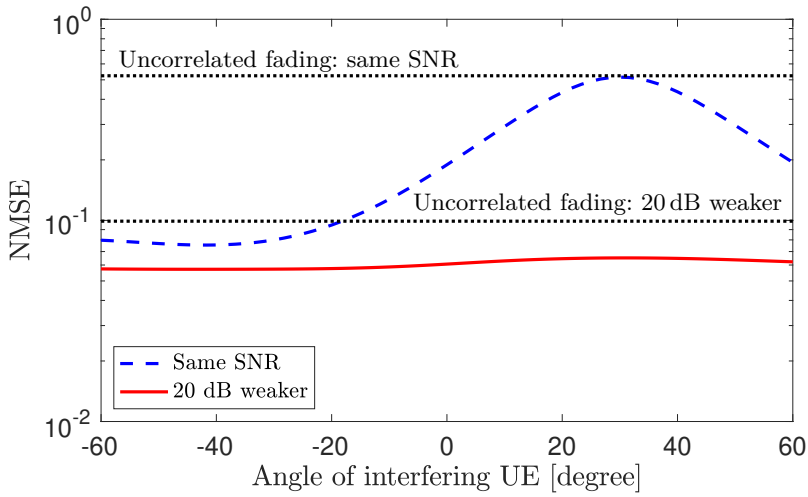


Figure 4.8: NMSE in the estimation of the desired UE's channel when there is an interfering UE, which uses the same pilot. There are $N = 4$ antennas. The local scattering model is used with $\theta_1 = \theta_2 = -15°$, $\sigma_\varphi = 10°$, and $\sigma_\theta = 0°$. The desired UE has a nominal azimuth angle of 30°, while the azimuth angle of the interfering UE is varied between $-60°$ and 60°. The NMSE with uncorrelated fading is shown as references.

a minor impact on the estimation quality when the UEs have well-separated eigenspaces. The NMSE increases when the UEs have similar angles and reaches a peak at $\varphi_1 = \varphi_2$. The variations are large when the interfering UE has a strong channel to the AP, while the variations are small when the interfering signal is 20 dB weaker. The figure also includes reference curves for the case of uncorrelated fading, where the NMSE is angle-independent. We notice that the NMSE with correlated fading is always lower (or equal) to the NMSE with uncorrelated fading, even in the worst-case scenario where the desired and interfering UEs have identical angles. Hence, spatial channel correlation is helpful in practice to improve the estimation quality under pilot contamination.

4.4 Pilot Assignment and Dynamic Cooperation Cluster Formation

The τ_p mutually orthogonal pilots must be reused among the UEs and can be assigned to them in an effort to limit the pilot contamination

effect. The examples given earlier in this section have demonstrated that this entails keeping a large physical distance between the pilot-sharing UEs, so that each AP that serves a given UE will be subject to pilot contamination that drowns in the receiver noise. This is a good way of picturing what we want to achieve when performing a pilot assignment but it is a simplification since the observations were made using a propagation model in which the SNR is a strictly decreasing function of the distance. In reality, the SNR and propagation distance are correlated, but there is also a fair amount of random variations that can be modeled as shadow fading. Hence, when designing pilot assignment algorithms, we should not take geometric parameters, such as the locations on a map, as input but instead consider the spatial correlation matrices $\mathbf{R}_{kl}$. Those matrices are describing both the average channel gains β_{kl}, including shadow fading, and the spatial channel characteristics.

The pilot assignment is a combinatorial problem. There are $(\tau_p)^K$ possible assignments in a setup with K UEs and τ_p pilots, thus the complexity of evaluating all of them grows exponentially with the number of UEs. The implication is that only suboptimal methods are feasible in practice. We will present one such algorithm in this section, and it will be utilized for performance evaluation in later sections. The algorithm will iteratively assign pilots to the UEs by always selecting the one that leads to the least pilot contamination. This is a so-called greedy algorithm and comes with no performance guarantees, but it will effectively avoid worst-case situations where closely spaced UEs (i.e., UEs with similar SNRs) are assigned to the same pilot. It is also practically attractive since the pilot assignment decision for a UE is made locally at a neighboring AP, instead of involving all APs in the network. We will provide a literature review of alternative pilot assignment methods in Section 7.4 on p. 427.

The formation of the DCC is tightly connected to the pilot assignment. We noted earlier in this section that a single-antenna AP can only reasonably serve one UE per pilot, namely the one having the strongest channel gain. The reason is that this UE would cause strong pilot contamination to all the pilot-sharing UEs if the AP tried to serve some of those UEs as well. Consequently, a basic algorithm for cluster formation

is to first assign pilots to the UEs and then let each AP serve exactly τ_p UEs; for every pilot, the AP serves the UE with the strongest channel gain among the subset of UEs that have been assigned that pilot. This is most likely a suboptimal algorithm, particularly for multiple-antenna APs which could use spatial correlation to distinguish between UEs that send the same pilot and thereby serve multiple UEs per pilot if they have nearly non-overlapping eigenspaces (recall Figure 4.8). However, it makes good practical sense since every AP can make its clustering decisions locally. This is the clustering algorithm that will be utilized for performance evaluation in later sections. We will provide a literature review of alternative methods in Section 7.5 on p. 431.

An Algorithm for Joint Pilot Assignment and Cluster Formation

The basic pilot assignment algorithm consists of two steps. First, the τ_p UEs with indices from 1 to τ_p are assigned mutually orthogonal pilots: UE k uses pilot k for $k = 1, \ldots, \tau_p$. The remaining UEs, with indices from $\tau_p + 1$ to K, are then assigned pilots one after the other. UE k begins by determining which AP it has the strongest channel to. The index of this AP is computed as

$$\ell = \underset{l \in \{1,\ldots,L\}}{\arg\max} \ \beta_{kl}. \tag{4.33}$$

Since AP ℓ is expected to contribute strongly to the service of UE k, it is preferable to assign UE k to the pilot for which AP ℓ experiences the least pilot contamination. Hence, for each pilot t, the AP can compute the sum of the average channel gains $\beta_{i\ell}$ of the UEs that have already been assigned to that pilot. The AP then identifies the index of the pilot where the pilot interference is minimized:

$$\tau = \underset{t \in \{1,\ldots,\tau_p\}}{\arg\min} \sum_{\substack{i=1 \\ t_i = t}}^{k-1} \beta_{i\ell}. \tag{4.34}$$

This pilot is assigned to the UE and the algorithm then continues with the next UE. When all the UEs have been assigned to pilots, the clusters can be created. Each AP goes through each pilot and identifies which of

Algorithm 4.1 Basic pilot assignment and cooperation clustering

Initialization: Set $\mathcal{M}_1 = \ldots = \mathcal{M}_K = \emptyset$
for $k = 1, \ldots, \tau_p$ **do**
    $t_k \leftarrow k$ ▷ Assign orthogonal pilots to first τ_p UEs
end for
for $k = \tau_p + 1, \ldots, K$ **do**
    $\ell \leftarrow \underset{l \in \{1,\ldots,L\}}{\arg\max}\ \beta_{kl}$ ▷ Find the best AP for UE k
    $\tau \leftarrow \underset{t \in \{1,\ldots,\tau_p\}}{\arg\min} \sum_{\substack{i=1 \\ t_i = t}}^{k-1} \beta_{i\ell}$ ▷ Find pilot with least interference at AP ℓ
    $t_k \leftarrow \tau$ ▷ Assign pilot τ to UE k
end for
for $l = 1, \ldots, L$ **do**
    for $t = 1, \ldots, \tau_p$ **do**
        $i \leftarrow \underset{k \in \{1,\ldots,K\}: t_k = t}{\arg\max}\ \beta_{kl}$ ▷ Find UE that AP l serves on pilot t
        $\mathcal{M}_i \leftarrow \mathcal{M}_i \cup \{l\}$
    end for
end for
Output: Pilot assignment $t_1, \ldots, t_K$ and DCCs $\mathcal{M}_1, \ldots, \mathcal{M}_K$

the UEs have the largest channel gain among those using that pilot. The AP will serve that UE. The procedure is summarized in Algorithm 4.1.

This algorithm resembles the ones proposed in [**Bjornson2020a**], [**Chen2016a**]. The key idea was that whenever a new UE is admitted to the network, it begins by identifying the AP that it has the strongest channel gain to. In practical systems, the APs are periodically broadcasting synchronization signals to convey basic access information. The channel gains can be measured based on these signals. The accessing UE computes ℓ according to (4.33) and appoints AP ℓ as its *Master AP*. Note that this selection is made in a user-centric manner. The UE also uses the broadcasted signal to synchronize to the AP. The UE can contact its Master AP via a standard random access procedure [**sanguinetti13bis**], [**Sesia09**]. The Master AP computes the preferred pilot τ using (4.34) and informs the surrounding APs of the existence of the new UE. The surrounding APs can then determine if they should change which UEs to serve on pilot τ. This is a dynamic variation of Algorithm 4.1 that can be applied when new UEs are entering the system. It can also be applied when a UE has moved to such a large extent that its channel statistics have changed and, therefore, it can be treated as a new UE [**Bjornson2020a**], [**Chen2016a**]. The three main steps of this initial access algorithm are illustrated in Figure 4.9.

The algorithm described above is distributed in the sense that each UE is managed separately and each AP makes its pilot assignment decisions and cluster formation locally. A key challenge for any distributed cluster formation algorithm, including Algorithm 4.1, is the risk that some UEs will not be served by any AP. This can happen when a particular AP has the strongest channel to more than τ_p UEs, while it is only allowed to serve at most τ_p of them. A solution to this problem is to require that each UE k is served by at least one AP, namely AP ℓ that is selected according to (4.33) [**Bjornson2020a**]. This will lead to pilot contamination but at least guarantees that each UE is being served. Since the intended operating regime of Cell-free Massive MIMO is the ultra-dense regime of $L \gg K$, each AP will on average only have the strongest channel to $K/L < 1$ UEs. Hence, if $\tau_p = 10$, there is a very low risk that one AP will have the strongest channel to more than ten UEs, but it can happen.

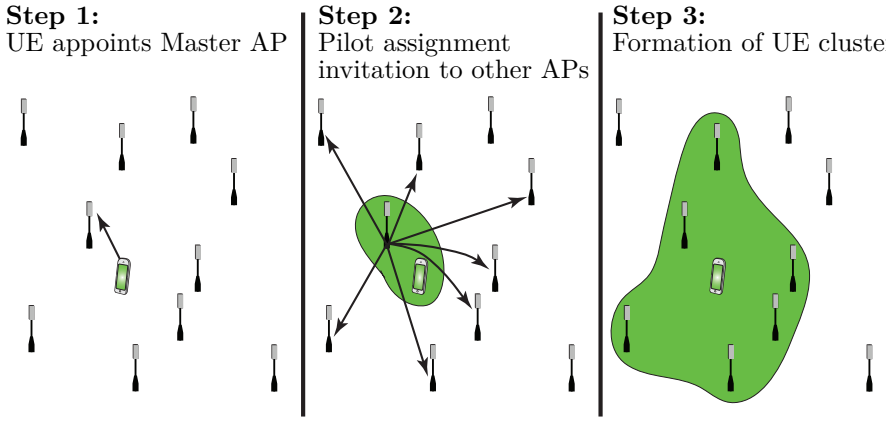


Figure 4.9: The procedure from [**Bjornson2020a**] for joint initial access, pilot assignment, and cluster formation. This is one way of implementing the steps in Algorithm 4.1 in a dynamic fashion.

Remark 4.2 (Pilot decontamination)**.** Specific signal processing algorithms have been developed for situations where pilot contamination is the main performance-limiting factor. The general aim is to expand the dimension of the pilot sequences and we will briefly describe three main categories of such *pilot decontamination* approaches. The first category utilizes channel statistics, particularly spatial channel correlation, to assign pilots to UEs to proactively limit the pilot contamination effect. Some key references from the Cellular Massive MIMO literature are [**BjornsonHS17**], [**Huh2012a**], [**Yin2013a**], [**Xu2015a**], but these are not directly applicable to cell-free networks where each AP has a small number of antennas. Approaches developed for cell-free networks will be described in Section 7.4 on p. 427. The second approach is to make use of the uplink data as additional pilots, which is known as semi-blind or data-aided estimation [**Carvalho1997a**]. The main complication is that the data is unknown, but since long sequences of random data samples are fairly orthogonal, the transmitted signals will fall into different subspaces that can be identified by the receiving AP. Some key references from the Cellular Massive MIMO literature are [**Ma2014a**], [**Mueller2014b**], [**Neumann2014a**], [**Ngo2012a**], [**Vinogradova2016a**], [**Yin2016a**]. The existing algorithms rely

on asymptotic properties that occur when N and τ_c are jointly growing large, thus they are not applicable in cell-free networks where N is small. The third approach is to superimpose low-power pilot sequences onto the uplink data, thereby making the pilot length equal to the number of uplink samples per coherence block. The benefit is that many more orthogonal pilot sequences can be generated, while the drawback is that the pilots and data will now mutually interfere. There are situations where it brings benefits, as exemplified in [**Upadhya2014a**], [**Upadhya2017b**], [**Verenzuela2017a**], [**Verenzuela2020a**] for Cellular Massive MIMO. However, it cannot increase the capacity scaling at high SNR, which is achievable by reserving certain dimensions for pilots [**Zheng2002a**]. We will not cover any of these methods in detail in this monograph, because it will later become apparent that pilot contamination is not a main limiting factor in Cell-free Massive MIMO.

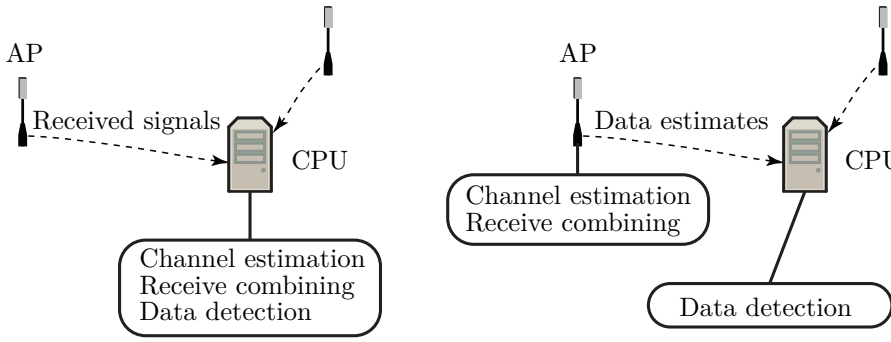
4.5 Summary of the Key Points in Section 4

- Channel estimation is fundamental to perform coherent transmission processing both between AP antennas and across cooperating APs. Since the channels are constant throughout one coherence block and change from one block to another, they need to be estimated once per each coherence block.
- Channel estimation is carried out by sending uplink pilots. The estimates can be either computed at the CPU or at the APs. Both options give the same estimation quality.
- The MMSE estimator exploits channel statistics to obtain good estimates. The interference generated by pilot-sharing UEs not only reduces the estimation quality but also makes the channel estimates statistically dependent. This phenomenon is called pilot contamination.
- Spatial correlation is helpful to improve the estimation quality. With pilot contamination, the correlation between channel estimates is low when the correlation matrices of the pilot-sharing UEs are sufficiently different.
- The estimation accuracy achieved with many single-antenna APs is higher than in the case of a single AP with a large co-located array. The highest estimation accuracy is achieved when the UE is served only by the APs that are close to it. However, coherent transmission processing among APs makes the network robust to lower-quality channel estimates.
- As a rule-of-thumb, the pilots should be assigned to the UEs so that each AP is close to (i.e., has a high effective SNR to) at most one UE per pilot. Optimal assignment is impractical but a sequential algorithm was provided in Algorithm 4.1.

5

Uplink Operation

This section describes two different uplink implementations of User-centric Cell-free Massive MIMO, which are characterized by different degrees of cooperation among the APs. Section 5.1 considers a centralized operation in which the pilot and data signals received at all APs are gathered (through the fronthaul links) at the CPU, which performs channel estimation and data detection. The achievable SE is derived based on the MMSE channel estimates and used to obtain the optimal, but unscalable, centralized receive combining. Alternative scalable combining schemes are then proposed. In Section 5.2, the SE analysis and design of scalable receive combining schemes are extended to a distributed operation where the local MMSE channel estimates are used at the APs to obtain local estimates of the UE data, which are then gathered and linearly combined at the CPU for final detection. To exemplify the performance and properties of cell-free networks under somewhat realistic conditions, a simulation setup is defined in Section 5.3 that will be used as a running example in the remainder of this monograph. Performance analysis and comparisons are carried out in Section 5.4, while a summary of the key points is provided in Section 5.5.



(a) Centralized operation (b) Distributed operation

Figure 5.1: The uplink signal processing tasks can be divided between the APs and CPU in different ways. In the centralized operation, the channel estimation, receive combining, and data detection are done at the CPU. In the distributed operation, everything, except the data detection, is done at the APs.

5.1 Centralized Uplink Operation

The most advanced uplink implementation of Cell-free Massive MIMO is a fully centralized operation, where the APs only act as remote-radio heads or relays that forward their received baseband signals to the CPU for processing. More precisely, each AP l sends the τ_p received pilot signals $\{\mathbf{y}_{tl}^{\text{pilot}} : t = 1, \ldots, \tau_p\}$ in (4.4) and the received uplink data signal $\mathbf{y}_l^{\text{ul}}$ in (2.3) to the CPU, which performs channel estimation, receive combining and data detection. The centralized operation is illustrated in Figure 5.1(a). Although we call it *centralized operation*, this should not be interpreted as building a network that gathers the information related to all UEs at a single point at the center of the network. As described in Section 1.1 on p. 169, the CPU is a logical entity and its tasks can be divided between many geographically distributed edge-cloud processors [**Interdonato2019a**], as illustrated in Figure 1.3 on p. 170. Different UEs might use different processors. Hence, in practice, centralized operation means that each UE is associated with one nearby processor which we refer to as "the CPU" and all its serving APs send their respective received uplink signals to it. We will now explain what processing is done at the CPU and quantify the achievable uplink SE.

For each UE k, the pilot signals are individually used at the CPU to compute the partial MMSE estimates $\{\mathbf{D}_k\widehat{\mathbf{h}}_i : i = 1, \ldots, K\}$ of the

collective channels $\{\mathbf{h}_i : i = 1, \ldots, K\}$ from all UEs to the APs that serve UE k (i.e., the APs with indices $l \in \mathcal{M}_k$). The estimation procedure was described in Section 4.2 on p. 266. Recall that $\mathbf{D}_k = \mathrm{diag}(\mathbf{D}_{k1}, \ldots, \mathbf{D}_{kL})$ is a block-diagonal matrix with $\mathbf{D}_{kl}$ being defined in (2.1) as $\mathbf{I}_N$ for APs that serve UE k and zero otherwise.

The received uplink data signals at the serving APs are jointly used at the CPU to compute an estimate $\widehat{s}_k$ of the signal s_k transmitted by UE k. From the signal model in Section 2.3.3 on p. 213, this is achieved by summing up the inner products between the effective receive combining vectors $\mathbf{D}_{kl}\mathbf{v}_{kl}$ and $\mathbf{y}_l^{\mathrm{ul}}$ for $l = 1, \ldots, L$. This yields the estimate

$$\begin{aligned} \widehat{s}_k &= \sum_{l=1}^{L} \widehat{s}_{kl} \\ &= \sum_{l=1}^{L} \mathbf{v}_{kl}^{\mathrm{H}} \mathbf{D}_{kl} \mathbf{y}_l^{\mathrm{ul}} = \mathbf{v}_k^{\mathrm{H}} \mathbf{D}_k \mathbf{y}^{\mathrm{ul}} \end{aligned} \tag{5.1}$$

where $\mathbf{v}_k = [\mathbf{v}_{k1}^{\mathrm{T}} \ \ldots \ \mathbf{v}_{kL}^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{C}^{LN}$ is the centralized combining vector and $\mathbf{y}^{\mathrm{ul}} \in \mathbb{C}^{LN}$ is the collective uplink data signal given by

$$\mathbf{y}^{\mathrm{ul}} = \begin{bmatrix} \mathbf{y}_1^{\mathrm{ul}} \\ \vdots \\ \mathbf{y}_L^{\mathrm{ul}} \end{bmatrix} = \sum_{i=1}^{K} \mathbf{h}_i s_i + \mathbf{n} \tag{5.2}$$

with $\mathbf{n} = [\mathbf{n}_1^{\mathrm{T}} \ \ldots \ \mathbf{n}_L^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{C}^{LN}$ being the collective noise vector. Notice that (5.2) depends on all the channel vectors $\{\mathbf{h}_i : i = 1, \ldots, K\}$ since all the APs receive the signal from all UEs. Since $\mathbf{D}_{kl} = \mathbf{0}_{N \times N}$ implies $\mathbf{v}_{kl}^{\mathrm{H}} \mathbf{D}_{kl} = \mathbf{0}_N^{\mathrm{T}}$ in (5.1), the CPU will however apply receive combining using only the data signals of the APs with $\mathbf{D}_{kl} \neq \mathbf{0}_{N \times N}$. In other words, (5.1) can be equivalently written as $\widehat{s}_k = \sum_{l \in \mathcal{M}_k} \mathbf{v}_{kl}^{\mathrm{H}} \mathbf{y}_l^{\mathrm{ul}}$ by including only the subset of APs that serves UE k in the summation. While this alternative expression is more compact at this point, it would require much more notation if used for performance analysis. That is why we utilize (5.1) in this section.

The expression in (5.2) is mathematically equivalent to the signal model of an uplink single-cell Massive MIMO system with correlated fading [**massivemimobook**], if one treats the CPU as a receiver equipped with LN antennas. However, there are some key differences:

1. Multiple UEs that are managed by the CPU are using the same pilot, which is normally avoided in single-cell systems.

2. The antennas are distributed over different geographical locations. Hence, the collective channel is distributed as $\mathbf{h}_k \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_{LN}, \mathbf{R}_k)$ where the spatial correlation matrix $\mathbf{R}_k = \text{diag}(\mathbf{R}_{k1}, \ldots, \mathbf{R}_{kL}) \in \mathbb{C}^{LN \times LN}$ has a block-diagonal structure, which is normally not the case in single-cell systems.

3. Not all antennas are being used for signal detection, but only those belonging to the APs that serve the particular UE.

Despite these distinct differences, we can compute an achievable uplink SE, based on the knowledge of the partial MMSE estimates $\{\mathbf{D}_k \widehat{\mathbf{h}}_i : i = 1, \ldots, K\}$, using the methodology from the Cellular Massive MIMO literature [**massivemimobook**]. From this expression, we can further obtain the optimal centralized receive combining vector $\mathbf{v}_k$. All the analysis in this section relies on the assumption that the spatial correlation matrices $\{\mathbf{R}_{kl} : k = 1, \ldots, K, l = 1, \ldots, L\}$ are perfectly known at the CPU to perform channel estimation, receive combining and data detection. We will return to this assumption in Section 5.1.5.

5.1.1 Spectral Efficiency With Centralized Operation

While the ergodic capacity of fully cooperative networks with perfect CSI is known in some cases [**Estella2019a**], it is generally unknown in the considered case with imperfect CSI. However, we can rigorously analyze the performance by using the standard capacity lower bounds described in Section 3.2 on p. 243. As a first step, we use (5.2) to rewrite (5.1) as

$$\widehat{s}_k = \underbrace{\mathbf{v}_k^{\mathrm{H}} \mathbf{D}_k \widehat{\mathbf{h}}_k s_k}_{\text{Desired signal over estimated channel}} + \underbrace{\mathbf{v}_k^{\mathrm{H}} \mathbf{D}_k \widetilde{\mathbf{h}}_k s_k}_{\text{Desired signal over unknown channel}} + \underbrace{\sum_{\substack{i=1 \\ i \neq k}}^{K} \mathbf{v}_k^{\mathrm{H}} \mathbf{D}_k \mathbf{h}_i s_i}_{\text{Interference}} + \underbrace{\mathbf{v}_k^{\mathrm{H}} \mathbf{D}_k \mathbf{n}}_{\text{Noise}} \tag{5.3}$$

where the desired signal term has been divided into two parts: one that is received over the known partially estimated channel $\mathbf{D}_k\widehat{\mathbf{h}}_k$ from UE k and one that is received over the unknown part of the channel, represented by the estimation error $\mathbf{D}_k\widetilde{\mathbf{h}}_k$. The former part can be utilized straight away for signal detection, while the latter part is less useful since only the distribution of the estimation error is known, as reported in Section 4.2.3 on p. 272. An achievable SE can be computed by treating the latter part as additional interference in the signal detection, by utilizing Lemma 3.5 on p. 248. If the first term in (5.3) is treated as the desired part while the other three terms are treated as noise in the receiver, the following SE is achievable.

Theorem 5.1. An achievable SE of UE k in the centralized operation is

$$\mathsf{SE}_k^{(\mathrm{ul,c})} = \frac{\tau_u}{\tau_c}\mathbb{E}\left\{\log_2\left(1+\mathsf{SINR}_k^{(\mathrm{ul,c})}\right)\right\} \quad \text{bit/s/Hz} \tag{5.4}$$

where the instantaneous effective SINR is given by

$$\mathsf{SINR}_k^{(\mathrm{ul,c})} = \frac{p_k\left|\mathbf{v}_k^{\mathrm{H}}\mathbf{D}_k\widehat{\mathbf{h}}_k\right|^2}{\sum\limits_{\substack{i=1\\i\neq k}}^{K} p_i\left|\mathbf{v}_k^{\mathrm{H}}\mathbf{D}_k\widehat{\mathbf{h}}_i\right|^2 + \mathbf{v}_k^{\mathrm{H}}\mathbf{Z}_k\mathbf{v}_k + \sigma_{\mathrm{ul}}^2\|\mathbf{D}_k\mathbf{v}_k\|^2} \tag{5.5}$$

with

$$\mathbf{Z}_k = \sum_{i=1}^{K} p_i\mathbf{D}_k\mathbf{C}_i\mathbf{D}_k \tag{5.6}$$

and the expectation is with respect to the channel estimates. The matrix $\mathbf{C}_i$ is the error correlation matrix of the collective channel $\mathbf{h}_i$.

Proof. The proof is available in Appendix C.2.1 on p. 449. □

The pre-log factor τ_u/τ_c in (5.4) is the fraction of each coherence block that is used for uplink data transmission. The term $\mathsf{SINR}_k^{(\mathrm{ul,c})}$ takes the form of an "effective instantaneous SINR" [**massivemimobook**], with the desired signal power received over the estimated channel $\mathbf{D}_k\widehat{\mathbf{h}}_k$ in the numerator and the interference plus noise in the denominator. More precisely, the numerator contains $p_k|\mathbf{v}_k^{\mathrm{H}}\mathbf{D}_k\widehat{\mathbf{h}}_k|^2$, while the first

term in the denominator is the interference that is received over the estimated parts of the interfering channels, $\mathbf{Z}_k$ is the combined spatial correlation matrix of the signals received over the unknown parts of the channels, and the last term in the denominator is the noise power after receive combining. Notice that the matrix $\mathbf{C}_i$ is the error correlation matrix of the collective channel $\mathbf{h}_i$; see Section 4.2.3 on p. 272. We call (5.5) an "effective" instantaneous SINR since it cannot be measured in the system at any particular point in time, because the averaging over the estimation errors does not occur instantaneously. However, the SE is effectively the same as that of a fading single-antenna single-user channel where (5.5) is the instantaneously measurable SNR and the receiver has perfect CSI. In particular, this means that the desired data signal can be encoded and the received signal can be decoded as if we would communicate over such a fading AWGN channel.

The SE expression in Theorem 5.1 is not in closed form since there is an expectation in front of the logarithm in (5.4). However, the SE can be easily computed numerically using Monte Carlo methods, which means that we approximate the expectation with an average over a large number of random realizations. More precisely, we can generate realizations of the channel estimates in a large set of coherence blocks and compute the value of the logarithm for each one of them, followed by taking the sample mean of these values.

Theorem 5.1 is rather general in the sense that the SE is derived using the DCC framework and considering multi-antenna APs with arbitrary correlated fading channels and arbitrary DCC. The case of $N = 1$ was considered in [**Attarifar2020a**], [**Nayebi2016a**], [**Riera2018a**], [**Yang2019a**], while [**Bashar2018a**], [**Chen2018b**] considered $N \geq 1$ with uncorrelated fading and [**Bjornson2019c**], [**Bjornson2020a**] considered $N \geq 1$ with correlated fading. In the extreme case when all APs are serving UE k (i.e., $\mathbf{D}_k = \mathbf{I}_{LN}$), the SINR in (5.5) can be simplified to

$$\mathsf{SINR}_k^{(\mathrm{ul,c})} = \frac{p_k \left| \mathbf{v}_k^{\mathrm{H}} \widehat{\mathbf{h}}_k \right|^2}{\sum\limits_{\substack{i=1 \\ i \neq k}}^{K} p_i \left| \mathbf{v}_k^{\mathrm{H}} \widehat{\mathbf{h}}_i \right|^2 + \mathbf{v}_k^{\mathrm{H}} \left(\sum\limits_{i=1}^{K} p_i \mathbf{C}_i + \sigma_{\mathrm{ul}}^2 \mathbf{I}_{LN} \right) \mathbf{v}_k}. \tag{5.7}$$

5.1.2 An Alternative Spectral Efficiency Expression

The SE expression in Theorem 5.1 is a lower bound on the capacity that is derived by utilizing the property that the channel estimate $\widehat{\mathbf{h}}_k$ and the estimation error $\widetilde{\mathbf{h}}_k = \mathbf{h}_k - \widehat{\mathbf{h}}_k$ are independent random vectors. This condition is satisfied when using the MMSE estimator and Rayleigh fading channels as assumed in this monograph. An alternative bound that can be applied along with any channel estimator and fading model is the so-called *use-and-then-forget (UatF)* that is widely used in Cellular Massive MIMO [**massivemimobook**], and also in Cell-free Massive MIMO [**Bashar2019a**], [**Bjornson2019c**], [**Nayebi2016a**] with $\mathbf{D}_{kl} = \mathbf{I}_N$ for all k, l and specific combining vectors. The name comes from the fact that the channel estimates are used for computing the receive combining vectors and then effectively "forgotten" before the signal detection takes place. In the considered network setting with multiple-antenna APs, arbitrary correlated fading channels, and arbitrary DCC, the following achievable SE is found by applying the UatF bound.

Theorem 5.2. An achievable SE of UE k in the centralized operation is

$$\mathsf{SE}_k^{(\mathsf{ul,c-UatF})} = \frac{\tau_u}{\tau_c} \log_2 \left(1 + \mathsf{SINR}_k^{(\mathsf{ul,c-UatF})}\right) \quad \text{bit/s/Hz} \tag{5.8}$$

where the effective SINR is given by

$$\mathsf{SINR}_k^{(\mathsf{ul,c-UatF})} = \frac{p_k \left|\mathbb{E}\left\{\mathbf{v}_k^{\mathrm{H}} \mathbf{D}_k \mathbf{h}_k\right\}\right|^2}{\sum\limits_{i=1}^{K} p_i \mathbb{E}\left\{\left|\mathbf{v}_k^{\mathrm{H}} \mathbf{D}_k \mathbf{h}_i\right|^2\right\} - p_k \left|\mathbb{E}\left\{\mathbf{v}_k^{\mathrm{H}} \mathbf{D}_k \mathbf{h}_k\right\}\right|^2 + \sigma_{\mathrm{ul}}^2 \mathbb{E}\left\{\left\|\mathbf{D}_k \mathbf{v}_k\right\|^2\right\}} \tag{5.9}$$

and the expectation is with respect to the channel realizations.

Proof. The proof is available in Appendix C.2.2 on p. 450. □

Intuitively, $\mathsf{SE}_k^{(\mathsf{ul,c-UatF})}$ in (5.8) should be smaller than $\mathsf{SE}_k^{(\mathsf{ul,c})}$ in (5.4) since it relies on a simplified implementation in which the channel estimates are not used at the CPU for signal detection. This conjecture is hard to prove analytically but has been verified by numerous numerical

experiments. Except for Section 5.4 where will compare $\mathsf{SE}_k^{(\mathsf{ul,c-UatF})}$ with $\mathsf{SE}_k^{(\mathsf{ul,c})}$, we will not use $\mathsf{SE}_k^{(\mathsf{ul,c-UatF})}$ in the remainder of this section, to avoid underestimating the achievable performance. However, we stress that one benefit with the simplified bound in Theorem 5.2 is that there is no expectation in front of the logarithm, because the bounding technique is treating the channel as deterministic. We will return to the expression in Section 7.1.1 on p. 394 where the uplink powers $\{p_k : k = 1, \ldots, K\}$ will be optimized by making use of the fact that there are only expectations at the inside of the logarithm.

5.1.3 Optimal Receive Combining

The SE expression in (5.4) holds for any receive combining vector $\mathbf{v}_k$ but we would like to use the one that maximizes its value. We first note that we can multiply $\mathbf{D}_k\mathbf{v}_k$ with any non-zero scalar without affecting the SINR value, since all the terms in the ratio are scaled equivalently. Hence, the goal of designing the receive combining is not to find the right length of the vector $\mathbf{D}_k\mathbf{v}_k$ but the right direction in the vector space. We notice that (5.5) has the form of a generalized Rayleigh quotient:

$$\mathsf{SINR}_k^{(\mathsf{ul,c})} = \frac{p_k \left| \mathbf{v}_k^{\mathrm{H}} \mathbf{D}_k \widehat{\mathbf{h}}_k \right|^2}{\mathbf{v}_k^{\mathrm{H}} \left(\sum\limits_{\substack{i=1 \\ i \neq k}}^{K} p_i \mathbf{D}_k \widehat{\mathbf{h}}_i \widehat{\mathbf{h}}_i^{\mathrm{H}} \mathbf{D}_k + \mathbf{Z}_k + \sigma_{\mathrm{ul}}^2 \mathbf{D}_k \right) \mathbf{v}_k} \tag{5.10}$$

where we have replaced $\mathbf{D}_k^2$ by $\mathbf{D}_k$ in the last term of the denominator by using the relation $\mathbf{D}_k^2 = \mathbf{D}_k$. Hence, the SINR can be maximized as described in Lemma 3.7 on p. 250. This leads to the following optimal combining vector.

Corollary 5.3. The instantaneous SINR in (5.5) for UE k is maximized by the MMSE combining vector

$$\mathbf{v}_k^{\mathrm{MMSE}} = p_k \left(\sum_{i=1}^{K} p_i \mathbf{D}_k (\widehat{\mathbf{h}}_i \widehat{\mathbf{h}}_i^{\mathrm{H}} + \mathbf{C}_i) \mathbf{D}_k + \sigma_{\mathrm{ul}}^2 \mathbf{I}_{LN} \right)^{-1} \mathbf{D}_k \widehat{\mathbf{h}}_k \tag{5.11}$$

which leads to the maximum value

$$\mathsf{SINR}_k^{(\mathrm{ul,c})} = p_k \widehat{\mathbf{h}}_k^{\mathrm{H}} \mathbf{D}_k \left(\sum_{\substack{i=1 \\ i \neq k}}^{K} p_i \mathbf{D}_k \widehat{\mathbf{h}}_i \widehat{\mathbf{h}}_i^{\mathrm{H}} \mathbf{D}_k + \mathbf{Z}_k + \sigma_{\mathrm{ul}}^2 \mathbf{I}_{LN} \right)^{-1} \mathbf{D}_k \widehat{\mathbf{h}}_k. \quad (5.12)$$

Proof. We can maximize (5.5) by using Lemma 3.7 on p. 250 with $\mathbf{v} = \mathbf{v}_k$, $\mathbf{h} = \sqrt{p_k} \mathbf{D}_k \widehat{\mathbf{h}}_k$, and $\mathbf{B} = \sum_{i=1, i \neq k}^{K} p_i \mathbf{D}_k \widehat{\mathbf{h}}_i \widehat{\mathbf{h}}_i^{\mathrm{H}} \mathbf{D}_k + \mathbf{Z}_k + \sigma_{\mathrm{ul}}^2 \mathbf{D}_k$. To handle that $\mathbf{B}$ might not be positive definite (but only positive semi-definite), we can regularize it by replacing $\sigma_{\mathrm{ul}}^2 \mathbf{D}_k$ with $\sigma_{\mathrm{ul}}^2 \mathbf{I}_{LN}$. This makes $\mathbf{B}$ invertible and will not affect the final result since the inverse is then multiplied with $\mathbf{D}_k$, which removes the extra noise that was added to the unused dimensions. □

The optimal combining vector in (5.11) consists of two parts: 1) the partial estimate $\mathbf{D}_k \widehat{\mathbf{h}}_k$ of the channel to UE k and 2) the inverse of a matrix that equals $\mathbb{E}\{\mathbf{D}_k \mathbf{y}^{\mathrm{ul}} (\mathbf{y}^{\mathrm{ul}})^{\mathrm{H}} \mathbf{D}_k | \{\widehat{\mathbf{h}}_i\}\}$ (plus the regularization term that is added to the unused dimensions), which is the conditional correlation matrix of the received signal, given the channel estimates. The first part identifies the combining vector that would maximize the desired signal power, while the second part rotates that vector to strike a balance between achieving a strong signal and suppressing interference. Note that the matrix inverse acts as a spatial whitening filter, as exemplified in Figure 5.2. The left graph in this figure shows how the received power of the desired signal and interference are arriving from different azimuth angles, while the noise is equally distributed over all angles. The goal of the receive combining is to identify the dashed direction where the SINR is maximized, which is clearly different from the 30° direction where the signal is the strongest. By applying a spatial whitening filter, we obtain the graph at the right where the sum of the signal, interference, and noise powers is constant in all directions. We can then easily maximize the SINR by selecting the "direction" in the transformed domain where the signal is stronger, which automatically implies that the sum of the interference plus noise is at its lowest value.

It can be shown that the combining vector that maximizes the instantaneous SINR also minimizes the MSE in the data detection,

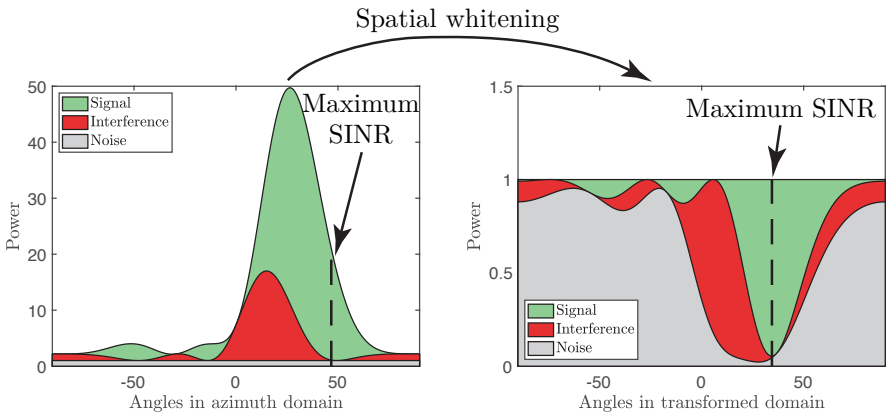


Figure 5.2: This figure exemplifies how MMSE combining in (5.11) is maximizing the effective SINR when using an $N = 4$ antenna AP. The left graph shows how the received signal power, interference, and noise are distributed over different azimuth angles. The curves are placed on top of each other, thus it is the colored height between them that measures the power. The SINR-maximizing direction of the combining vector finds a nontrivial tradeoff between having a strong signal and having low interference. This direction can be easily identified by transforming the received vector using a spatial whitening filter based on the conditional correlation matrix $\mathbb{E}\{\mathbf{D}_k \mathbf{y}^{\mathrm{ul}} (\mathbf{y}^{\mathrm{ul}})^{\mathrm{H}} \mathbf{D}_k | \{\widehat{\mathbf{h}}_i\}\}$. This results in the graph to the right where the sum of the signal, interference, and noise is the same in all directions. The maximum SINR is then achieved when the signal power is maximized. By transforming this signal back to the original domain, the MMSE combining is obtained.

which is defined as

$$\mathrm{MSE}_k = \mathbb{E}\left\{ |s_k - \widehat{s}_k|^2 \,|\, \left\{ \widehat{\mathbf{h}}_i : i = 1, \ldots, K \right\} \right\}. \tag{5.13}$$

This is the conditional MSE between the true data signal s_k and the estimate $\widehat{s}_k$ of it from (5.3); see [**massivemimobook**] for a deeper discussion. This is the reason why $\mathbf{v}_k^{\mathrm{MMSE}}$ is generally called *MMSE combining*. This type of receive combining maximizes the mutual information of many types of channels with multiple receive antennas [**Park2013a**]. However, the particular expression in (5.11) is unique for Cell-free Massive MIMO with the centralized operation and the DCC framework.

Note that $\mathbf{D}_k \widehat{\mathbf{h}}_i$ in (5.11) is always non-zero except when $\mathbf{D}_k = \mathbf{0}_{LN \times LN}$, which is an uninteresting special case since the CPU only needs to apply receive combining when at least one AP serves the UE. This implies that the CPU needs to compute all the K MMSE

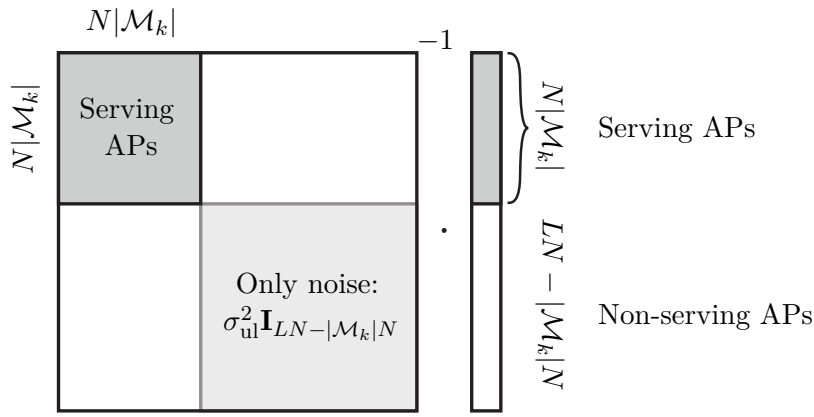


Figure 5.3: Only the serving APs need to be considered when computing the MMSE combining vector $\mathbf{v}_k^{\text{MMSE}}$ of UE k in (5.11). By ordering the APs so that the $|\mathcal{M}_k|$ serving ones are listed first followed by the $L - |\mathcal{M}_k|$ non-serving APs, a block-diagonal matrix needs to be inverted. Only the shaded upper block of dimension $N|\mathcal{M}_k| \times N|\mathcal{M}_k|$ needs to be inverted and multiplied with a vector of length $N|\mathcal{M}_k|$. The computational complexity is identical to the case of having only $|\mathcal{M}_k|$ APs.

channel estimates $\{\widehat{\mathbf{h}}_{il} : i = 1, \ldots, K\}$ corresponding to any AP l that is serving UE k (i.e., APs with index $l \in \mathcal{M}_k$). The total number of complex multiplications required for channel estimation is reported in Table 5.1 and is computed as discussed above (4.11). In addition, we need to account for the complexity of computing the combining vector $\mathbf{v}_k^{\text{MMSE}}$ in (5.11) for $k = 1, \ldots, K$ once per coherence block. This complexity can be computed using the framework described in [**massivemimobook**] and exploiting the fact that only a subset of the APs takes part in estimating the transmitted signal s_k; see Remark 5.1 below for further details. Table 5.1 summarizes the total number of complex multiplications required for the computation of the MMSE combining vector of a generic UE k. Since this number grows linearly with K, the optimal MMSE combining makes the network unscalable according to Definition 2.2 on p. 217. We will, therefore, look for alternatives.

Remark 5.1 (Computation of MMSE combining)**.** The complexity of computing the MMSE combining vector of UE k in (5.11) depends on the number of serving APs $|\mathcal{M}_k|$, not the total number of APs. To

observe this, recall that $\mathbf{D}_k = \text{diag}(\mathbf{D}_{k1}, \ldots, \mathbf{D}_{kL})$ is a block-diagonal matrix with $\mathbf{D}_{kl}$ being $\mathbf{I}_N$ if AP l serves UE k (i.e., $l \in \mathcal{M}_k$) and zero otherwise. Therefore, the elements of $\{\widehat{\mathbf{h}}_i : i = 1, \ldots, K\}$ and $\mathbf{Z}_k$ whose indices correspond to the zeros of $\mathbf{D}_k$ do not affect any term in (5.11). In particular, suppose the APs with index $1, \ldots, |\mathcal{M}_k|$ are the serving ones. The computations will then be as shown in Figure 5.3, where only the shaded parts are non-zero. Only the upper block of size $N|\mathcal{M}_k| \times N|\mathcal{M}_k|$ needs to be inverted and then multiplied with a vector of length $N|\mathcal{M}_k|$. The total number of complex multiplications required for these two operations can be obtained by using the framework described in [**massivemimobook**] and is summarized in Table 5.1. If the $|\mathcal{M}_k|$ serving APs have arbitrary indices, the computational complexity remains the same since the implementation can simply reorder the APs so that the serving ones are listed first.

5.1.4 Scalable Combining Schemes for Centralized Operation

We will now develop centralized receive combining schemes that are scalable, in the sense that the computational complexity per UE is independent of K. The simplest solution is to use MR combining with

$$\mathbf{v}_k^{\text{MR}} = \mathbf{D}_k \widehat{\mathbf{h}}_k \tag{5.14}$$

which maximizes the numerator $|\mathbf{v}_k^{\text{H}} \mathbf{D}_k \widehat{\mathbf{h}}_k|^2$ of the effective SINR in (5.5). In other words, the power of the desired signal is maximized while the existence of interference and estimation errors is ignored. Since the MMSE channel estimate is directly used as the MR combining vector, the computational complexity is equal to that of obtaining the partial MMSE channel estimate $\mathbf{D}_k \widehat{\mathbf{h}}_k$. This requires a total of $(N\tau_p + N^2)|\mathcal{M}_k|$ complex multiplications, which is summarized in Table 5.1. Since no additional computations are needed, MR combining is a scalable solution with low complexity. However, we will observe later that the performance can be poor in the presence of interference, since a high degree of favorable propagation cannot be guaranteed between all UEs (see Section 2.6.2 on p. 234). Two scalable combining schemes that are superior to MR, in terms of dealing with interference, are therefore derived next by taking inspiration from the optimal MMSE combining in (5.11).

Scheme	Channel estimation	Combining vector computation
MMSE	$(N\tau_p + N^2)K\lvert\mathcal{M}_k\rvert$	$\frac{(N\lvert\mathcal{M}_k\rvert)^2+N\lvert\mathcal{M}_k\rvert}{2}K + (N\lvert\mathcal{M}_k\rvert)^2 + \frac{(N\lvert\mathcal{M}_k\rvert)^3-N\lvert\mathcal{M}_k\rvert}{3}$
P-MMSE	$(N\tau_p + N^2)\lvert\mathcal{S}_k\rvert\lvert\mathcal{M}_k\rvert$	$\frac{(N\lvert\mathcal{M}_k\rvert)^2+N\lvert\mathcal{M}_k\rvert}{2}\lvert\mathcal{S}_k\rvert + (N\lvert\mathcal{M}_k\rvert)^2 + \frac{(N\lvert\mathcal{M}_k\rvert)^3-N\lvert\mathcal{M}_k\rvert}{3}$
P-RZF	$(N\tau_p + N^2)\lvert\mathcal{S}_k\rvert\lvert\mathcal{M}_k\rvert$	$\frac{\lvert\mathcal{S}_k\rvert^2+\lvert\mathcal{S}_k\rvert}{2}N\lvert\mathcal{M}_k\rvert + \lvert\mathcal{S}_k\rvert^2 + \lvert\mathcal{S}_k\rvert N\lvert\mathcal{M}_k\rvert + \frac{\lvert\mathcal{S}_k\rvert^3-\lvert\mathcal{S}_k\rvert}{3}$
MR	$(N\tau_p + N^2)\lvert\mathcal{M}_k\rvert$	–

Table 5.1: The number of complex multiplications required per coherence block to compute the combining vector of UE k, including the computation of the channel estimates. Different combining schemes for the centralized operation are considered.

Partial MMSE Combining

In a large cell-free network with distributed UEs, it is reasonable to believe that the interference that affects UE k is mainly generated by a small subset of the other UEs, which are located in the neighborhood of UE k. We want to utilize this observation to reduce the computational complexity of the optimal MMSE combining and thereby achieve a scalable alternative. To this end, we assume that only the UEs that are served by partially the same APs as UE k should be included in the inverse matrix in (5.11). These UEs have indices in the set

$$\mathcal{S}_k = \{i : \mathbf{D}_k \mathbf{D}_i \neq \mathbf{0}_{LN \times LN}\}. \tag{5.15}$$

By utilizing $\mathcal{S}_k$, an alternative *partial MMSE (P-MMSE)* scheme can be defined as follows [**Attarifar2020a**], [**Bjornson2020a**], [**Nayebi2016a**]:

$$\mathbf{v}_k^{\mathrm{P-MMSE}} = p_k \left(\sum_{i \in \mathcal{S}_k} p_i \mathbf{D}_k \widehat{\mathbf{h}}_i \widehat{\mathbf{h}}_i^{\mathrm{H}} \mathbf{D}_k + \mathbf{Z}_{\mathcal{S}_k} + \sigma_{\mathrm{ul}}^2 \mathbf{I}_{LN} \right)^{-1} \mathbf{D}_k \widehat{\mathbf{h}}_k \tag{5.16}$$

with

$$\mathbf{Z}_{\mathcal{S}_k} = \sum_{i \in \mathcal{S}_k} p_i \mathbf{D}_k \mathbf{C}_i \mathbf{D}_k. \tag{5.17}$$

This scheme is not designed to be optimal from an SE perspective but as a scalable approximation of the optimal MMSE combining. We note that P-MMSE combining coincides with MMSE combining only when UE k is served by all APs; that is, $\mathbf{D}_k = \mathbf{I}_{LN}$ and $\mathcal{S}_k = \{1, \ldots, K\}$. Another way to view this is that P-MMSE combining would be optimal if only the UEs with indices in $\mathcal{S}_k$ were active. However, MMSE combining and P-MMSE combining are generally different schemes.

In Section 4.4 on p. 286, a DCC selection method was described where each AP serves exactly one UE per pilot sequence, which guarantees $|\mathcal{D}_l| = \tau_p$ for all l. Using this scheme, $|\mathcal{S}_k| = \tau_p$ if all the APs that serve UE k co-serve the same set of UEs. This is the smallest value that $|\mathcal{S}_k|$ can take. Moreover, it holds that $|\mathcal{S}_k| \leq (\tau_p - 1)|\mathcal{M}_k| + 1$, where the upper bound is achieved in the unlikely case that all the APs in $\mathcal{M}_k$ serve UE k but otherwise serve entirely non-overlapping sets of

UEs. Importantly, even this unlikely upper bound is independent of K. The number of complex multiplications required by P-MMSE for channel estimation and combining vector computation is reported in Table 5.1. In computing those numbers, we have taken into account the discussion from Remark 5.1 that only the number of serving APs affects the size of the matrix inverse and matrix-vector multiplication. The exact expression is rather lengthy but, importantly, it is independent of K, which makes P-MMSE a scalable combining scheme that supports an arbitrarily large number of UEs.

In practice, some fine-tuning can be done to improve the performance of P-MMSE by including additional UEs in $\mathcal{S}_k$ to deal with strongly interfering UEs that are only served by other APs. A proper selection of these extra UEs will increase network performance without violating the scalability. However, in this monograph, we stick to the definition of $\mathcal{S}_k$ in (5.15) for brevity.

Partial Regularized Zero-Forcing Combining

The complexity of computing the P-MMSE combining vector scales with $(N|\mathcal{M}_k|)^3$. Each AP is supposed to have a relatively small number of antennas N in cell-free networks, but $|\mathcal{M}_k|$ might be large when there are many APs in the vicinity of UE k. Since the combining vector is computed at the CPU, which is assumed to have high computational capability, the complexity might not be an issue. However, it is nevertheless interesting to explore if a further complexity reduction can be achieved without affecting the performance to any great extent. We will therefore present an alternative receive combining scheme, which is obtained as a simplification of P-MMSE. To this end, we note that if the channel conditions of the interfering UEs in $\mathcal{S}_k$ are good, all the corresponding estimation error correlation matrices $\{\mathbf{C}_i : i \in \mathcal{S}_k\}$ in $\mathbf{Z}_{\mathcal{S}_k}$ will be small. If we neglect $\mathbf{Z}_{\mathcal{S}_k}$ in (5.16), we obtain

$$\mathbf{v}_k^{\mathrm{P-RZF}} = p_k \left(\sum_{i \in \mathcal{S}_k} p_i \mathbf{D}_k \widehat{\mathbf{h}}_i \widehat{\mathbf{h}}_i^{\mathrm{H}} \mathbf{D}_k + \sigma_{\mathrm{ul}}^2 \mathbf{I}_{LN} \right)^{-1} \mathbf{D}_k \widehat{\mathbf{h}}_k. \tag{5.18}$$

This change has a negligible impact on the complexity but it enables further reformulation. Let us construct $\widehat{\mathbf{H}}_{\mathcal{S}_k} \in \mathbb{C}^{LN \times |\mathcal{S}_k|}$ by stacking

all the vectors $\widehat{\mathbf{h}}_i$ with indices $i \in \mathcal{S}_k$ with the first column being $\widehat{\mathbf{h}}_k$. We further define $\mathbf{P}_{\mathcal{S}_k} \in \mathbb{R}^{|\mathcal{S}_k| \times |\mathcal{S}_k|}$ as a diagonal matrix containing the transmit powers p_i for $i \in \mathcal{S}_k$, listed in the same order as the columns of $\widehat{\mathbf{H}}_{\mathcal{S}_k}$. By letting $[\cdot]_{:,1}$ denote the operation of only keeping the first column of its matrix argument, we can reformulate (5.18) as

$$\begin{aligned} \mathbf{v}_k^{\mathrm{P-RZF}} &= \left[\left(\mathbf{D}_k \widehat{\mathbf{H}}_{\mathcal{S}_k} \mathbf{P}_{\mathcal{S}_k} \widehat{\mathbf{H}}_{\mathcal{S}_k}^{\mathrm{H}} \mathbf{D}_k + \sigma_{\mathrm{ul}}^2 \mathbf{I}_{LN} \right)^{-1} \mathbf{D}_k \widehat{\mathbf{H}}_{\mathcal{S}_k} \mathbf{P}_{\mathcal{S}_k} \right]_{:,1} \\ &= \left[\mathbf{D}_k \widehat{\mathbf{H}}_{\mathcal{S}_k} \mathbf{P}_{\mathcal{S}_k} \left(\widehat{\mathbf{H}}_{\mathcal{S}_k}^{\mathrm{H}} \mathbf{D}_k \mathbf{D}_k \widehat{\mathbf{H}}_{\mathcal{S}_k} \mathbf{P}_{\mathcal{S}_k} + \sigma_{\mathrm{ul}}^2 \mathbf{I}_{|\mathcal{S}_k|} \right)^{-1} \right]_{:,1} \\ &= \left[\mathbf{D}_k \widehat{\mathbf{H}}_{\mathcal{S}_k} \left(\widehat{\mathbf{H}}_{\mathcal{S}_k}^{\mathrm{H}} \mathbf{D}_k \widehat{\mathbf{H}}_{\mathcal{S}_k} + \sigma_{\mathrm{ul}}^2 \mathbf{P}_{\mathcal{S}_k}^{-1} \right)^{-1} \right]_{:,1} \end{aligned} \tag{5.19}$$

where the second equality follows from the first identity of Lemma B.2 on p. 445 and the third equality utilizes the fact that $\mathbf{D}_k \mathbf{D}_k = \mathbf{D}_k$.

We call (5.19) *partial regularized zero-forcing (P-RZF) combining*. This name indicates that the matrix expression in (5.19) has the same form as the pseudo-inverse $\mathbf{D}_k \widehat{\mathbf{H}}_{\mathcal{S}_k} (\widehat{\mathbf{H}}_{\mathcal{S}_k}^{\mathrm{H}} \mathbf{D}_k \widehat{\mathbf{H}}_{\mathcal{S}_k})^{-1}$ to the partial channel matrix $\widehat{\mathbf{H}}_{\mathcal{S}_k}^{\mathrm{H}} \mathbf{D}_k$. The only difference is that the inverse has been regularized by adding the matrix $\sigma_{\mathrm{ul}}^2 \mathbf{P}_{\mathcal{S}_k}^{-1}$ containing the noise variance and transmit powers. Combining vectors based on pseudo-inverses force the interference between the UEs to zero, but might lead to large losses in the desired signal power when the UEs have similar channels. P-RZF is balancing between suppressing interference and maintaining strong desired signal powers.

The total number of complex multiplications required for channel estimation and combining vector computation with P-RZF is reported in Table 5.1. As expected, this number is independent of K. Hence, this scheme is also scalable. The main benefit over the P-MMSE is that an $|\mathcal{S}_k| \times |\mathcal{S}_k|$ matrix is inverted in (5.19) instead of an $N|\mathcal{M}_k| \times N|\mathcal{M}_k|$ matrix as in (5.16), which can substantially reduce the complexity since we typically have $N|\mathcal{M}_k| \geq |\mathcal{S}_k|$ in cell-free networks. Most likely, this benefit comes with an SE loss since, in general, the channel conditions will not be so good to every interfering UE so that the estimation errors can be ignored. The tradeoff will be evaluated later.

5.1.5 Fronthaul Signaling Load for Centralized Operation

We wrap up the analysis of the centralized operation by quantifying the required fronthaul signaling. We will do this by counting the number of complex scalars that must be sent over the fronthaul links. In practice, these scalars will have to be quantized and different types of information can be quantized using a different number of bits per scalar. However, if all the scalars are sent over the fronthaul using the same bit precision, counting the number of scalars is sufficient to compare the signal load of different methods.

In each coherence block, AP l needs to send $\tau_p N$ complex scalars representing the pilot signals $\{\mathbf{y}_{tl}^{\text{pilot}} : t = 1, \ldots, \tau_p\}$ and $\tau_u N$ complex scalars representing the received data signal $\mathbf{y}_l^{\text{ul}}$. As summarized in Table 5.2, this requires a total of $(\tau_p + \tau_u)NL$ complex scalars, irrespective of the combining scheme used for detection at the CPU.[1] Since this value does not grow with K, we can conclude that the fronthaul signaling is scalable.

Notice that the implementations of the MMSE channel estimator and the P-MMSE combining scheme require knowledge of the spatial correlation matrices $\{\mathbf{R}_{kl} : k = 1, \ldots, K, l = 1, \ldots, L\}$. In a centralized network, where the APs only act as relays, this information can be acquired directly at the CPU by using one of the pilot signaling methods that already exist in Cellular Massive MIMO literature (see [**Sanguinetti2019a**] for an overview). Hence, there is no need to exchange the spatial correlation matrices through the fronthaul links. As discussed later in Section 5.2.3, the situation is different in the distributed implementation presented next, where the pilot signals are not sent to the CPU but used locally at the APs for channel estimation. Statistical information must then be acquired at the APs and gathered at the CPU using the fronthaul.

[1] Note that $\tau_p N$ is an upper bound on the fronthaul signaling load for pilot signals, which is achieved in the case that AP l assigns all the τ_p available pilots to the UEs that it is serving. This was assumed by the pilot assignment scheme presented in Section 4.4 on p. 286, but is not the case in general. In other cases, AP l exchanges a lower number of complex scalars for the pilot signals through the fronthaul links.

Table 5.2: The number of complex scalars to send from the APs to the CPU over the fronthaul either in each coherence block or for each realization of the channel statistics (e.g., user locations). Both centralized and distributed operation are considered.

Implementation	Per coherence block	Statistics
Centralized	$(\tau_p + \tau_u)NL$	$--$
Distributed: opt LSFD	$\tau_u \sum_{l=1}^{L} \lvert\mathcal{D}_l\rvert$	$\frac{3K+1}{2} \sum_{l=1}^{L} \lvert\mathcal{D}_l\rvert$
Distributed: n-opt LSFD	$\tau_u \sum_{l=1}^{L} \lvert\mathcal{D}_l\rvert$	$\sum_{l=1}^{L} \sum_{k\in\mathcal{D}_l} \frac{3\lvert\mathcal{S}_k\rvert+1}{2}$
Distributed: no LSFD	$\tau_u \sum_{l=1}^{L} \lvert\mathcal{D}_l\rvert$	$--$

5.2 Distributed Uplink Operation

Instead of entirely delegating the channel estimation and data detection to the CPU, a User-centric Cell-free Massive MIMO network can be implemented to carry out as much of the signal processing as possible in a distributed manner. A motivating factor is that each AP can easily be equipped with a baseband processor that is (at least) as powerful as those in the UEs, thus algorithms that can make efficient use of it should be developed. The goal is to enable an ultra-dense deployment with $L \gg K$, where the CPU's capabilities are dimensioned based on the number of UEs, but largely unaffected by the number of APs. In other words, we should be able to add new APs to the network without having to upgrade the CPUs, thanks to the property that each AP comes with a local processor. Another reason is to reduce the fronthaul signaling compared to the centralized operation and to avoid the fronthaul quantization distortion, which exists in practice but is neglected in this monograph. Since a key property of cell-free networks is that multiple APs are involved in the service of each UE, the final data detection must be carried out at a point where the inputs from multiple APs are combined. Recall that the CPU is a logical entity so the unavoidable CPU tasks can be physically assigned to a nearby edge-cloud processor (or even one of the serving APs), so there is no need for a physical central unit. The following two-stage procedure is called *distributed operation* in this monograph and is schematically

described in Figure 5.1(b).

In the first stage, each AP l uses its received pilot signals $\{\mathbf{y}_{tl}^{\text{pilot}} : t = 1, \ldots, \tau_p\}$ to locally estimate the channels $\{\widehat{\mathbf{h}}_{il} : i \in \mathcal{D}_l\}$. For each UE k, the AP can then use the local estimates to select a local receive combining vector $\mathbf{v}_{kl}$. According to (2.5), the AP then computes its local estimate of s_k as

$$\widehat{s}_{kl} = \mathbf{v}_{kl}^{\text{H}} \mathbf{D}_{kl} \mathbf{y}_l^{\text{ul}}. \tag{5.20}$$

Note that this estimate is zero for $k \notin \mathcal{D}_l$, thus AP l only needs to compute (5.20) for the UEs that is serves.

In the second stage, the local data estimates of all APs are gathered at the CPU where they are combined/fused into a final estimate of the UE data. The CPU computes its estimate as a linear combination of the local estimates:

$$\widehat{s}_k = \sum_{l=1}^{L} a_{kl}^{\star} \widehat{s}_{kl} = \sum_{l=1}^{L} a_{kl}^{\star} \mathbf{v}_{kl}^{\text{H}} \mathbf{D}_{kl} \mathbf{y}_l^{\text{ul}} \tag{5.21}$$

where $a_{kl} \in \mathbb{C}$ is the weight that the CPU assigns to the local signal estimate that AP l has of the signal from UE k. Note that we include all the AP indices since $\widehat{s}_{kl} = 0$ for $l \notin \mathcal{M}_k$ and the results do not change. To limit the fronthaul signaling, the APs are only sending the local data estimates to the CPU and not the channel estimates. Hence, the CPU needs to select the weights $\{a_{kl} : k = 1, \ldots, K, l \in \mathcal{M}_k\}$ as a deterministic function of the channel statistics. Intuitively, the CPU should assign a larger weight to an AP having a large SNR to the UE than an AP having a small SNR, but also the interference situation and receive combining scheme should be taken into consideration.

The structure of the second stage resembles what is called *large-scale fading decoding (LSFD)* and was originally used in Cellular Massive MIMO [**Adhikary2017a**], [**ashikhmin2012pilot**]. Unlike the centralized operation, the two-stage procedure outlined above allows the network to make use of the distributed processing capabilities of the APs. The received signals are primarily processed at the points where they were received. Even if it the operation is not entirely distributed, it is as distributed as a cell-free network, utilizing coherent cooperation

between APs, can become since only the final signal detection is done at the CPU.

The achievable SE for arbitrary local receive combining $\{\mathbf{v}_{kl} : k = 1, \ldots, K, l \in \mathcal{M}_k\}$ and LSFD weights will be derived next. After that, we will show how to make a judicious selection of the local combining vectors and how to compute the optimal LSFD weights. Since the preferred solutions turn out to be unscalable for networks with many UEs, we will provide scalable alternatives.

5.2.1 Spectral Efficiency and Optimal LSFD Weights

We will now compute an SE expression for UE k in the distributed operation. Plugging (2.3) into (5.21) yields

$$\widehat{s}_k = \left(\sum_{l=1}^{L} a_{kl}^{\star} \mathbf{v}_{kl}^{\mathrm{H}} \mathbf{D}_{kl} \mathbf{h}_{kl} \right) s_k + \sum_{\substack{i=1 \\ i \neq k}}^{K} \left(\sum_{l=1}^{L} a_{kl}^{\star} \mathbf{v}_{kl}^{\mathrm{H}} \mathbf{D}_{kl} \mathbf{h}_{il} \right) s_i + n'_k \tag{5.22}$$

where $n'_k = \sum_{l=1}^{L} a_{kl}^{\star} \mathbf{v}_{kl}^{\mathrm{H}} \mathbf{D}_{kl} \mathbf{n}_l$ represents the noise. For brevity, we define the vector $\mathbf{g}_{ki} \in \mathbb{C}^L$ computed as

$$\mathbf{g}_{ki} = \begin{bmatrix} \mathbf{v}_{k1}^{\mathrm{H}} \mathbf{D}_{k1} \mathbf{h}_{i1} \\ \vdots \\ \mathbf{v}_{kL}^{\mathrm{H}} \mathbf{D}_{kL} \mathbf{h}_{iL} \end{bmatrix}. \tag{5.23}$$

This is the vector with the receive-combined channels between UE i and all APs that serve UE k. Using this notation, (5.22) can be expressed as

$$\widehat{s}_k = \mathbf{a}_k^{\mathrm{H}} \mathbf{g}_{kk} s_k + \sum_{\substack{i=1 \\ i \neq k}}^{K} \mathbf{a}_k^{\mathrm{H}} \mathbf{g}_{ki} s_i + n'_k \tag{5.24}$$

where $\mathbf{a}_k = [a_{k1} \ \ldots \ a_{kL}]^{\mathrm{T}} \in \mathbb{C}^L$ is the LSFD weight vector of UE k. The realizations of $\mathbf{g}_{ki}$ change in every coherence block, while $\mathbf{a}_k$ is supposed to be a deterministic vector that the CPU can select without knowing the channel realizations. We notice that (5.24) has the structure of a single-antenna channel where $\{\mathbf{a}_k^{\mathrm{H}} \mathbf{g}_{ki} : i = 1, \ldots, K\}$ represent the effective channels of the different UEs. Although the effective channel $\mathbf{a}_k^{\mathrm{H}} \mathbf{g}_{kk}$ is unknown at the CPU, we notice that its average $\mathbb{E}\{\mathbf{a}_k^{\mathrm{H}} \mathbf{g}_{kk}\} = \mathbf{a}_k^{\mathrm{H}} \mathbb{E}\{\mathbf{g}_{kk}\}$

is deterministic and non-zero if the receive combining is properly selected. Therefore, it can be assumed known[2] and used to compute the following achievable SE.

Theorem 5.4. An achievable SE of UE k in the distributed operation is

$$\mathsf{SE}_k^{(\mathrm{ul,d})} = \frac{\tau_u}{\tau_c} \log_2 \left(1 + \mathsf{SINR}_k^{(\mathrm{ul,d})} \right) \quad \text{bit/s/Hz} \tag{5.25}$$

where the effective SINR is given by

$$\mathsf{SINR}_k^{(\mathrm{ul,d})} = \frac{p_k \left| \mathbf{a}_k^{\mathrm{H}} \mathbb{E}\{\mathbf{g}_{kk}\} \right|^2}{\mathbf{a}_k^{\mathrm{H}} \left(\sum\limits_{i=1}^{K} p_i \mathbb{E}\{\mathbf{g}_{ki} \mathbf{g}_{ki}^{\mathrm{H}}\} - p_k \mathbb{E}\{\mathbf{g}_{kk}\} \mathbb{E}\{\mathbf{g}_{kk}^{\mathrm{H}}\} + \mathbf{F}_k \right) \mathbf{a}_k} \tag{5.26}$$

and

$$\mathbf{F}_k = \sigma_{\mathrm{ul}}^2 \mathrm{diag} \left(\mathbb{E} \left\{ \left\| \mathbf{D}_{k1} \mathbf{v}_{k1} \right\|^2 \right\}, \ldots, \mathbb{E} \left\{ \left\| \mathbf{D}_{kL} \mathbf{v}_{kL} \right\|^2 \right\} \right) \in \mathbb{R}^{L \times L}. \tag{5.27}$$

Proof. The proof is given in Appendix C.2.3 on p. 451. □

The pre-log factor τ_u/τ_c in (5.25) is the fraction of each coherence block that is used for uplink data transmission. The term $\mathsf{SINR}_k^{(\mathrm{ul,d})}$ takes the form of an effective SINR and is the ratio of the signal power term $p_k \left| \mathbf{a}_k^{\mathrm{H}} \mathbb{E}\{\mathbf{g}_{kk}\} \right|^2$ and a term containing noise and interference (including some self-interference from the desired signal due to the imperfect channel knowledge). Recall that the word "effective" refers to that this SE is effectively the same as that of a deterministic single-antenna single-user channel where (5.26) is the SNR and the receiver has perfect CSI. In particular, this means that the desired data signal can be encoded and the received signal can be decoded as if we would communicate over such an AWGN channel.

The SE expression in (5.25) is very general. Although this monograph considers correlated Rayleigh fading channels and MMSE channel

[2]When dealing with ergodic capacities, all deterministic parameters can be assumed known without loss of generality, because these can be estimated using a finite number of transmission resources, while the capacity is only achieved as the amount of transmission resources goes to infinity. Hence, the estimation overhead for obtaining deterministic parameters is negligible.

estimation, (5.25) is actually a valid SE for any channel estimator and channel model, in contrast to the SE expression in (5.4) which is explicitly using those assumptions. This was previously discussed in Section 5.1.2.

The achievable SE in (5.25) holds for any local receive combining and LSFD vector $\mathbf{a}_k$, but we are naturally seeking a combination that makes the SE as high as possible. We start with considering the local receive combining vectors. Since AP l locally selects $\mathbf{v}_{kl}$ for $k \in \mathcal{D}_l$, without receiving input from other APs, there is no way for its combining scheme to be optimal from a network-wide perspective. However, the AP can do as well as it can by optimizing a local performance metric. Recall from Section 5.1.3 that the optimal receive combining in a centralized implementation minimizes the MSE in the data detection. In analogy with this result, we can let AP l select the receive combining giving the best local estimate $\widehat{s}_{kl} = \mathbf{v}_{kl}^{\mathrm{H}} \mathbf{D}_{kl} \mathbf{y}_l^{\mathrm{ul}}$ in terms of minimal conditional MSE:

$$\mathbb{E}\left\{ \left| s_k - \mathbf{v}_{kl}^{\mathrm{H}} \mathbf{D}_{kl} \mathbf{y}_l^{\mathrm{ul}} \right|^2 \,\middle|\, \left\{ \widehat{\mathbf{h}}_{il} : i = 1, \ldots, K \right\} \right\}. \tag{5.28}$$

The combining vector $\mathbf{v}_{kl}$ that minimizes (5.28) is

$$\mathbf{v}_{kl}^{\mathrm{L-MMSE}} = p_k \left(\sum_{i=1}^{K} p_i \left(\widehat{\mathbf{h}}_{il} \widehat{\mathbf{h}}_{il}^{\mathrm{H}} + \mathbf{C}_{il} \right) + \sigma_{\mathrm{ul}}^2 \mathbf{I}_N \right)^{-1} \mathbf{D}_{kl} \widehat{\mathbf{h}}_{kl} \tag{5.29}$$

where $\mathbf{C}_{il}$ is the error correlation matrix given in (4.9). This can be proved by computing the conditional expectation in (5.28) and equating its first derivative with respect to $\mathbf{v}_{kl}$ to zero. One can also show that (5.29) is the combining vector that would maximize the SE if AP l detected the data signal s_k locally. We call (5.29) *Local MMSE (L-MMSE) combining* to distinguish it from the MMSE combining in (5.11), which is applied at the CPU in the centralized operation. L-MMSE and MMSE combining coincide (except for that the latter is padded with zeros) in the special case when only AP l serves UE k (i.e., $\mathcal{M}_k = \{l\}$). Hence, the computational complexity of L-MMSE is the same as for MMSE combining in Table 5.1 but with $|\mathcal{M}_k| = 1$. Since this complexity grows with K, we notice that L-MMSE combining is not a scalable combining scheme.

We now shift focus to the LSFD weights $\{\mathbf{a}_k : k = 1, \ldots, K\}$. For any given local receive combining vectors, $\mathbf{a}_k$ can be optimized by the CPU to maximize the SE. Since it is a deterministic vector, it is optimized as a function of the channel statistics. We notice that (5.26) is a generalized Rayleigh quotient with respect to $\mathbf{a}_k$. Hence, we can maximize the effective SINR by using Lemma 3.7 on p. 250. The optimal $\mathbf{a}_k$ is thus obtained as follows.

Corollary 5.5. The effective SINR in (5.26) for UE k is maximized by

$$\mathbf{a}_k^{\text{opt}} = p_k \left(\sum_{i=1}^{K} p_i \mathbb{E}\left\{ \mathbf{g}_{ki} \mathbf{g}_{ki}^{\text{H}} \right\} + \mathbf{F}_k + \widetilde{\mathbf{D}}_k \right)^{-1} \mathbb{E}\left\{ \mathbf{g}_{kk} \right\} \tag{5.30}$$

where $\widetilde{\mathbf{D}}_k \in \mathbb{R}^{L \times L}$ is the diagonal matrix with the (l, l)th element being one if $l \notin \mathcal{M}_k$ and zero otherwise. This leads to the maximum value

$$\mathsf{SINR}_k^{(\mathsf{ul},\mathsf{d})} = p_k \mathbb{E}\left\{ \mathbf{g}_{kk}^{\text{H}} \right\} \left(\sum_{i=1}^{K} p_i \mathbb{E}\left\{ \mathbf{g}_{ki} \mathbf{g}_{ki}^{\text{H}} \right\} - p_k \mathbb{E}\left\{ \mathbf{g}_{kk} \right\} \mathbb{E}\left\{ \mathbf{g}_{kk}^{\text{H}} \right\} + \mathbf{F}_k + \widetilde{\mathbf{D}}_k \right)^{-1} \mathbb{E}\left\{ \mathbf{g}_{kk} \right\}. \tag{5.31}$$

Proof. We can maximize (5.26) by using Lemma 3.7 on p. 250 with $\mathbf{v} = \mathbf{a}_k$, $\mathbf{h} = \sqrt{p_k} \mathbb{E}\left\{ \mathbf{g}_{kk} \right\}$, and $\mathbf{B} = \sum_{i=1}^{K} p_i \mathbb{E}\{ \mathbf{g}_{ki} \mathbf{g}_{ki}^{\text{H}} \} - p_k \mathbb{E}\{ \mathbf{g}_{kk} \} \mathbb{E}\{ \mathbf{g}_{kk}^{\text{H}} \} + \mathbf{F}_k + \widetilde{\mathbf{D}}_k$. The term $\widetilde{\mathbf{D}}_k$ is not present in (5.26) but has been added to make $\mathbf{B}$ positive definite and thereby invertible.[3] This addition will not affect the result since $\widetilde{\mathbf{D}}_k \mathbb{E}\left\{ \mathbf{g}_{kk} \right\} = \mathbf{0}_L$. □

This corollary provides the optimal (opt) LSFD vector that maximizes the effective SINR in (5.26). The optimal vector has a structure resembling that of MMSE combining since it is also computed to maximize a generalized Rayleigh quotient. The evaluation of $\mathbf{a}_k^{\text{opt}}$ in (5.30) requires knowledge of the L-dimensional statistical vector $\mathbb{E}\left\{ \mathbf{g}_{kk} \right\}$, the $L \times L$ Hermitian complex matrix $\mathbb{E}\left\{ \mathbf{g}_{ki} \mathbf{g}_{ki}^{\text{H}} \right\}$, for $i = 1, \ldots, K$, and the L diagonal elements of the real-valued matrix $\mathbf{F}_k$. These statistical vectors and matrices can be computed at the APs, based on the received uplink

[3] Alternatively, pseudo-inverses could have been used in (5.30) and (5.31), but the end result will be the same since $\widetilde{\mathbf{D}}_k \mathbb{E}\left\{ \mathbf{g}_{kk} \right\} = \mathbf{0}_L$.

signals and their choices of local combining vectors. The CPU is unable to compute them locally since it does not have access to those signals, thus the APs need to send these statistical parameters to the CPU through the fronthaul links. We will quantify the fronthaul signaling load in Section 5.2.3 but it is clear that the number of parameters grows with K, thus making the use of the opt LSFD vector unscalable according to Definition 2.2 on p. 217.

We notice that the elements of $\mathbf{g}_{ki}$, for all i, and $\mathbf{F}_k$ whose indices correspond to the zero rows of $\mathbf{D}_{kl}$ are all zero. By following the approach outlined in Remark 5.1, the computation of $\mathbf{a}_k^{\text{opt}}$ can be carried out while ignoring the $L - |\mathcal{M}_k|$ APs that are not serving UE k (see also Figure 5.3). The corresponding elements of $\mathbf{a}_k$ can be set to zero and the computation of $\mathbf{a}_k^{\text{opt}}$ only requires the inversion of a $|\mathcal{M}_k| \times |\mathcal{M}_k|$ matrix and multiplication of it by an $|\mathcal{M}_k|$-length vector, which corresponds to $|\mathcal{M}_k|^2 + \frac{|\mathcal{M}_k|^3 - |\mathcal{M}_k|}{3}$ complex multiplications. This is reported in Table 5.3.

Remark 5.2 (Tightness of the capacity bound)**.** The SE expression presented in Theorem 5.4 is derived based on the UatF bound from the Cellular Massive MIMO literature [**massivemimobook**], [**Marzetta2016a**] in which the channel estimates are used for receive combining but "forgotten" when detecting the data. While this bounding principle is artificial when applied in cellular networks, where the combining and signal detection are carried out at the same place, it makes good sense in cell-free networks where each AP carries out local combining based on its local CSI and the CPU carries out the data detection without CSI.

Signal detection without access to channel knowledge generally leads to rather poor performance since the receiver does not know the desired channel realization $\mathbf{a}_k^{\text{H}}\mathbf{g}_{kk}$ but only its mean value $\mathbf{a}_k^{\text{H}}\mathbb{E}\{\mathbf{g}_{kk}\}$. However, if there is a high degree of channel hardening so that $\mathbf{a}_k^{\text{H}}\mathbf{g}_{kk} \approx \mathbf{a}_k^{\text{H}}\mathbb{E}\{\mathbf{g}_{kk}\}$, the capacity bound in Theorem 5.4 is expected to be close to the SE that we would get if $\mathbf{a}_k^{\text{H}}\mathbf{g}_{kk}$ is perfectly known. As shown in Section 2.6.1 on p. 229, there is no guarantee that this will be the case in cell-free networks; the degree of channel hardening can be low when having a small number of antennas per AP [**Chen2018b**]. In that case, the SE

expression in (5.25) underestimates the practically achievable SE, but it is anyway the state-of-the-art capacity lower bound and will be used in this monograph. We will evaluate the tightness of the bound later in this section.

5.2.2 Scalable Schemes for Distributed Operation

While the L-MMSE combining in (5.29) and the opt LSFD vector in (5.30) are the preferable schemes in the distributed operation, none of these schemes is scalable. The former one has a computational complexity that increases with the number of UEs in the whole network and the latter one requires a fronthaul signaling load that grows in the same way. Therefore, we will now present suboptimal but scalable alternatives that are inspired by the aforementioned methods.

Scalable Local Receive Combining

We begin by revisiting the problem of selecting the combining vectors $\mathbf{v}_{kl}$ for $k \in \mathcal{D}_l$ at AP l. Recall that L-MMSE in (5.29) is not a scalable scheme since it makes use of estimates of the channels from all K UEs. To achieve scalability, the selection can instead be made as a function of the estimates $\{\widehat{\mathbf{h}}_{il} : i \in \mathcal{D}_l\}$ of the channels from the UEs that AP l is serving.

The simplest solution is to utilize MR combining [**Nayebi2016a**], [**Ngo2017b**], which is given by

$$\mathbf{v}_{kl}^{\mathrm{MR}} = \mathbf{D}_{kl}\widehat{\mathbf{h}}_{kl}. \tag{5.32}$$

One key benefit of using this scheme is that the combining vector follows directly from the channel estimates so that no additional computations are needed. The total number of complex multiplications is $\left(N\tau_p + N^2\right)|\mathcal{M}_k|$ and is reported in Table 5.3. Another benefit is that all the expectations in Theorem 5.4 can be computed analytically, so the SE is available in closed form.

Corollary 5.6. If MR combining with $\mathbf{v}_{kl}^{\mathrm{MR}} = \mathbf{D}_{kl}\widehat{\mathbf{h}}_{kl}$ is used, then the

expectations in (5.26) become

$$[\mathbb{E}\{\mathbf{g}_{ki}\}]_l = \begin{cases} \sqrt{\eta_k \eta_i}\tau_p \mathrm{tr}\left(\mathbf{D}_{kl}\mathbf{R}_{il}\boldsymbol{\Psi}_{t_k l}^{-1}\mathbf{R}_{kl}\right) & i \in \mathcal{P}_k \\ 0 & i \notin \mathcal{P}_k \end{cases} \tag{5.33}$$

with $[\mathbb{E}\{\mathbf{g}_{ki}\mathbf{g}_{ki}^{\mathrm{H}}\}]_{lr} = [\mathbb{E}\{\mathbf{g}_{ki}\}]_l [\mathbb{E}\{\mathbf{g}_{ki}^{\star}\}]_r$ for $r \neq l$ and

$$\begin{aligned}[\mathbb{E}\{\mathbf{g}_{ki}\mathbf{g}_{ki}^{\mathrm{H}}\}]_{ll} =& \eta_k \tau_p \mathrm{tr}\left(\mathbf{D}_{kl}\mathbf{R}_{il}\mathbf{R}_{kl}\boldsymbol{\Psi}_{t_k l}^{-1}\mathbf{R}_{kl}\right) \\ &+ \begin{cases} \eta_k \eta_i \tau_p^2 \left|\mathrm{tr}\left(\mathbf{D}_{kl}\mathbf{R}_{il}\boldsymbol{\Psi}_{t_k l}^{-1}\mathbf{R}_{kl}\right)\right|^2 & i \in \mathcal{P}_k \\ 0 & i \notin \mathcal{P}_k \end{cases} \end{aligned} \tag{5.34}$$

while

$$[\mathbf{F}_k]_{ll} = \sigma_{\mathrm{ul}}^2 \eta_k \tau_p \mathrm{tr}\left(\mathbf{D}_{kl}\mathbf{R}_{kl}\boldsymbol{\Psi}_{t_k l}^{-1}\mathbf{R}_{kl}\right). \tag{5.35}$$

Proof. The proof is available in Appendix C.2.4 on p. 452. □

The results from Corollary 5.6 can be inserted into the instantaneous effective SINR in (5.26) to obtain a closed-form SE expression. However, the final expression is lengthy so we will only present it in the case of $N = 1$, where it can be substantially simplified.

Corollary 5.7. If each AP has $N = 1$ antenna, then the MMSE estimate of the scalar channel $\mathbf{h}_{kl} \in \mathbb{C}$ has variance

$$\gamma_{kl} = \frac{\eta_k \tau_p \beta_{kl}^2}{\sum_{i \in \mathcal{P}_k} \eta_i \tau_p \beta_{il} + \sigma_{\mathrm{ul}}^2}. \tag{5.36}$$

The instantaneous effective SINR, $\mathsf{SINR}_k^{(\mathrm{ul,d})}$, in (5.26) then becomes

$$\frac{p_k \left|\sum\limits_{l \in \mathcal{M}_k} a_{kl}^{\star}\gamma_{kl}\right|^2}{\sum\limits_{i=1}^{K} p_i \sum\limits_{l \in \mathcal{M}_k} |a_{kl}|^2 \beta_{il}\gamma_{kl} + \sum\limits_{i \in \mathcal{P}_k \setminus \{k\}} p_i \left|\sum\limits_{l \in \mathcal{M}_k} a_{kl}^{\star}\gamma_{kl}\sqrt{\frac{\eta_i}{\eta_k}}\frac{\beta_{il}}{\beta_{kl}}\right|^2 + \sigma_{\mathrm{ul}}^2 \sum\limits_{l \in \mathcal{M}_k} |a_{kl}|^2 \gamma_{kl}}. \tag{5.37}$$

Proof. Using the notation in (5.36), the expectations in Corollary 5.6 can be rewritten as $[\mathbb{E}\{\mathbf{g}_{ki}\}]_l = \sqrt{\frac{\eta_i}{\eta_k}}\frac{\beta_{il}}{\beta_{kl}}\gamma_{kl}$, for $i \in \mathcal{P}_k$ and zero for

other i, $[\mathbf{F}_k]_{ll} = \sigma^2_{\text{ul}}\gamma_{kl}$,

$$[\mathbb{E}\{\mathbf{g}_{ki}\mathbf{g}_{ki}^{\text{H}}\}]_{ll} = \beta_{il}\gamma_{kl} + \begin{cases} \frac{\eta_i\beta_{il}^2}{\eta_k\beta_{kl}^2}\gamma_{kl}^2 & i \in \mathcal{P}_k \\ 0 & i \notin \mathcal{P}_k \end{cases} \tag{5.38}$$

for $l \in \mathcal{M}_k$ and zero for other l. Inserting these values into (5.26) yields the SINR expression in (5.37). □

The closed-form effective SINR in (5.37) shows the key behaviors of the cell-free operation. The signal term $p_k|\sum_{l\in\mathcal{M}_k} a_{kl}^\star\gamma_{kl}|^2$ in the numerator contains the transmit power p_k multiplied with a term containing the weighted sum of the contributions from the serving APs. Each AP contributes with a term proportional to the variance γ_{kl} of its channel estimate, since the combining vector is based on that estimate. The coherent combination of the contributions from the different APs is represented by the square of the sum. The first term in the denominator contains non-coherent interference, which means that it is a summation of the powers $p_i\beta_{il}$ of the individual interfering signals at the serving APs $l \in \mathcal{M}_k$ without any squares. Each term is multiplied with a scaling factor $|a_{kl}|^2\gamma_{kl}$ containing the LSFD weight and the scaling made by the MR combining. The second term in the denominator is the additional coherent interference caused by pilot contamination, which has a similar shape as the signal term. The third term is the noise power.

If local MR combining is used along with the LSFD vector $\mathbf{a}_k = [1 \ldots 1]^{\text{T}}$, then the result of the distributed operation is equivalent to MR combining in the centralized operation. In other words, centralized MR combining can be implemented in a distributed fashion. However, MR is expected to perform poorly in the presence of interference, unless a high degree of favorable propagation is achieved, which is not guaranteed in cell-free networks (see Section 2.6.2 on p. 234). Therefore, we will also consider combining schemes that can suppress interference while still being scalable.

Inspired by the scalable P-MMSE combining in (5.16) that was designed for centralized operation, we can define the local P-MMSE (LP-MMSE) at AP l based on only the channel estimates and statistics of UEs that are served by this AP (i.e., $\{\widehat{\mathbf{h}}_{il} : i \in \mathcal{D}_l\}$). By starting from

Scheme	Channel estimation	Combining vector computation
opt (and n-opt) LSFD	–	$\lvert\mathcal{M}_k\rvert^2 + \frac{\lvert\mathcal{M}_k\rvert^3 - \lvert\mathcal{M}_k\rvert}{3}$
L-MMSE	$(N\tau_p + N^2)K\lvert\mathcal{M}_k\rvert$	$\frac{N^2+N}{2}K\lvert\mathcal{M}_k\rvert + N^2\lvert\mathcal{M}_k\rvert + \frac{N^3-N}{3}\lvert\mathcal{M}_k\rvert$
LP-MMSE	$(N\tau_p + N^2)\sum_{l\in\mathcal{M}_k}\lvert\mathcal{D}_l\rvert$	$\frac{N^2+N}{2}\sum_{l\in\mathcal{M}_k}\lvert\mathcal{D}_l\rvert + N^2\lvert\mathcal{M}_k\rvert + \frac{N^3-N}{3}\lvert\mathcal{M}_k\rvert$
MR	$(N\tau_p + N^2)\lvert\mathcal{M}_k\rvert$	–

Table 5.3: The number of complex multiplications required per coherence block to compute the local combining vectors of UE k, including the computation of the channel estimates. Different combining schemes for the distributed operation are considered.

(5.29) but taking a summation over only the UEs in $\mathcal{D}_l$, instead of all UEs, we obtain the local combining vector

$$\mathbf{v}_{kl}^{\mathrm{LP-MMSE}} = p_k \left(\sum_{i \in \mathcal{D}_l} p_i \left(\widehat{\mathbf{h}}_{il} \widehat{\mathbf{h}}_{il}^{\mathrm{H}} + \mathbf{C}_{il} \right) + \sigma_{\mathrm{ul}}^2 \mathbf{I}_N \right)^{-1} \mathbf{D}_{kl} \widehat{\mathbf{h}}_{kl}. \quad (5.39)$$

The number of complex multiplications required by LP-MMSE is reported in Table 5.3. Since this number is independent of K, LP-MMSE is a scalable solution as $K \to \infty$. Compared to P-MMSE in (5.16) that was proposed for centralized operation, LP-MMSE has a substantially lower complexity per UE since it requires to compute the inverse of an $N \times N$ matrix, rather than an $N|\mathcal{M}_k| \times N|\mathcal{M}_k|$ matrix. Note that LP-MMSE coincides with L-MMSE when AP l serves all the UEs (i.e., $\mathcal{D}_l = \{1, \ldots, K\}$). Unlike MR, the expectations in (5.26) and (5.31) cannot be computed in closed form when using LP-MMSE, but can be computed numerically using Monte Carlo methods, as will be done later in this section.

Remark 5.3 (Local P-RZF)**.** Since the complexity of P-MMSE is rather high in the centralized operation, we presented an alternative simplified P-RZF scheme with lower complexity. A similar simplification is not needed in the distributed operation since LP-MMSE already has a rather low complexity.

Scalable Large-Scale Fading Decoding

The opt LSFD vector in Corollary 5.5 is also not scalable for implementation in large networks. The bottleneck originates from the fact that all UEs in the network affect the interference levels at all APs, thereby determining how accurate the local data estimates are. To identify a scalable way to select $\mathbf{a}_k$ for UE k, we need to limit how many interfering UEs are considered in the computation.

The simplest solution is to give all the serving APs equal importance by setting $\mathbf{a}_k = [1 \ \ldots \ 1]^{\mathrm{T}}$ [**Nayebi2017a**], [**Ngo2017b**]. In this case, the CPU takes the local estimates of the data from UE k and sum them

up to create the final data estimate

$$\widehat{s}_k = \sum_{l \in \mathcal{M}_k} \widehat{s}_{kl}. \tag{5.40}$$

We call this approach *no LSFD* since the LSFD weights are basically omitted. The reason is that the original works on Cell-free Massive MIMO considered this solution, while LSFD weights were introduced only later. The key benefit is that no statistical parameters are needed at the CPU to compute (5.40).[4] The main drawback is that equal importance is given to all APs when computing (5.40), even if some APs might be subject to high interference or low SNR. It can even happen that an AP reduces rather than increases the SE due to the suboptimal fusing of information at the CPU [**Buzzi2017b**].

The CPU should use some side-information of how accurate the respective local data estimates are, because this is what the opt LSFD vector does. Similar to the approach taken when P-MMSE combining was developed in Section 5.1.4, we can compute an approximately optimal solution by taking only the UEs that are served by partially the same APs as UE k into consideration. If the AP selection is done properly, these should be the UEs that cause the vast majority of the interference and, thus, affect the quality of the local data estimates the most. Recall that $\mathcal{S}_k$ in (5.15) denotes the set of UEs that have at least one serving AP in common with UE k, including itself. We then approximate the opt LSFD vector in (5.30) as

$$\mathbf{a}_k^{\mathrm{n-opt}} = p_k \left(\sum_{i \in \mathcal{S}_k} p_i \mathbb{E}\left\{ \mathbf{g}_{ki} \mathbf{g}_{ki}^{\mathrm{H}} \right\} + \mathbf{F}_k + \widetilde{\mathbf{D}}_k \right)^{-1} \mathbb{E}\left\{ \mathbf{g}_{kk} \right\} \tag{5.41}$$

which we call the *nearly optimal (n-opt) LSFD vector* [**globecom_nopt**]. The "nearly optimal" name will be demonstrated numerically later in this section. The number of complex multiplications required for computing $\mathbf{a}_k^{\mathrm{n-opt}}$ is the same as those needed for computing $\mathbf{a}_k^{\mathrm{opt}}$ since the matrices and vectors have the same dimensions. The key difference is

[4] To be more precise, no statistical parameters are needed to compute $\mathbf{a}_k$ or (5.40), but the data detection leading to the SE in Theorem 5.4 requires that the signal strength and interference variance are known at the CPU.

that the n-opt LSFD scheme requires knowledge of a lower number of statistical parameters, which is independent of K. This makes its fronthaul signaling load scalable, as we will elaborate on next.

5.2.3 Fronthaul Signaling Load for Distributed Operation

The fronthaul signaling required by the distributed operation can be quantified as follows. Each AP l needs to send the local estimates $\widehat{s}_{kl}$ for $k \in \mathcal{D}_l$ to the CPU, which corresponds to $\tau_u |\mathcal{D}_l|$ complex scalars per coherence block. This number equals $\tau_u \tau_p$ if we use the pilot assignment scheme from Section 4.4 on p. 286, where each AP serves τ_p UEs. This value does not grow with K so we can conclude that this part of the distributed operation is scalable. The total fronthaul signaling is summarized in Table 5.2.

Moreover, the statistical parameters for the computation of the LSFD vectors are required to be sent to the CPU and this number is different for opt LSFD and n-opt LSFD as we quantify in the following. Recall that $\mathbf{g}_{ki} = [\mathbf{v}_{k1}^{\mathrm{H}} \mathbf{D}_{k1} \mathbf{h}_{i1} \ \ldots \ \mathbf{v}_{kL}^{\mathrm{H}} \mathbf{D}_{kL} \mathbf{h}_{iL}]^{\mathrm{T}}$. Moreover, we notice that $\mathbb{E}\{[\mathbf{g}_{ki}]_l [\mathbf{g}_{ki}^{\star}]_r\} = \mathbb{E}\{[\mathbf{g}_{ki}]_l\} \mathbb{E}\{[\mathbf{g}_{ki}^{\star}]_r\}$ for $l \neq r$, by utilizing the independence of the channels corresponding to different APs. Hence, each AP can individually send its first- and second-order statistics to the CPU, which can then combine them to compute the required parameters. The following statistical parameters are needed to be sent from AP l to the CPU for the computation of the opt LSFD vector in (5.30) for a generic UE $k \in \mathcal{D}_l$:

- $\mathbb{E}\{[\mathbf{g}_{ki}]_l\}$, for $i = 1, \ldots, K$;
- $\mathbb{E}\left\{|[\mathbf{g}_{ki}]_l|^2\right\}$, for $i = 1, \ldots, K$;
- $[\mathbf{F}_k]_{ll}$.

This sums up to $(3K+1)/2$ complex scalars. Each AP sends these complex scalars for each of its $|\mathcal{D}_l|$ UEs and, hence, $(3K+1)/2 \sum_{l=1}^{L} |\mathcal{D}_l|$ complex scalars are needed in total. These values are summarized in Table 5.2. As anticipated earlier in this section, the number of statistical parameters that need to be shared when using opt LSFD increases with K, thus making it unscalable.

If the n-opt LSFD vector in (5.41) is used instead, then AP l is required to send $(3|\mathcal{S}_k|+1)/2$ complex parameters to the CPU. In that case, the required total number of complex scalars becomes $\sum_{l=1}^{L}\sum_{k\in\mathcal{D}_l}(3|\mathcal{S}_k|+1)/2$, as summarized in Table 5.2. Fortunately, this number does not grow with K (see the discussion after (5.17) for further details) and we can conclude that the combination of n-opt LSFD and any scalable local combining scheme leads to a completely scalable distributed operation in the uplink.

Finally, in the case when no LSFD vector is used (i.e., all its elements are equal), no channel statistics need to be sent to the CPU. This case is also listed in Table 5.2 as a worst-case baseline.

5.3 Running Example

To exemplify the performance of User-centric Cell-free Massive MIMO and make comparisons with cellular networks under somewhat realistic conditions, we will now define a network setup that will be used as a running example in the remainder of this monograph. The key parameters are given in Table 5.4. The total coverage area is $1\,\text{km} \times 1\,\text{km}$ and the total number of antennas is $M = LN = 400$. The number of APs is either $L = 400$ or $L = 100$. In the former case, the number of antennas per AP will be $N = 1$ whereas it is $N = 4$ for the latter case. A wrap-around topology is used to mimic a large network deployment without edges, where all APs and UEs are receiving interference from all directions. Unless otherwise stated, we assume that the APs are deployed uniformly at random in the coverage area. The communication takes place over a 20 MHz bandwidth with a total receiver noise power of $-94\,\text{dBm}$ (consisting of thermal noise and a noise figure of 7 dB in the receiver hardware) at both the APs and UEs. The maximum uplink transmit power of each UE is 100 mW while the maximum downlink transmit power of each AP is 200 mW. The difference models the fact that APs are connected to the electricity grid and thus less power limited. Each coherence block consists of $\tau_c = 200$ samples. This value matches well with a 2 ms coherence time and a 100 kHz coherence bandwidth, which correspond to high mobility and large channel dispersion in sub-6 GHz bands, as exemplified in Section 2.3.1 on p. 208. Hence, we are

Parameter	Value
Network area	$1\,\text{km} \times 1\,\text{km}$
Network layout	Random deployment (with wrap-around)
Number of APs	400 or 100
Number of antennas per AP	1 or 4
Number of total antennas	$M = LN = 400$
Bandwidth	$B = 20\,\text{MHz}$
Receiver noise power	$\sigma^2_{\text{ul}} = -94\,\text{dBm}$
Maximum uplink transmit power	$100\,\text{mW}$
Maximum downlink transmit power	$200\,\text{mW}$
Samples per coherence block	$\tau_c = 200$
Channel gain at 1 km	$\Upsilon = -140.6\,\text{dB}$
Pathloss exponent	$\alpha = 3.67$
Height difference between AP and UE	$10\,\text{m}$
Standard deviation of shadow fading	$\sigma_{\text{sf}} = 4$

Table 5.4: Key parameters of running example.

considering an example that supports both user mobility and outdoor propagation conditions.

The large-scale fading coefficient (channel gain) is computed in dB on the basis of the model presented in Section 2.5.2 on p. 223, and reported here for convenience:

$$\beta_{kl}\,[\text{dB}] = -30.5 - 36.7 \log_{10}\left(\frac{d_{kl}}{1\,\text{m}}\right) + F_{kl} \tag{5.42}$$

where d_{kl} is the three-dimensional distance between AP l and UE k. The APs are deployed 10 m above the plane where the UEs are located, which acts as the minimum distance. This model matches with the 3GPP Urban Microcell model in [**LTE2017a**]. The shadow fading is $F_{kl} \sim \mathcal{N}(0, 4^2)$ and the terms from an AP to different UEs are correlated as [**LTE2017a**]

$$\mathbb{E}\{F_{kl}F_{ij}\} = \begin{cases} 4^2 2^{-\delta_{ki}/9\,\text{m}} & l = j \\ 0 & l \neq j \end{cases} \tag{5.43}$$

where δ_{ki} is the distance between UE k and UE i. The intuition behind

the shadow fading correlation is that if two UEs are closely located and one of them is subject to strong shadowing caused by a large object in the propagation environment, then the other UE will likely also be subject to strong shadowing. The second row in (5.43) implies that each UE achieves independent shadow fading realizations from different APs, since these are deployed at fairly different locations. The large-scale fading model in (5.42) corresponds to a median channel gain of $-140.6\,\text{dB}$ at $1\,\text{km}$ (the median is achieved by $F_{kl} = 0\,\text{dB}$) and has a pathloss exponent $\alpha = 3.67$, as reported in Table 5.4.

The spatial correlation matrices are generated using the local scattering model defined in Section 2.5.3 on p. 224. The multipath components are Gaussian distributed around the nominal azimuth and elevation angles according to (2.19), where the nominal angles are computed by drawing a line between the AP and UE. We let σ_φ and σ_θ denote the ASDs in the azimuth and elevation domains, respectively. Different values are considered throughout the simulations. We will also consider uncorrelated fading with $\mathbf{R}_{kl} = \beta_{kl}\mathbf{I}_N$ as a reference case.

To assign the pilots and form the DCC, the joint pilot assignment and AP selection scheme from Section 4.4 on p. 286 is used.

5.3.1 Benchmark Schemes

Most of the performance evaluations will consider the achievable SE expressions that were provided in Theorem 5.1 for the centralized operation and in Theorem 5.4 for the distributed operation. However, we will also compare these results with some benchmark schemes. For completeness, these will be briefly presented below.

Cellular Network

To compare the performance of a cell-free network with that of a corresponding cellular network, we can treat the cellular network as a degenerate case of a cell-free network where each UE is only served by one of the APs; that is, $|\mathcal{M}_k| = 1$. To compare with the best conceivable small-cell implementation, we assume the APs use the optimal receive combining. In this case, the SE can be computed as follows.

Corollary 5.8. If UE k is served only by AP l, an achievable SE is

$$\mathsf{SE}_{kl}^{(\mathrm{ul,cellular})} = \frac{\tau_u}{\tau_c}\mathbb{E}\left\{\log_2\left(1+\mathsf{SINR}_{kl}^{(\mathrm{ul,cellular})}\right)\right\} \tag{5.44}$$

where the instantaneous effective SINR is

$$\mathsf{SINR}_{kl}^{(\mathrm{ul,cellular})} = \frac{p_k\left|\mathbf{v}_{kl}^{\mathrm{H}}\widehat{\mathbf{h}}_{kl}\right|^2}{\sum\limits_{\substack{i=1\\i\neq k}}^{K} p_i\left|\mathbf{v}_{kl}^{\mathrm{H}}\widehat{\mathbf{h}}_{il}\right|^2 + \mathbf{v}_{kl}^{\mathrm{H}}\left(\sum\limits_{i=1}^{K} p_i\mathbf{C}_{il}+\sigma_{\mathrm{ul}}^2\mathbf{I}_N\right)\mathbf{v}_{kl}} \tag{5.45}$$

for an arbitrary receive combining $\mathbf{v}_{kl}\in\mathbb{C}^N$. The maximum value in (5.45) is achieved with the L-MMSE combining in (5.29), leading to

$$\mathsf{SINR}_{kl}^{(\mathrm{ul,cellular})} = p_k\widehat{\mathbf{h}}_{kl}^{\mathrm{H}}\left(\sum_{\substack{i=1\\i\neq k}}^{K} p_i\widehat{\mathbf{h}}_{il}\widehat{\mathbf{h}}_{il}^{\mathrm{H}} + \sum_{i=1}^{K} p_i\mathbf{C}_{il}+\sigma_{\mathrm{ul}}^2\mathbf{I}_N\right)^{-1}\widehat{\mathbf{h}}_{kl}. \tag{5.46}$$

Proof. This result follows from Theorem 5.1 and Corollary 5.3 in the special case of $\mathcal{M}_k=\{l\}$. □

The SE-maximizing receive combining is L-MMSE, which is not scalable but will anyway be used as a benchmark to compare cell-free networks with the best kind of cellular network implementation.[5]

The results presented above can be directly used to consider a cellular network with the same AP locations as in the cell-free counterpart. However, it can also be used to consider a Cellular Massive MIMO system, which is characterized by a substantially smaller L and a substantially larger N than in the cell-free network.

In the simulations, the APs in the small-cell setup are deployed in the same locations as in the cell-free network. The system-specific parameter values for the small-cell systems are reported in Table 5.5. Hence, the number of APs will be either $L=400$ or $L=100$. In the former case, the number of antennas per AP will be $N=1$ whereas it will be $N=4$ for the latter case. The same set of UEs is served by both

[5]The L-MMSE scheme is more commonly referred to as multi-cell MMSE combining in the literature on Cellular Massive MIMO [**BjornsonHS17**], [**massivemimobook**].

Parameter	Value
Number of small cells	400 or 100
Coverage area per cell	50 m $\times$ 50 m or 100 m $\times$ 100 m
Number of antennas per AP	1 or 4

Table 5.5: The parameters of the running example for the small-cell system. The APs are distributed either on a square grid or uniformly at random. The specified coverage areas are exact in the former case and computed on the average in the latter case.

network setups, in order to achieve a fair comparison. The small-cell system uses the same pilot assignment from Section 4.4 on p. 286 as in the cell-free case even if only one AP serves each UE. The index of the single AP that is serving UE k is determined as

$$\ell = \underset{l \in \mathcal{M}_k}{\arg\max} \ \beta_{kl} \tag{5.47}$$

where $\mathcal{M}_k$ is the DCC for UE k that is found by implementing Algorithm 4.1 on p. 289 for the cell-free case.

In the Cellular Massive MIMO setup, we assume each UE is connected to the AP with the best average channel gain, i.e., β_{kl} for UE k and AP l, and we will specify the system parameters and how the pilot assignment is carried out later in the simulation part.

Genie-Aided SEs With Centralized and Distributed Operations

The SE expressions presented in Theorem 5.1 for centralized operation and in Theorem 5.4 for distributed operation are computed using the state-of-the-art capacity lower bounds. The bound for the centralized operation is reliable but Remark 5.2 emphasized that the bound for the distributed operation might underestimate the actually achievable SE. This is because the bound is only tight when there is a sufficient degree of channel hardening. Unfortunately, there is no good definition of what this means. To evaluate these SE expressions, we can compare them to upper bounds. It is easy to obtain practically unachievable upper bounds by neglecting interference or changing the signal processing entirely, but this will not provide much insight into the tightness of

the SE expressions presented earlier in this section. We will instead use the following methodology: we select the receive combining using the estimated channels as before, but in the final detection step we inject perfect "genie-aided" channel knowledge. The SE is still computed by treating interference as noise, but the perfect CSI assumption leads to higher SE values.

We begin by considering the centralized operation and assume the channels $\{\mathbf{h}_k : k = 1, \ldots, K\}$ are available at the CPU after the receive combining has been carried out. We can then rewrite the soft estimate of UE k's uplink data in (5.3) as

$$\widehat{s}_k = \underbrace{\mathbf{v}_k^{\mathrm{H}} \mathbf{D}_k \mathbf{h}_k s_k}_{\text{Desired signal}} + \underbrace{\sum_{\substack{i=1 \\ i \neq k}}^{K} \mathbf{v}_k^{\mathrm{H}} \mathbf{D}_k \mathbf{h}_i s_i}_{\text{Interference}} + \underbrace{\mathbf{v}_k^{\mathrm{H}} \mathbf{D}_k \mathbf{n}}_{\text{Noise}} \tag{5.48}$$

where the desired signal term now includes the channel $\mathbf{h}_k$ itself rather than its estimate. We then have the following result.

Corollary 5.9. A genie-aided SE of UE k in the centralized operation is

$$\mathsf{SE}_k^{(\text{gen}-\text{ul},\mathsf{c})} = \frac{\tau_u}{\tau_c} \mathbb{E}\left\{ \log_2 \left(1 + \mathsf{SINR}_k^{(\text{gen}-\text{ul},\mathsf{c})} \right) \right\} \tag{5.49}$$

where

$$\mathsf{SINR}_k^{(\text{gen}-\text{ul},\mathsf{c})} = \frac{p_k \left| \mathbf{v}_k^{\mathrm{H}} \mathbf{D}_k \mathbf{h}_k \right|^2}{\sum\limits_{\substack{i=1 \\ i \neq k}}^{K} p_i \left| \mathbf{v}_k^{\mathrm{H}} \mathbf{D}_k \mathbf{h}_i \right|^2 + \sigma_{\text{ul}}^2 \left\| \mathbf{D}_k \mathbf{v}_k \right\|^2}. \tag{5.50}$$

Proof. This follows from utilizing Lemma 3.5 on p. 248. □

Similarly, in the distributed operation, we can rewrite the data estimate of UE k at the CPU in (5.24) for final data detection as

$$\widehat{s}_k = \underbrace{\mathbf{a}_k^{\mathrm{H}} \mathbf{g}_{kk} s_k}_{\text{Desired signal}} + \underbrace{\sum_{\substack{i=1 \\ i \neq k}}^{K} \mathbf{a}_k^{\mathrm{H}} \mathbf{g}_{ki} s_i}_{\text{Interference}} + \underbrace{n'_k}_{\text{Noise}} \tag{5.51}$$

where $\mathbf{a}_k \in \mathbb{C}^L$ is the LSFD vector, $n'_k = \sum_{l=1}^{L} a^{\star}_{kl} \mathbf{v}_{kl}^{\mathrm{H}} \mathbf{D}_{kl} \mathbf{n}_l$, and $\mathbf{g}_{ki} = [\mathbf{v}_{k1}^{\mathrm{H}} \mathbf{D}_{k1} \mathbf{h}_{i1} \ \ldots \ \mathbf{v}_{kL}^{\mathrm{H}} \mathbf{D}_{kL} \mathbf{h}_{iL}]^{\mathrm{T}} \in \mathbb{C}^L$. If perfect CSI is available in the final data detection step, we can treat the first term in (5.51) as the desired signal obtained over a known channel and thus get the following result.

Corollary 5.10. A genie-aided SE of UE k in the distributed operation is

$$\mathsf{SE}_k^{(\mathrm{gen-ul,d})} = \frac{\tau_u}{\tau_c} \mathbb{E} \left\{ \log_2 \left(1 + \mathsf{SINR}_k^{(\mathrm{gen-ul,d})} \right) \right\} \tag{5.52}$$

where

$$\mathsf{SINR}_k^{(\mathrm{gen-ul,d})} = \frac{p_k \left| \mathbf{a}_k^{\mathrm{H}} \mathbf{g}_{kk} \right|^2}{\sum\limits_{\substack{i=1 \\ i \neq k}}^{K} p_i |\mathbf{a}_k^{\mathrm{H}} \mathbf{g}_{ki}|^2 + \mathbf{a}_k^{\mathrm{H}} \mathbf{F}'_k \mathbf{a}_k} \tag{5.53}$$

with

$$\mathbf{F}'_k = \sigma^2_{\mathrm{ul}} \mathrm{diag} \left(\| \mathbf{D}_{k1} \mathbf{v}_{k1} \|^2, \ldots, \| \mathbf{D}_{kL} \mathbf{v}_{kL} \|^2 \right) \in \mathbb{C}^{L \times L}.$$

Proof. This follows from utilizing Lemma 3.5 on p. 248. □

We reiterate that, in the simulations, the combining vectors and LSFD vectors are computed using the estimated channels, as described earlier in this section. It is only in the final data detection step that we inject the receiver with perfect CSI. In this way, we can evaluate the tightness of the capacity bounds in Theorem 5.1 for centralized operation and in Theorem 5.4 for distributed operation, for a given receiver operation.

5.4 Numerical Performance Evaluation

In this section, we will quantify the uplink SE achieved by the centralized and distributed operations of User-centric Cell-free Massive MIMO, using the different combining schemes presented earlier in this section. The running example from Section 5.3 will be considered in all the simulations. Comparisons will be made with cellular networks with either small cells or Massive MIMO. In addition to showing the

SE, we will exemplify the fronthaul signaling load and the computational complexity of the combining schemes in the centralized and distributed operations, to demonstrate the difference between scalable and unscalable implementations.

In all the simulations in this section, we assume each UE transmits with full power both in the pilot and data transmission phases:

$$\eta_k = p_k = p_{\max}, \quad k = 1, \ldots, K. \tag{5.54}$$

We will consider other optimized and heuristic power control methods in Section 7 on p. 393, where it is also shown that full power transmission approximately maximizes the sum SE in the considered setup. All the APs are randomly distributed in the coverage area following an independent and uniform distribution. There are $K = 40$ UEs in all the considered setups, which implies that $L \gg K$. The pilot sequence length is $\tau_p = 10$ and the remaining $\tau_u = \tau_c - \tau_p = 190$ samples of each coherence block are used for uplink data. The Gaussian local scattering model is used to generate the spatial correlation matrices with ASD $\sigma_\varphi = \sigma_\theta = 15^\circ$.

We use the following Monte Carlo simulation methodology to generate the numerical results for the cell-free and small-cell networks:

1. Deploy the APs in the simulation area either in a random manner with independent uniform distribution or using a square grid.
2. Randomly drop the UEs one by one.
3. Compute the distance from the considered UE to each AP by using the wrap-around topology.[6]
4. Compute the channel gain from the considered UE to each AP by using (5.42). The shadow fading realizations are generated using the conditional Gaussian distribution recursively from [**Kay1993a**] to simulate the shadowing according to (5.43).

[6]This means that the north edge of the simulation area is connected to the south edge, while the west edge is connected to the east edge. Hence, there are multiple ways to draw a line between an AP and a UE, but we always consider the shortest option, which might go over an edge. The reason for having a wrap-around topology is to simulate a large-scale scenario where all the UEs effectively are in the center of the simulation area and are subject to interference from all directions.

5. Generate spatial correlation matrices $\mathbf{R}_{kl}$ and estimation error correlation matrices $\mathbf{C}_{kl}$.

6. Determine the pilot assignment and DCC by implementing Algorithm 4.1 on p. 289.

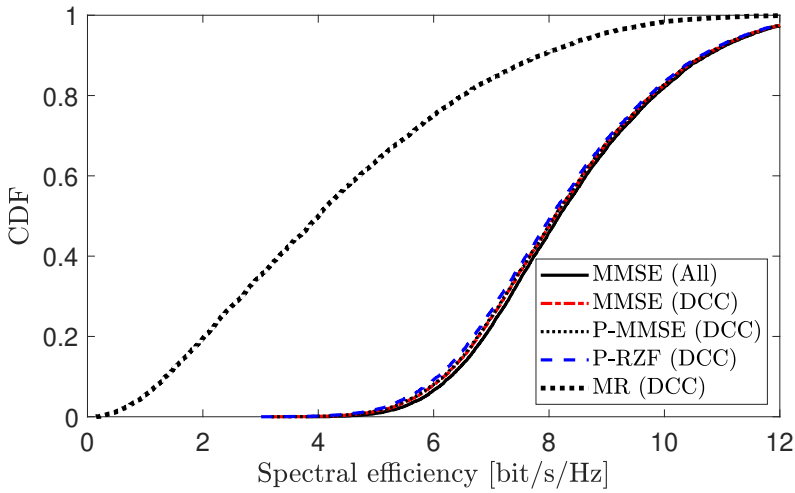
7. Generate the estimated channels $\widehat{\mathbf{h}}_{kl}$ and use them to compute sample averages that approximate all the expectations in the SE expressions.

In each figure, the legend will indicate the schemes that are being used. In some figures, we will compare the scalable operation obtained using Algorithm 4.1 on p. 289 with a Cell-free Massive MIMO network, where each UE is served by all the APs. In these cases, we use "(DCC)" to denote the scalable operation and "(All)" to denote the case where all APs serve all UEs.

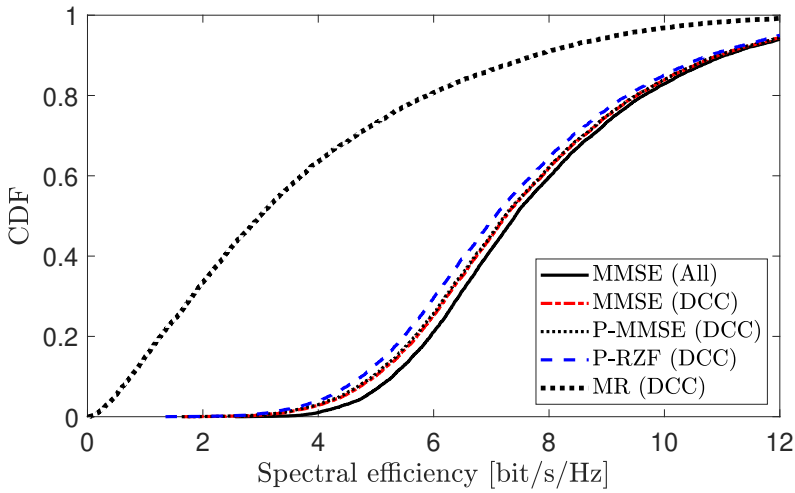
5.4.1 Performance With Centralized Operation

In Figure 5.4, we show the CDF of the SE per UE with a centralized operation of User-centric Cell-free Massive MIMO. The sources of randomness that give rise to the CDF are the random AP and UE locations, as well as the shadow fading realizations. The SE is computed using (5.4) for the different combining schemes described in Section 5.1: MMSE, P-MMSE, P-RZF, and MR.

Figure 5.4(a) shows the scenario with $L = 400$ APs and $N = 1$ antenna per AP and Figure 5.4(b) shows the scenario with $L = 100$ APs with $N = 4$ antennas per AP. In both scenarios, the largest SE is achieved by "MMSE (All)", which corresponds to that all APs serve all UEs using MMSE combining. This is expected since this is the optimal scheme, however, it is not a scalable solution. We notice that the use of "MMSE (DCC)" leads to a negligible performance loss, thus it is sufficient to only involve a subset of the APs in the signal detection. If we switch to using the scalable P-MMSE combining scheme, denoted as "P-MMSE (DCC)", we arrive at a fully scalable solution and the gap to the optimal scheme remains negligible. The P-RZF combining scheme, which is also scalable, results in a slightly lower SE than the P-MMSE



(a) $L = 400$, $N = 1$.



(b) $L = 100$, $N = 4$.

Figure 5.4: CDF of the uplink SE per UE in the centralized operation. We consider $K = 40$ UEs, $\tau_p = 10$, and spatially correlated Rayleigh fading with ASD $\sigma_\varphi = \sigma_\theta = 15°$. Different scalable and unscalable combining schemes are compared.

scheme. Hence, by capitalizing on the large channel gain variations in cell-free networks, a properly designed scalable DCC implementation can attain almost the same performance as the optimal unscalable solution. We note that MR combining achieves significantly smaller SEs than the interference-suppressing schemes. Recall that MR requires a high degree of favorable propagation to achieve good results, which apparently is not available anywhere in the network.

By comparing the performance of the two scenarios in Figures 5.4(a) and 5.4(b), we notice that the former case with many single-antenna APs is more desirable when having a centralized operation. The difference is particularly large at the lower end of the CDF curves, which explains that it is beneficial to spread out the antennas as much as possible to obtain the so-called macro-diversity gain. The reduction in the lowest SE values when switching from $L = 400$, $N = 1$ to $L = 100$, $N = 4$ are more apparent for the interference-suppressing schemes than for MR combining. Interestingly, the performance loss of using P-RZF combining is higher with $N = 4$ than in the case of $N = 1$. The reason is that in the first scenario with single-antenna APs, the collective estimation error correlation matrices $\mathbf{C}_k$ are diagonal whereas in the second case they have a block-diagonal structure with some non-zero off-diagonal elements due to the spatial correlation. When computing the P-RZF combining vector in (5.18), we neglect the estimation error correlation matrices $\mathbf{C}_k$ and when these are non-diagonal, this creates a more significant performance drop in comparison to the combining schemes that utilize the spatial structure of $\mathbf{C}_k$. On the other hand, only neglecting the matrices $\mathbf{C}_k$ for UEs that cause little interference, as done with P-MMSE combining, is associated with a negligible performance loss.

Tightness of the SE Expressions

The SE expression for centralized operation in (5.4) is a lower bound on the capacity, where the signals received over the unknown parts of the channel vectors are treated as noise. To quantify the tightness of this lower bound, Figure 5.5 considers the same simulation setup as in Figure 5.4(b) but includes the genie-aided SE from (5.49). We focus on P-MMSE combining, which is the scalable scheme that achieves the highest SE. Since the genie-aided SE is obtained with perfect CSI at the CPU,

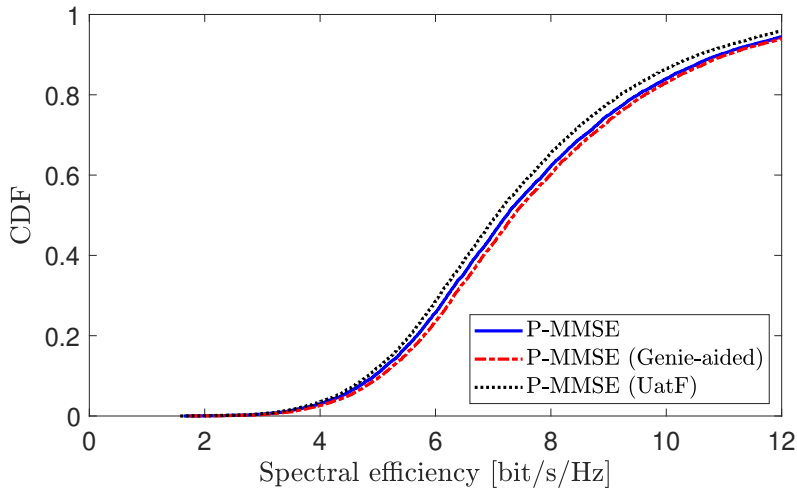


Figure 5.5: CDF of the uplink SE per UE in the centralized operation in the same scenario as in Figure 5.4(b). We consider $L = 100$, $N = 4$, $K = 40$, $\tau_p = 10$, and spatially correlated Rayleigh fading with ASD $\sigma_\varphi = \sigma_\theta = 15°$. The SE expression from (5.4) is compared with genie-aided SE in (5.49) and the UatF bound in (5.8).

a higher SE is achieved compared to (5.4). However, the performance gap is very small. In the figure, we also consider the alternative SE from (5.8) that is obtained by the UatF bounding technique, where no CSI is available in the final signal detection. This simplification results in lower SE values compared to the original SE expression in (5.4) but the gap is almost negligible. We will anyway continue using the original expression in the remainder of this section, but the UatF bound will be used for optimized power control in Section 7.1 on p. 394. In conclusion, the two SE expressions provided in Theorem 5.1 and in Theorem 5.2 are both representative of the practically achievable performance with centralized operation.

5.4.2 Performance With Distributed Operation

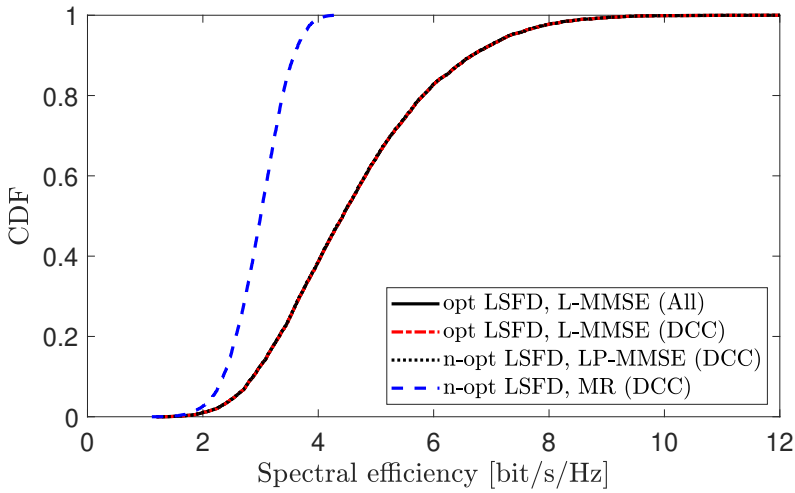
We will now focus on the distributed operation. Figure 5.6 shows the CDF of SE per UE in the same scenarios as in Figure 5.4. Hence, Figure 5.6(a) considers $L = 400$ and $N = 1$, while Figure 5.6(b) considers $L = 100$ and $N = 4$. The most advanced distributed implementation

is to use L-MMSE combining at each AP, which is locally optimal, and then to apply the optimal LSFD weights "opt LSFD" at the CPU. This combination gives the highest SEs in Figure 5.6, but it is not a scalable solution since the complexity grows unboundedly with K. The figure shows that we can achieve essentially the same SEs even if we make three simplifications: 1) involve only a subset of the APs in the data detection; 2) use the scalable n-opt LSFD (instead of opt LSFD); and 3) switch from L-MMSE to LP-MMSE combining. Hence, the performance loss required to achieve a scalable implementation is negligible also in the distributed operation. The figure also considers MR combining and we notice that there is a large SE loss compared to LP-MMSE combining. This is expected in the case of $N = 4$ since LP-MMSE is using the antennas to suppress interference. However, there is also a substantial performance loss in the case of $N = 1$. The reason is that MR creates larger variations in the interference power [**Bjornson2020a**].

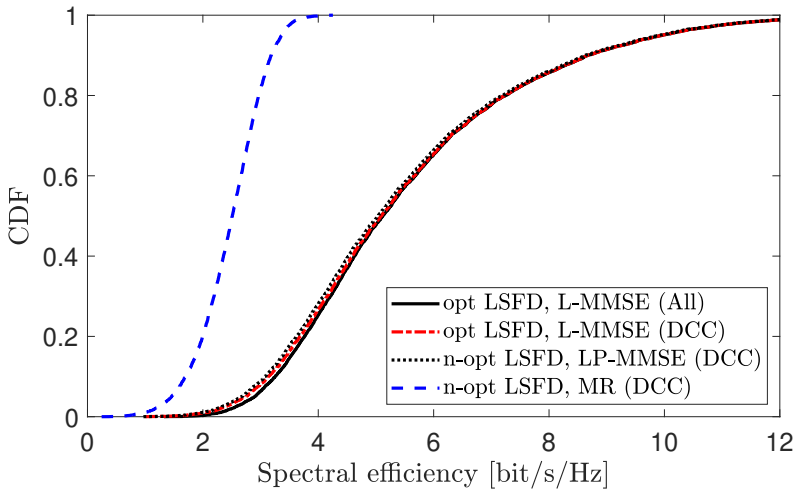
By comparing the two scenarios in Figure 5.6, we notice that the SE curves with $N = 4$ are more spread out than the curves with $N = 1$. The UEs with the worst channel conditions experience lower SE since there are larger gaps between the AP locations, while the UEs with the best channel conditions experience higher SE since these UEs are sensitive to interference and each AP can locally suppress interference using its array of antennas. Note that this behavior is different from the centralized case in Figure 5.4, where the gap between the two scenarios was larger and it was always preferable to have many APs since interference could be suppressed at the CPU using the received signals at multiple APs. One advantage of deploying $N = 4$ antennas per AP rather than $N = 1$ antenna (when $M = LK = 400$ is fixed) is that the channel estimation is improved at each AP owing to the spatial correlation between the channel coefficients observed at the antennas in the same array. This was previously discussed in Section 4.3.3 on p. 281.

Tightness of the SE Expressions

As discussed in Remark 5.2, the SE expressions used in the distributed operation are developed under the assumption that the CPU does not have access to the instantaneous channel realizations but only



(a) $L = 400$, $N = 1$.



(b) $L = 100$, $N = 4$.

Figure 5.6: CDF of the uplink SE per UE in the distributed operation. We consider $K = 40$ UEs, $\tau_p = 10$, and spatially correlated Rayleigh fading with ASD $\sigma_\varphi = \sigma_\theta = 15°$. Different LSFD and local combining schemes are compared.

to the channel statistics. Such capacity bounds might underestimate the achievable SE, unless there is a sufficiently high degree of channel hardening and small interference variations. To investigate this potential issue, Figure 5.7 shows the CDFs of the SE with the scalable schemes from Figure 5.6(b) with $L = 100$ and $N = 4$. We compare these curves with the genie-aided SE from (5.52), which assumes that the same processing schemes are used but the CPU has perfect CSI when detecting the data. The gap between the genie-aided SE and the achievable SE values is larger than in the centralized case (see Figure 5.5), but the difference is reasonably small when using LP-MMSE combining. However, there is a substantial gap when using MR combining. The reason for this is that MR gives rise to larger power variations in both the desired and interfering signals [**Bjornson2020a**], [**Interdonato2016a**], which the detector can deal with if they are known (as in the genie-aided case) but not when it lacks CSI. We conclude that the SE expression that we presented is practically achievable and fairly close to what could be achieved with perfect CSI at the CPU when using LP-MMSE combining (or other interference-suppressing schemes). The gap with MR can be reduced by normalizing it as explained in [**massivemimobook**] and [**Interdonato2016a**].

Benefits of LSFD

In the distributed cell-free operation, it is important to give different priorities to the local data estimates obtained by different APs using LSFD at the CPU. To demonstrate this, we will now compare the performance with and without LSFD, where the latter is represented by $\mathbf{a}_k = [1 \ \ldots \ 1]^{\mathrm{T}}$. Figure 5.8 compares the CDFs of the SE per UE with MR and L-MMSE combining when all APs serve all UE with their scalable alternatives, i.e., user-centric selection of APs with MR and LP-MMSE combining. The four left-most curves are without LSFD, where all the serving APs are given equal priority, while the results for the scalable n-opt LSFD with LP-MMSE combining from Figure 5.6(b) is included as a reference curve. The first observation is that all UEs experience a much higher SE with LSFD than without LSFD, which demonstrates that the LSFD feature is essential. The SE is very low with MR, which

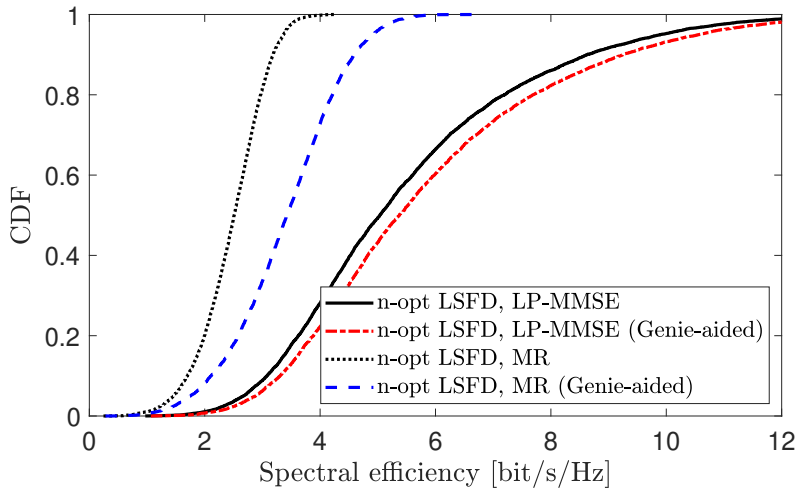


Figure 5.7: CDF of the uplink SE per UE in the distributed operation with LSFD in the same scenario as in Figure 5.6(b). We consider $L = 100$, $N = 4$, $K = 40$, $\tau_p = 10$, and spatially correlated Rayleigh fading with ASD $\sigma_\varphi = \sigma_\theta = 15°$. The SE expression from (5.25) is compared with genie-aided SE in (5.52).

can be partially addressed by adjusting the power control [**Ngo2017b**], but it is more efficient to use LP-MMSE combining in combination with LSFD [**Bjornson2019c**], [**Bjornson2020a**]. Note that the user-centric AP selection leads to minor performance degradation when using LP-MMSE compared to L-MMSE (All), while the loss is negligible when using MR since the network is strongly interference-limited in that case.

5.4.3 Comparison Between Cell-Free and Cellular Networks

In this section, we have thus far only evaluated the performance of cell-free networks with different types of operation and different combining schemes. We will now compare cell-free networks with cellular networks, thereby extending the preliminary comparison that was made in Section 1.3 on p. 188 under idealized conditions. It is inherently hard to make a fair comparison between different types of networks, but we will make an attempt by letting the total number of antennas be the same: $LN = 400$. We consider both a Cellular Massive MIMO with many antennas per AP and a small-cell system with the same AP configuration as in the cell-free network. The propagation model

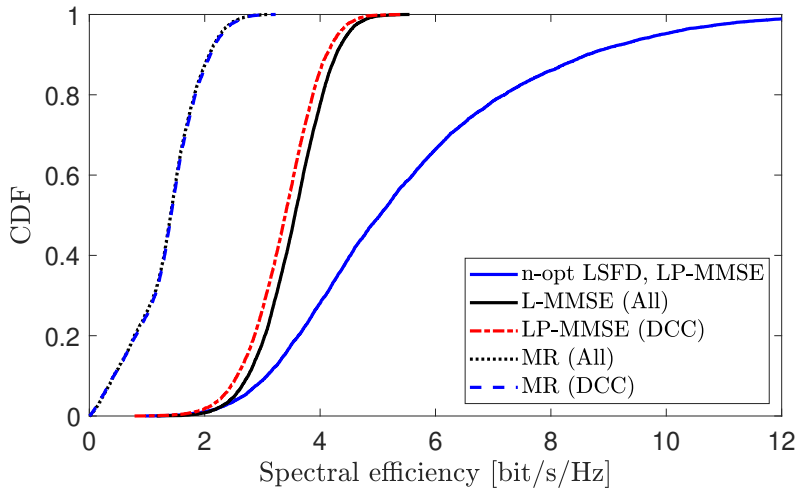


Figure 5.8: CDF of the uplink SE per UE in the distributed operation with and without LSFD. We consider the same scenario as Figure 5.6(b) with $L = 100$, $N = 4$, $K = 40$, $\tau_p = 10$, and spatially correlated Rayleigh fading with ASD $\sigma_\varphi = \sigma_\theta = 15°$. Different local combining schemes are compared. The scalable LSFD scheme with LP-MMSE combining from Figure 5.6(b) is provided as a benchmark.

described in Section 5.3 is used in all cases. In Table 5.6, we report the parameters of the Cellular Massive MIMO system where there are 4 square cells, each having an area of 500 m $\times$ 500 m. The APs are deployed at the center of the cells and they are each equipped with 100 antennas. For the cell-free network and small-cell system, either $L = 400$ APs with $N = 1$ antennas or $L = 100$ APs with $N = 4$ antennas are deployed on a symmetric square grid.

We drop the UEs so that each AP in the Cellular Massive MIMO system serves the same number of UEs. Note that this does not mean that these 10 UEs are geometrically located in the corresponding square cell. Due to the random shadow fading, a UE can occasionally get a better channel to another AP than the one located in its own square. To make a fair comparison for all the considered systems, we use the same UE locations and pilot assignments. In the Cellular Massive MIMO case, each UE in a cell is assigned to a unique pilot sequence from $\tau_p = 10$ mutually orthogonal pilot sequences. To guarantee this condition and

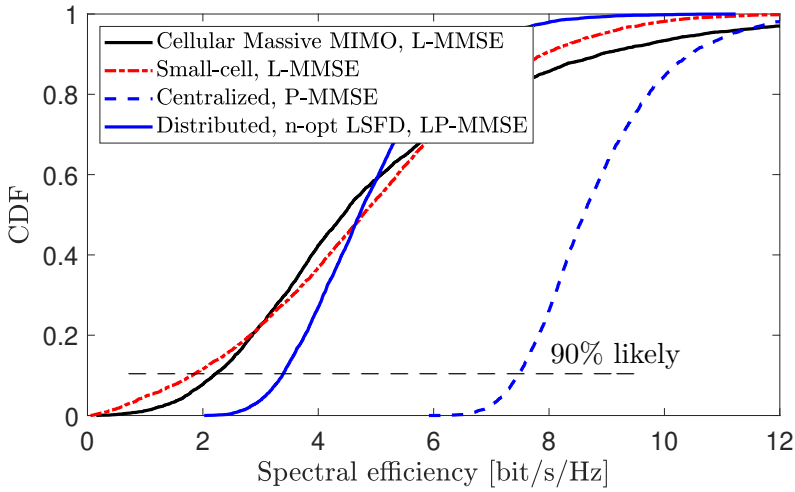
Parameter	Value
Number of cells	4
Cell area	500 m $\times$ 500 m
Number of antennas per AP	100
Number of UEs per cell	10

Table 5.6: The parameters for the Cellular Massive MIMO system used for performance comparison. Each cell covers a square area of 500 m $\times$ 500 m and is deployed on a grid of 2 $\times$ 2 cells.

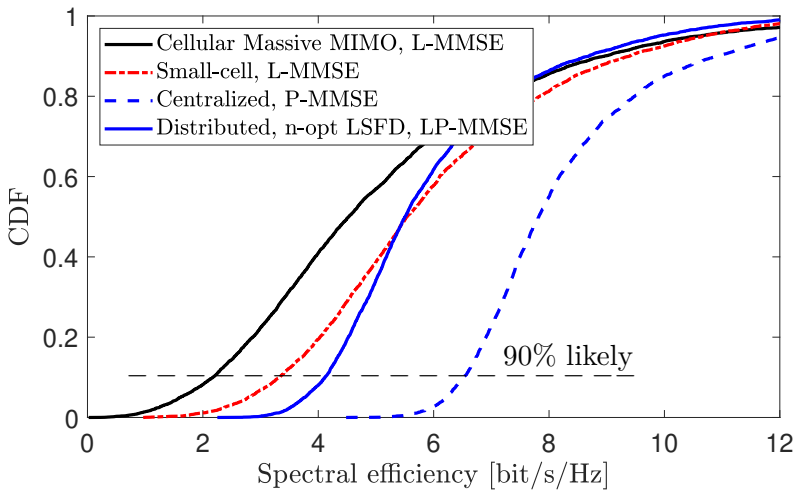
to be able to use the same pilot assignment, we drop the UEs in the coverage area one by one and assign the pilot to each according to Algorithm 4.1 on p. 289. Then, we check whether the serving AP in the Cellular Massive MIMO system can serve that UE with that assigned pilot (i.e., whether there is not any other UE that is served by this AP using the same pilot sequence). If this is the case, we assign this UE to the considered AP. Otherwise, we remove this randomly located UE and drop a new UE. At the end of this UE dropping methodology, it is guaranteed that all the setups consider the same UE locations and that the Cellular Massive MIMO system is well operating: each UE is connected to the AP with the largest channel gain and the 10 UEs in each cell are using mutually orthogonal pilot sequences.

Different scalable operations of User-centric Cell-free Massive MIMO are compared with Cellular Massive MIMO and small-cell systems in Figure 5.9. The figure shows the CDF of the SE achieved by UEs at different random locations. For the cellular systems, L-MMSE combining is used, which is the best scheme although it is unscalable. We have also observed that the gap between the achievable and genie-aided SEs for the small-cell system is negligibly small and we will present only the genie-aided SE results for the small-cell systems in the remainder of this section to avoid any underestimation of the small-cell system performance.

Figure 5.9(a) considers the case with $L = 400$ and $N = 1$. Among all the considered network architectures and processing methods, the centralized cell-free operation with P-MMSE combining provides the largest SEs. The CDF curve is to the right of all the other curves, which implies that all UEs benefit from using this mode of operation. The



(a) $L = 400$, $N = 1$ for the cell-free and small-cell setups.



(b) $L = 100$, $N = 4$ for the cell-free and small-cell setups.

Figure 5.9: CDF of the uplink SE per UE for Cellular Massive MIMO, a small-cell setup, and different operations of scalable Cell-free Massive MIMO. We consider $K = 40$, $\tau_p = 10$, and spatially correlated Rayleigh fading with ASD $\sigma_\varphi = \sigma_\theta = 15°$. For the Cellular Massive MIMO case, there are $L = 4$ APs with $N = 100$ antennas per AP.

distributed cell-free operation with n-opt LSFD and LP-MMSE provides nearly the same median SE as the small-cell and Cellular Massive MIMO systems but it gives substantially higher SE for the weakest UEs (i.e., more fairness). We can quantify the fairness gains by considering the 90% likely SE value (where the CDF is 0.1), which represents the SE that can be provided to 90% of all UEs. The 90% likely SE with the distributed cell-free operation is around 86% and 55% larger than in the small-cell and Cellular Massive MIMO systems, respectively. This is a natural consequence of both the architectural benefits of cell-free networks listed in Section 1.3 on p. 188 and the increased channel estimation quality observed from Figure 4.3 on p. 278. Qualitatively speaking, the results in this figure are consistent with the preliminary comparison in Figure 1.11 on p. 198. The centralized cell-free operation greatly outperforms the cellular networks, while the small-cell network is slightly preferred over Cellular Massive MIMO except in the upper and lower tails of the CDF curve. The distributed cell-free operation was not considered in Figure 1.11 and provides substantial gains over the cellular networks in the lower tail, but at the expense of lower performance in the upper tail.

Figure 5.9(b) considers the case with $L = 100$ and $N = 4$. The cellular curve is the same as before, the curves for the small-cell and distributed cell-free operations are shifted to the right, while the curve for the centralized cell-free operation is shifted to the left. In other words, the gap between the centralized and distributed cell-free operations reduces, thus showing that any distributed implementation should make use of multiple antennas per AP. Interestingly, the small-cell network outperforms Cellular Massive MIMO, because it combines the SNR benefits of a distributed operation with the ability to suppress interference with $N = 4$ antennas per AP. The 90% likely SE with the distributed cell-free operation is around 24% and 90% larger than in the small-cell and Cellular Massive MIMO systems, respectively. In conclusion, User-centric Cell-free Massive MIMO provides a more uniform service quality among the UEs (as seen from the steeper CDF curves) and, particularly, improves the service for the UEs with the weakest channel conditions (as seen from the locations of the lower tails of the curves). The centralized operation is by far the most desirable

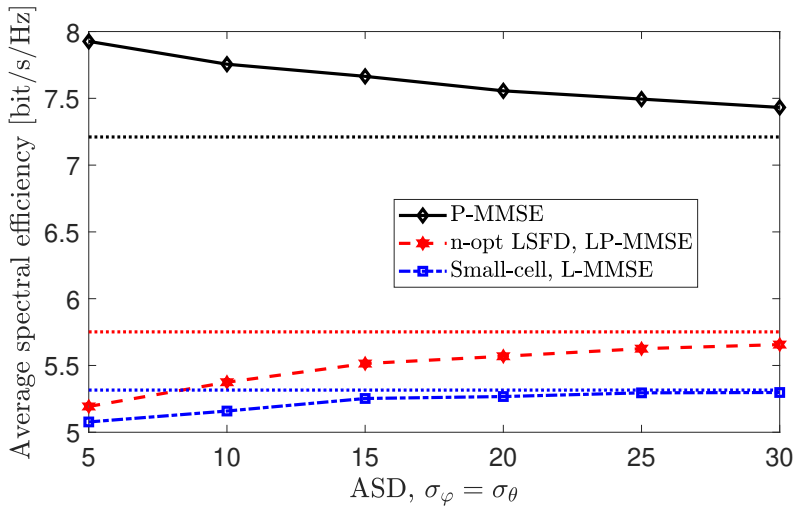


Figure 5.10: The average uplink SE per UE as a function of ASD for azimuth and elevation angles, $\sigma_\varphi = \sigma_\theta$ for different operations of Cell-free Massive MIMO and small-cell systems. We consider $L = 100$, $N = 4$, $K = 40$, and $\tau_p = 10$. The results for uncorrelated Rayleigh fading are shown as reference by the dotted lines.

from an SE performance perspective.

5.4.4 Impact of Spatial Correlation

In the remaining simulations of this section, we will consider $L = 100$ APs with $N = 4$ antennas per AP. Figure 5.10 analyzes the impact of spatial channel correlation. This figure shows the average SE per UE as a function of the ASD of the azimuth and elevation angular spread. We compare a cell-free network with scalable centralized and distributed operation to a small-cell setup with L-MMSE combining. The ASD for azimuth and elevation angles are the same, $\sigma_\varphi = \sigma_\theta$, and varied on the horizontal axis. The straight dotted lines present the reference average SE in the case of uncorrelated Rayleigh fading, where the spatial correlation matrices are scaled identities. We can think of uncorrelated fading as the limit case where the ASDs tend to infinity.

Note that smaller ASD means more highly correlated channels. We notice from Figure 5.10 that spatial correlation improves the average SE in the centralized operation whereas it degrades the average SE

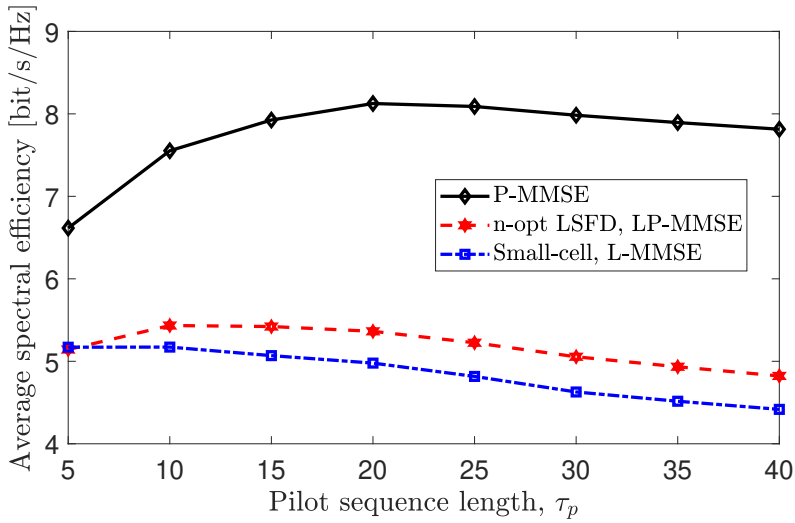


Figure 5.11: The average uplink SE per UE as a function of pilot sequence length, τ_p, for different operations of Cell-free Massive MIMO and small-cell setup. We consider $L = 100$, $N = 4$, $K = 40$, and spatially correlated Rayleigh fading with ASD $\sigma_\varphi = \sigma_\theta = 15°$.

in the distributed operation and small-cell system. When the ASD increases, the average SE with spatially correlated Rayleigh fading channels approaches the reference lines for the uncorrelated case; the difference is small when the ASDs are 30°. The intuition behind these results is that the centralized operation can exploit the spatial correlation to separate the UEs better (as is the case in Cellular Massive MIMO [**massivemimobook**]) since it co-processes the signals over many antennas. However, in the distributed operation and small-cell implementation, local combining is applied at each AP and then we are mainly subject to the negative effects of spatial correlation.

Another interesting observation is that the average SE of the small-cell network and the distributed Cell-free Massive MIMO are nearly the same when the channels are highly correlated (i.e., almost LoS), while the cell-free system benefits more from having lower spatial correlation.

5.4.5 Impact of the Pilot Sequence Length

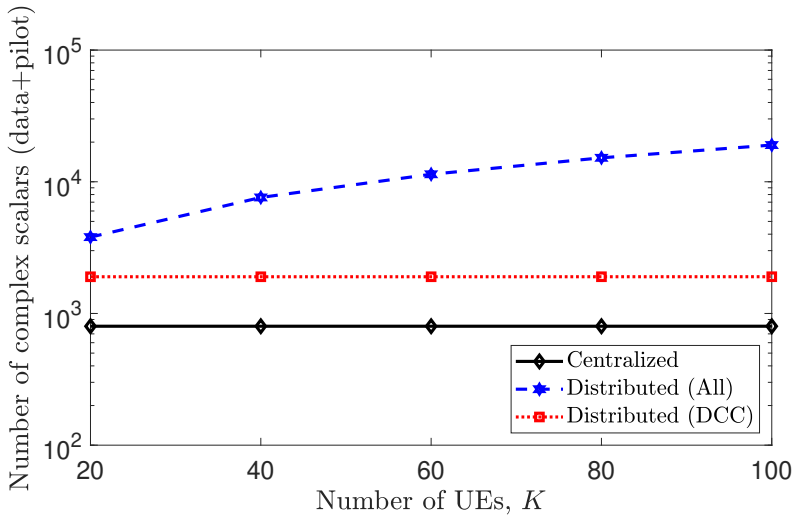
Figure 5.11 shows how the pilot sequence length, τ_p, impacts the average SE per UE. We recall that a larger τ_p will result in less pilot contamination when K is maintained fixed but also in a pre-log penalty since there is less room for data in every coherence block. For both the centralized and distributed cell-free operation, increasing the pilot sequence length improves the SE until a certain point, and then it decays with τ_p. There are several reasons for this. First, reducing pilot contamination improves the channel estimation quality and reduces the coherent interference. Moreover, each UE is now served by more APs since each AP is allowed to serve up to τ_p UEs in the considered scalable DCC and pilot assignment scheme. However, after some point (i.e., $\tau_p = 20$ in the centralized operation), the improvement saturates since the number of transmission symbols, τ_u, allocated to uplink data transmission also decreases. If we continue to increase the pilot length, the SE will eventually reduce because of the pre-log penalty. The saturation point occurs at a smaller value (i.e., $\tau_p = 10$) in the distributed cell-free operation, while $\tau_p = 5$ gives the highest SE in the small-cell system. The reason is the reduced ability to suppress interference, which does not improve much by increasing the estimation quality.

5.4.6 Scalability of Fronthaul and Computations

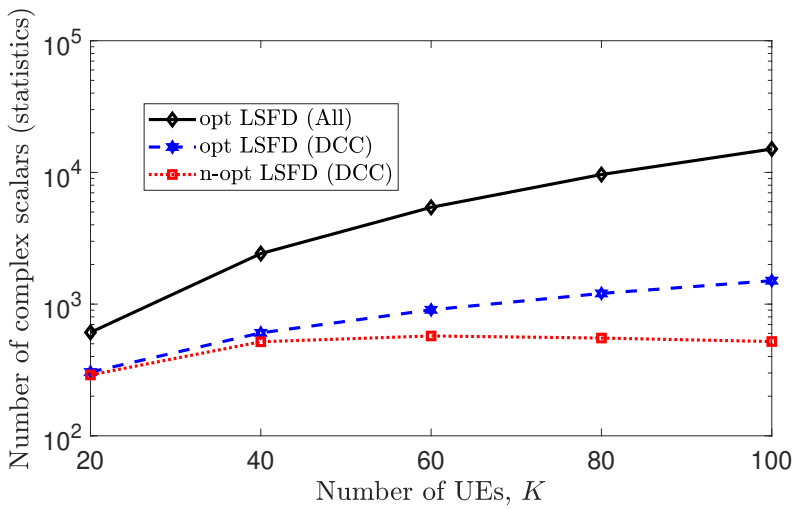
We will now consider the fronthaul signaling and computational complexity associated with different types of centralized and distributed schemes. We consider several random network realizations by dropping APs and UEs randomly in the simulation area and all the results in this section present the signaling and complexity per AP. Hence, they are obtained by averaging over both network realizations and all the APs. In addition, we count the operations and complex numbers once when they are common for more than one UE.

Fronthaul Signaling

To quantify the fronthaul signaling load of different cell-free operations, we present the average number of complex scalars to be sent from the APs to the CPU per AP in each coherence block in Figure 5.12(a).



(a) Data plus pilot signals per coherence block.



(b) Statistical parameters per channel statistics realization. No statistical parameters are exchanged with centralized operation.

Figure 5.12: The average number of complex scalars to send from the APs to the CPU over the fronthaul either in each coherence block or for each realization of the channel statistics as a function of the number K of UEs. We consider $L = 100$, $N = 4$, and $\tau_p = 10$. Both centralized and distributed operations are considered.

These scalars can represent both data and pilots, depending on the operation. The number of UEs is shown on the horizontal axis and a scalable fronthaul operation is represented by having a number that does not grow with K. We note that, for the centralized operation, the number of complex scalars to be sent from the APs to the CPU in each coherence block is the same irrespective of the number of UEs and DCC scheme, thus scalability is not an issue in terms of fronthaul capacity. For the distributed operation, the number of complex scalars related to the data signals depends on the DCC. We see that it grows with K if all APs serve all UEs, while it does not grow when using the proposed scalable DCC solution where each AP serves τ_p UEs. Interestingly, the fronthaul load is larger in the distributed operations than in the centralized operation. The reason is that every AP is serving more UEs than it has antennas (i.e., $\tau_p > N$), thus the local receive combining is increasing the dimension of the signals that need to be sent to the CPU. Hence, the main purpose of a distributed operation cannot be to reduce the fronthaul capacity but to make use of distributed processing capabilities.

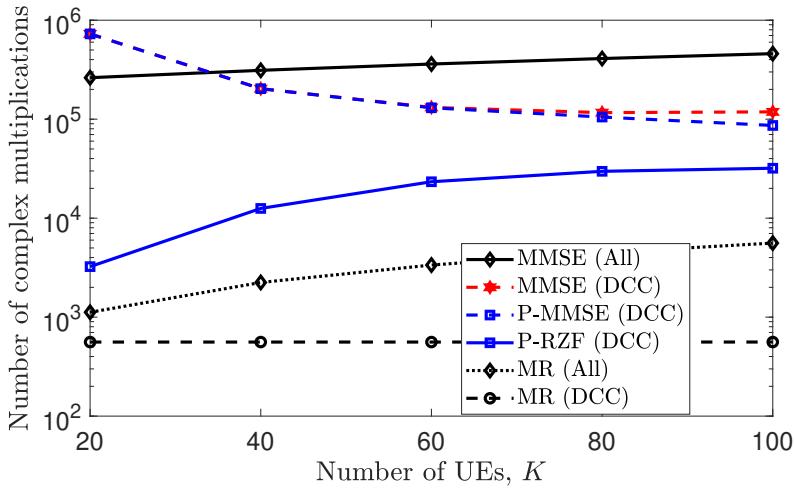
The number of statistical parameters that must be sent to the CPU in one realization of the network is shown in Figure 5.12(b). We note that, for the centralized operation, no statistical parameters need to be exchanged so there are no curves for this case. For the distributed operation with LSFD, there is a certain number of statistical parameters to be sent from each AP to the CPU in each realization of the channel statistics. The largest number of parameters is needed when using opt LSFD and serving all UEs using all APs. The number of scalars grows rapidly with the number of UEs. Even if we introduce DCC, the number of scalars grows with K if opt LSFD is used, which indicates that the fronthaul signaling load is not scalable. However, n-opt LSFD using DCC is a scalable alternative. We notice that there is a slight decrease in the number of statistical parameters that need to be sent to the CPU in this case. The reason is that as the number of UEs increases, the average number of APs serving a generic UE k decreases since each AP at most serves τ_p UEs. This in turn results in a decrease in the number of UEs that are partially served by the same APs as UE k, i.e., $|\mathcal{S}_k|$ in

Table 5.2.

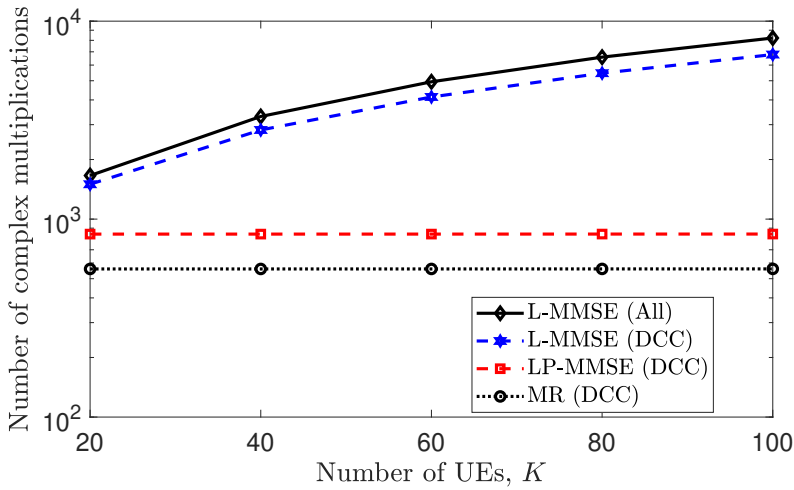
Computational Complexity

Figure 5.13 shows the average number of complex multiplications required for the computation of different centralized and distributed combining schemes. The results for the centralized operation in Figure 5.13(a) are based on Table 5.1 whereas the results for the distributed operation in Figure 5.13(b) are obtained from Table 5.3. Note that the matrix in (5.11) that needs to be inverted is the same when all APs serve all UEs and as all the other common operations, this is counted once for all the UEs. However, with DCC, the matrix inverse is taken separately for each UE resulting in more complex multiplications for $K = 20$ compared to MMSE combining where all APs serve all UEs, as can be seen in Figure 5.13(a). However, as the number of UEs increases, the computational complexity decreases in the DCC case while it increases when all APs serve all UEs. Hence, for a large number of UEs, the computational complexity can be reduced efficiently with a well-designed DCC. We know from previous figures that scalable methods can be utilized with a negligible SE loss and the benefit of doing that is the vast reduction in complexity observed in this figure. Another observation is that P-RZF might be useful when the number of UEs is relatively small, however, when K increases, the gap between P-MMSE and P-RZF decreases. The reason that the complexity of P-RZF increases with K is that the number of UEs $|\mathcal{S}_k|$ that are served by partially the same APs as UE k increases on the average when we increase K from 20 to 60. Although $|\mathcal{S}_k|$ decreases after $K = 60$, the overall effect of the other parameters from Table 5.1 results in such a pattern in the computational complexity of P-RZF combining.

From Figure 5.13(b), we observe that the average number of complex multiplications does not grow with K for the scalable combining LP-MMSE and MR schemes. However, the total complexity grows with K when using the unscalable L-MMSE schemes. Hence, scalability is important also in the distributed operation. Note that the vertical scale is substantially smaller in the distributed operation than in the centralized operation, because the distributed schemes are less computationally



(a) Centralized operation.



(b) Distributed operation.

Figure 5.13: The average number of complex multiplications required per coherence block to compute the local combining vectors, including the computation of the channel estimates as a function of the number K of UEs. We consider $L = 100$, $N = 4$, and $\tau_p = 10$. Different combining schemes for the centralized and distributed operations are considered.

complex.

5.5 Summary of the Key Points in Section 5

- The uplink of a cell-free network can be implemented with either centralized or distributed operation.
- In the centralized operation, the APs are forwarding all the uplink signals to the CPU, which carries out the channel estimation and receive combining. This operation enables centralized combining where the signals received at all the APs that serve UE k can be used to balance between obtaining a strong signal and suppressing interference.
- The uplink SE with the centralized operation is given in Theorem 5.1. It is maximized by MMSE combining, but this is not a scalable scheme. The alternative P-MMSE, P-RZF, and MR combining schemes are scalable and represent different tradeoffs between complexity and performance.
- In the distributed operation, each AP uses its locally received signals to compute local channel estimates and estimates of the uplink data. The local data estimates are then forwarded to the CPU, which computes a weighted average of the local estimates and then detects the data. The latter stage is called LSFD and should give higher priority to APs with good channel conditions.
- The uplink SE with the distributed operation is given in Theorem 5.4 and depends both on the local receive combining and on the LSFD weights. The optimal weights can be computed but are not scalable to implement, thus a nearly optimal but scalable alternative was derived. L-MMSE is the local receive combining that minimizes the MSE of the local data estimate, but it is not scalable. LP-MMSE and MR combining are scalable alternatives.

- We defined a running example and used it to evaluate the performance of cell-free networks. We noticed that a scalable operation can be achieved with a negligible SE loss compared to the case when all APs serve all UEs, but with a substantial reduction in computational complexity.
- The centralized operation can achieve substantially higher SE than the distributed operation, but both implementations outperform Cellular Massive MIMO and small-cell networks. The gain is particularly large for the UEs with the worst channel conditions, which demonstrates the ability of the cell-free architecture to increase the uniformity of the service. The centralized operation is associated with higher computational complexity than the distributed operation, but less fronthaul signaling.
- If the total number of AP antennas is fixed, a centralized operation prefers to have many single-antenna APs since this reduces the average distance between a UE and its closest APs, while the interference is mitigated by centralized receive combining. In a distributed operation, there is another tradeoff: having many single-antenna APs is beneficial for the UEs with weak channel conditions, while having fewer multi-antenna APs gives a local interference suppression capability at each AP that is beneficial for the UEs with good channel conditions.

6

Downlink Operation

This section describes two different downlink implementations of User-centric Cell-free Massive MIMO, which are the counterparts of the two uplink operations described in Section 5. Section 6.1 considers centralized operation in which the precoding is carried out at the CPU, based on the channel estimates that are obtained from the uplink pilot signals gathered from all APs. In this case, the CPU performs all the signal processing for channel estimation and precoding. The achievable SE is derived and different unscalable and scalable centralized transmit precoding schemes are presented. An uplink-downlink duality result is presented to motivate the precoding selection. Distributed operation is then considered in Section 6.2. In this case, each AP applies precoding on the basis of locally obtained channel estimates, while the CPU is only encoding the data. The SE is analyzed and different local precoding schemes are presented. A performance comparison is provided in Section 6.3, where the different schemes are analyzed in terms of SE. The key points are summarized in Section 6.4.

6.1 Centralized Downlink Operation

As in the uplink, the most advanced downlink implementation of Cell-free Massive MIMO is a fully centralized operation, where the APs only act as remote-radio heads or relays that transmit signals that the CPU has generated and sent out over the fronthaul links. To explain the setup in further detail, we first recall the downlink system model from Section 2.3.4 on p. 215. The received downlink signal at UE k is

$$y_k^{\mathrm{dl}} = \sum_{l=1}^{L} \mathbf{h}_{kl}^{\mathrm{H}} \mathbf{x}_l + n_k = \sum_{l=1}^{L} \mathbf{h}_{kl}^{\mathrm{H}} \left(\sum_{i=1}^{K} \mathbf{D}_{il} \mathbf{w}_{il} \varsigma_i \right) + n_k \tag{6.1}$$

where $n_k \sim \mathcal{N}_{\mathbb{C}}(0, \sigma_{\mathrm{dl}}^2)$ is the receiver noise and

$$\mathbf{x}_l = \sum_{i=1}^{K} \mathbf{D}_{il} \mathbf{w}_{il} \varsigma_i \tag{6.2}$$

is the signal transmitted by AP l. This signal is created as the sum of the UEs' signals where each term intended for UE i consists of the unit-power downlink data signal $\varsigma_i \in \mathbb{C}$ (with $\mathbb{E}\{|\varsigma_i|^2\} = 1$) and the effective transmit precoding vector

$$\mathbf{D}_{il} \mathbf{w}_{il} = \begin{cases} \mathbf{w}_{il} & l \in \mathcal{M}_i \\ \mathbf{0}_N & l \notin \mathcal{M}_i. \end{cases} \tag{6.3}$$

Recall that $\mathbf{D}_{il} = \mathbf{0}_{N \times N}$ implies $\mathbf{D}_{il} \mathbf{w}_{il} = \mathbf{0}_N$, thus AP l will only transmit to UE i in the downlink if $\mathbf{D}_{il} \neq \mathbf{0}_{N \times N}$. By utilizing the effective precoding vectors defined in (6.3), the signal in (6.2) can be equivalently expressed as a summation only over the $|\mathcal{D}_l|$ UEs served by AP l:

$$\mathbf{x}_l = \sum_{i=1}^{K} \mathbf{D}_{il} \mathbf{w}_{il} \varsigma_i = \sum_{i \in \mathcal{D}_l} \mathbf{w}_{il} \varsigma_i. \tag{6.4}$$

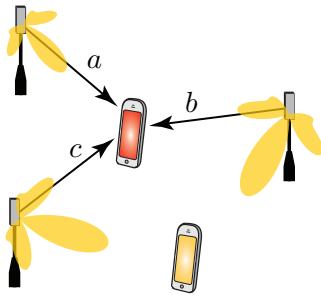
However, we will utilize the original formulation in (6.2) since it allows us to write the received signal in (6.1) in compact form as

$$y_k^{\mathrm{dl}} = \sum_{i=1}^{K} \begin{bmatrix} \mathbf{h}_{k1} \\ \vdots \\ \mathbf{h}_{kL} \end{bmatrix}^{\mathrm{H}} \begin{bmatrix} \mathbf{D}_{i1} \mathbf{w}_{i1} \\ \vdots \\ \mathbf{D}_{iL} \mathbf{w}_{iL} \end{bmatrix} \varsigma_i + n_k = \sum_{i=1}^{K} \mathbf{h}_k^{\mathrm{H}} \mathbf{D}_i \mathbf{w}_i \varsigma_i + n_k \tag{6.5}$$

where $\mathbf{h}_k = [\mathbf{h}_{k1}^{\mathrm{T}} \ \ldots \ \mathbf{h}_{kL}^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{C}^{LN}$ is the collective channel to UE k from all APs, $\mathbf{w}_i = [\mathbf{w}_{i1}^{\mathrm{T}} \ \ldots \ \mathbf{w}_{iL}^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{C}^{LN}$ is the collective precoding vector assigned to UE i, and the block-diagonal matrix $\mathbf{D}_i = \mathrm{diag}(\mathbf{D}_{i1}, \ldots, \mathbf{D}_{iL})$ identifies which APs are transmitting to UE i.

The system model formulation in (6.5) provides a network-wide perspective on the downlink transmission and is particularly useful when describing the centralized operation. The CPU estimates the downlink channels by exploiting the uplink-downlink channel reciprocity, which says that the uplink and downlink channels are identical within a coherence block.[1] Hence, we can utilize the estimation results from Section 4.2 on p. 266 to compute the MMSE estimates of the collective channels in the downlink. The estimates are utilized by the CPU to compute the collective precoding vectors $\{\mathbf{D}_i\mathbf{w}_i : i = 1, \ldots, K\}$ for all the K UEs in the network. The signal $\mathbf{x}_l$ in (6.2) is generated for each AP using these precoding vectors and the downlink data $\{\varsigma_i : i = 1, \ldots, K\}$. Each AP l must only be aware of its N-dimensional piece $\mathbf{D}_{il}\mathbf{w}_{il}$ of the collective precoding vector $\mathbf{D}_i\mathbf{w}_i$ of UE i, but the CPU can design the collective vector so that the different pieces fit well together. In particular, the CPU can select the precoding so that the APs can cancel out each others' interference in a way that each AP is unable to figure out individually. This feature is illustrated in Figure 6.1, where three APs are transmitting to the yellow UE but the signals are creating interference to the red UE. If the interference from the APs are represented by $a, b, c \in \mathbb{C}$, then the CPU can adjust the APs' transmit powers and phases so that $a + b + c \approx 0$. Hence, a viable solution is that each AP creates a large amount of interference but it is canceled by an equally strong amount of interference with the opposite sign from another AP: $b = -a$ and $|a| = |b| \gg 0$. Such multi-AP interference cancelation is not possible to implement in a distributed operation

[1] Even if the physical channel is the same in the uplink and the downlink, different pieces of transceiver hardware are being used in the two directions. This can create a mismatch between the uplink and downlink channels in practice. This mismatch can be estimated and compensated for using reciprocity calibration algorithms. We will not cover that in this monograph but refer to [**Guillaud2005a**], [**Nishimori2001a**], [**Rogalin2014a**], [**Shepard2012a**], [**Vieira2017a**], [**Vieira2014b**], [**Zetterberg2011a**] for details and algorithms.



Centralized interference cancelation:

$$a + b + c \approx 0$$

Distributed interference cancelation:

$$a \approx 0,\ b \approx 0,\ c \approx 0$$

Figure 6.1: When three APs transmit to a UE, they will cause interference to other neighboring UEs. In the centralized operation, the CPU can design the precoding so that the APs cancel out others' interference contributions: $a + b + c \approx 0$. In a distributed operation, where a, b, c are selected without cooperation, the only way to cancel interference is to make the individual contributions small: $a \approx 0$, $b \approx 0$, $c \approx 0$.

where the APs are unaware of each others' channels, making $a, b, c \approx 0$ the only viable way to limit the interference. We will return to the centralized and distributed precoding design later in this section.

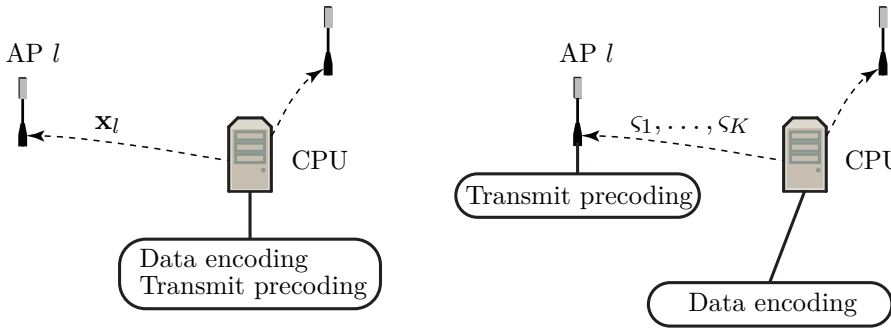
Figure 6.2(a) illustrates how the downlink signal processing is carried at the CPU in the centralized operation. We will describe the centralized operation as if the CPU performs the baseband processing while the conversion from digital baseband to analog passband signals is carried out at each AP, but it is also possible that each AP receives a passband signal from the CPU, which is then amplified before transmission.[2]

The expression in (6.5) is mathematically equivalent to the signal model of a downlink single-cell Massive MIMO system with correlated fading [**massivemimobook**] if one treats the CPU as a transmitter equipped with LN antennas. However, some key differences exist as also previously described in the uplink:

1. Multiple UEs that are managed by the CPU are using the same pilot, which is normally avoided in single-cell systems.

2. The antennas are distributed at different geographical locations. Hence, the collective channel is distributed as $\mathbf{h}_k \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_{LN}, \mathbf{R}_k)$ where the spatial correlation matrix $\mathbf{R}_k = \text{diag}(\mathbf{R}_{k1}, \ldots, \mathbf{R}_{kL}) \in \mathbb{C}^{LN \times LN}$ has a block-diagonal structure, which is normally not

[2]This can be implemented using radio-over-fiber technology.



(a) Centralized operation (b) Distributed operation

Figure 6.2: The downlink signal processing tasks can be divided between the APs and CPU in different ways. In the centralized operation, the data encoding and transmit precoding are done at the CPU. In the distributed operation, everything except the data encoding is done at the APs.

the case in single-cell systems.

3. Many elements of the effective precoding vector $\mathbf{D}_k\mathbf{w}_k$ are zero, namely the ones corresponding to APs that are not serving UE k.

4. The precoding vectors should be chosen to satisfy per-AP power constraints, instead of a total power constraint as in single-cell Massive MIMO systems. More precisely, we require that

$$\mathbb{E}\left\{\|\mathbf{x}_l\|^2\right\} \leq \rho_{\max}, \quad l = 1, \ldots, L \tag{6.6}$$

where $\rho_{\max} \geq 0$ denotes the maximum downlink transmit power of an AP. For notational convenience, $\rho_{\max}$ is assumed to be the same for all APs. Note that the expectation in (6.6) is over both the data signals and channel realizations. The motivation is that we consider a system where the available bandwidth resources are divided into many coherence blocks. The AP's power amplifier will simultaneously transmit signals over these many blocks with independent channel realizations, thus we can exploit the ability to assign different amounts of energy to different coherence blocks by using a power constraint of the type in (6.6).

These differences play an important role in the operation of cell-free

networks and differentiate the SE analysis in this monograph from what can be found in the literature on Cellular Massive MIMO.

6.1.1 Spectral Efficiency With Centralized Operation

We will now derive an achievable SE expression that applies when using any precoding scheme. The received signal in (6.5) for UE k can be divided into three terms:

$$y_k^{\mathrm{dl}} = \underbrace{\mathbf{h}_k^{\mathrm{H}} \mathbf{D}_k \mathbf{w}_k \varsigma_k}_{\text{Desired signal}} + \underbrace{\sum_{\substack{i=1 \\ i \neq k}}^{K} \mathbf{h}_k^{\mathrm{H}} \mathbf{D}_i \mathbf{w}_i \varsigma_i}_{\text{Inter-user interference}} + \underbrace{n_k}_{\text{Noise}}. \tag{6.7}$$

The UE wants to extract the data from the first term under the presence of inter-user interference and noise, which are represented by the latter two terms. The expression $\mathbf{h}_k^{\mathrm{H}} \mathbf{D}_k \mathbf{w}_k$ that is multiplied with the desired signal ς_k is called the *effective downlink channel*. This terminology refers to the fact that the precoding is effectively turning the multiple-antenna channel $\mathbf{h}_k$ into the effective single-antenna channel $\mathbf{h}_k^{\mathrm{H}} \mathbf{D}_k \mathbf{w}_k$. The CPU has partial knowledge about the effective channel. It knows the precoding vector (since it is the one selecting it) and it also knows the partial MMSE estimate

$$\mathbf{D}_k \widehat{\mathbf{h}}_k = \begin{bmatrix} \mathbf{D}_{k1} \widehat{\mathbf{h}}_{k1} \\ \vdots \\ \mathbf{D}_{kL} \widehat{\mathbf{h}}_{kL} \end{bmatrix} \sim \mathcal{N}_{\mathbb{C}} \left(\mathbf{0}_{LN}, \eta_k \tau_p \mathbf{D}_k \mathbf{R}_k \boldsymbol{\Psi}_{t_k}^{-1} \mathbf{R}_k \mathbf{D}_k \right) \tag{6.8}$$

of the channel. This partial MMSE channel estimate was previously defined in (4.20) along with the corresponding estimation error $\mathbf{D}_k \widetilde{\mathbf{h}}_k = \mathbf{D}_k \mathbf{h}_k - \mathbf{D}_k \widehat{\mathbf{h}}_k \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_{LN}, \mathbf{D}_k \mathbf{C}_k)$ with $\mathbf{C}_k = \mathrm{diag}(\mathbf{C}_{k1}, \ldots, \mathbf{C}_{kL})$.

UE k can transmit in the uplink without knowing the channel but it cannot decode the downlink data without utilizing some information about the effective channel $\mathbf{h}_k^{\mathrm{H}} \mathbf{D}_k \mathbf{w}_k$. We have not defined any functionality for estimating this term. This is not as strange as it might seem but it is a common practice in communication theory to only send pilot signals in one direction (e.g., the uplink) and then let the entity that obtains those estimates transmit in such a way that the effective

channel becomes (approximately) a positive scalar that the receiver can estimate without the need for sending pilot signals.

Remark 6.1 (Pilots in only uplink direction)**.** Consider a scalar channel $g \in \mathbb{C}$ which is known at the AP but not at the UE. The AP can then precode the data signal ς using $g^\star/|g|$ so that the received signal becomes $gg^\star/|g|\varsigma = |g|\varsigma$. The UE still does not know the channel, but it knows that the effective channel $|g|$ is positive and it knows the power of the data signal ς. The UE can therefore deduce the effective channel by computing the sample average power of the received signal over the received data signals in a coherence block (where the channel $|g|$ is fixed but ς is changing) [**Ngo2017a**]. However, if the channel is only partially known at the AP (e.g., due to estimation errors or imperfect hardware calibration between uplink and downlink), then it might be necessary to transmit one or multiple downlink pilots as well [**Interdonato2019b**] (e.g., dedicated demodulation reference signals or using a differential modulation scheme). For historical reasons, this is commonly done in practical networks but will not be considered in this monograph.

We will compute an SE expression for the case when the UE knows the average value $\mathbb{E}\{\mathbf{h}_k^{\mathrm{H}}\mathbf{D}_k\mathbf{w}_k\}$ of the effective channel. This is a deterministic number and can, thus, be easily obtained in practice; deterministic numbers are always assumed known in the capacity analysis since the data transmission spans over infinitely many transmission symbols and coherence blocks. We cannot reuse the capacity bound that was provided in Theorem 5.1 on p. 298 for the centralized operation in the uplink, but instead, we will utilize the alternative SE expression in Theorem 5.2 on p. 300. We then obtain the following result.

Theorem 6.1. An achievable SE of UE k in the downlink with centralized operation is

$$\mathsf{SE}_k^{(\mathrm{dl,c})} = \frac{\tau_d}{\tau_c} \log_2 \left(1 + \mathsf{SINR}_k^{(\mathrm{dl,c})}\right) \quad \text{bit/s/Hz} \tag{6.9}$$

where $\mathsf{SINR}_k^{(\mathrm{dl,c})}$ is the effective SINR

$$\mathsf{SINR}_k^{(\mathrm{dl,c})} = \frac{\left|\mathbb{E}\left\{\mathbf{h}_k^{\mathrm{H}}\mathbf{D}_k\mathbf{w}_k\right\}\right|^2}{\sum\limits_{i=1}^{K} \mathbb{E}\left\{\left|\mathbf{h}_k^{\mathrm{H}}\mathbf{D}_i\mathbf{w}_i\right|^2\right\} - \left|\mathbb{E}\left\{\mathbf{h}_k^{\mathrm{H}}\mathbf{D}_k\mathbf{w}_k\right\}\right|^2 + \sigma_{\mathrm{dl}}^2}. \tag{6.10}$$

Proof. The proof is given in Appendix C.3.1 on p. 453. □

The pre-log factor τ_d/τ_c in (6.9) is the fraction of each coherence block that is used for downlink data transmission. The term $\mathsf{SINR}_k^{(\mathrm{dl,c})}$ in (6.10) takes the form of an effective SINR, which means that the SE matches the capacity of a deterministic single-antenna single-user AWGN channel with an SNR that equals $\mathsf{SINR}_k^{(\mathrm{dl,c})}$. It also means that the data signal can be encoded and the received signal can be decoded as if we would be communicating over such an AWGN channel. Hence, there is a known way to achieve the SE stated in Theorem 6.1. The numerator of (6.10) contains the square of the average effective channel. The denominator equals the total power $\mathbb{E}\{|y_k^{\mathrm{dl}}|^2\} = \sum_{i=1}^{K} \mathbb{E}\{|\mathbf{h}_k^{\mathrm{H}}\mathbf{D}_i\mathbf{w}_i|^2\} + \sigma_{\mathrm{dl}}^2$ of the received signal minus the useful term from the numerator.

The SE expression in Theorem 6.1 holds for any transmit precoding vector $\mathbf{w}_k$ and selection of the DCC. In fact, it also holds for any channel distribution and not only correlated Rayleigh fading, as assumed in this monograph. The expression can be computed for any $\mathbf{w}_k$ by using Monte Carlo methods, which means that each expectation is approximated with the sample average over a large number of random realizations. More precisely, we can generate realizations of the channel estimates in a large set of coherence blocks, compute each term in (6.10) for each realization, and then take the average over all these computations. This is the approach that we will use when evaluating the SE numerically later in this section.

6.1.2 Centralized Transmit Precoding

The SE expression in Theorem 6.1 depends on the precoding of all UEs in the entire network. This makes the selection of transmit precoding vectors much more complicated than selecting receive combining vectors, which can be designed/optimized for one UE at a time. In particular, we noticed that the uplink SINR in (5.10) took the form of a generalized Rayleigh quotient and, thus, could be maximized with respect to the receive combining in closed form. There is no corresponding result for the downlink. The intuition behind this difference is that the precoding determines how the signals are emitted from the APs. For whatever

selected precoding, every UE will be interfered with by the transmission to the other UEs, even if the impact can hopefully be made marginal. In contrast, in the uplink, the received signals are not affected by the receive combining. However, after the reception, we take the received signals from a set of APs and process them to detect the signal from a particular UE. We can use the same received signals to detect the signals from another UE using another combining vector.

The effective downlink SINR in (6.10) has a very different structure than the uplink SINR in (5.9) or (5.10), hence, the precoding cannot be optimized on a per-UE basis. In fact, there is not a single collection of optimal precoding vectors but it all depends on what tradeoff between the SEs achieved by the different UEs that we want to achieve. Finding the optimal transmit precoding is computationally complicated in most cases [**Bjornson2013d**] but there is a common heuristic that can be utilized to deduce the structure of the optimal precoding [**Bjornson2014d**]. It is obtained from the following uplink-downlink duality result [**Bjornson2020a**], which particularizes the classical duality results [**Boche2002a**], [**Tse2005a**], [**Viswanath2003a**] for cell-free network architectures.

Theorem 6.2. Let $\{\mathbf{D}_i\mathbf{v}_i : i = 1, \ldots, K\}$ and $\{p_i : i = 1, \ldots, K\}$ denote the set of combining vectors and transmit powers, respectively, used in the uplink. If the precoding vectors are selected as

$$\mathbf{w}_i = \sqrt{\rho_i} \frac{\mathbf{v}_i}{\sqrt{\mathbb{E}\left\{\|\mathbf{D}_i\mathbf{v}_i\|^2\right\}}} \tag{6.11}$$

then there exists a downlink power allocation policy $\{\rho_i : i = 1, \ldots, K\}$ with $\sum_{i=1}^{K} \rho_i/\sigma_{\text{dl}}^2 = \sum_{i=1}^{K} p_i/\sigma_{\text{ul}}^2$ for which

$$\mathsf{SINR}_k^{(\text{dl},\mathsf{c})} = \mathsf{SINR}_k^{(\text{ul},\mathsf{c-UatF})}, \quad k = 1, \ldots, K \tag{6.12}$$

where $\mathsf{SINR}_k^{(\text{ul},\mathsf{c-UatF})}$ is the effective uplink SINR of UE k from (5.9).

Proof. The proof is given in Appendix C.3.2 on p. 454. □

This theorem shows that the effective SINRs that are achieved in the uplink (when using the SE expression from Theorem 5.2 on p. 300)

are also achievable in the downlink. Consequently, we have that an achievable downlink SE for UE k is

$$\mathsf{SE}_k^{(\mathrm{dl,c})} = \frac{\tau_d}{\tau_c} \log_2 \left(1 + \mathsf{SINR}_k^{(\mathrm{ul,c-UatF})} \right). \tag{6.13}$$

To achieve this SE, the precoding must be selected according to (6.11), which implies that the precoding vector for UE i should be a scaled version of the combining vector that is assigned to this UE in the uplink. The scaling factors $\{\rho_i : i = 1, \ldots, K\}$ represent the downlink power allocation. The intuition behind the uplink-downlink duality in Theorem 6.2 is that the direction from which the signal is best received in the uplink coincides with the direction in which the signal should be transmitted in the downlink. Note that the word "direction" does not refer to a distinct angular direction in our three-dimensional world but the direction of a vector in the LN-dimensional vector space where the received and transmitted signals exist.

If the noise is the same in uplink and downlink (i.e., $\sigma_{\mathrm{ul}}^2 = \sigma_{\mathrm{dl}}^2$), then Theorem 6.2 implies that the total transmit power is the same in both directions but is allocated differently between the UEs. The exact way of selecting the downlink power allocation $\{\rho_i : i = 1, \ldots, K\}$ can be found in (C.26) on p. 455 but we will not use this expression because there are multiple reasons why this power allocation is not used in practice. On the one hand, each AP might be allowed to transmit with substantially higher power than each UE and we also expect that more APs than UEs exist in cell-free networks. Hence, the total available downlink power might far exceed the total uplink power and this should be utilized to achieve the maximum downlink performance. On the other hand, the power allocation obtained from the uplink-downlink duality might require that some APs transmit with very high power (beyond what is allowed by the power constraint in (6.6)) while other APs might transmit with very low power. The bottom line is that we will utilize Theorem 6.2 as a motivation to heuristically select the downlink precoding vectors based on the uplink combining vectors according to (6.11). However, we will optimize the downlink transmit power separately instead of relying on the duality theorem. We will consider an arbitrary downlink power allocation in this section and then optimize it in Section 7.1.2 on p. 399.

In Section 5.1.3 on p. 301 and Section 5.1.4 on p. 305, we presented several receive combining schemes for the centralized operation which we can utilize as the basis for the downlink precoding, motivated by the uplink-downlink duality. To make the transformation simple, we define the centralized precoding vector for UE k as

$$\mathbf{w}_k = \sqrt{\rho_k} \frac{\bar{\mathbf{w}}_k}{\sqrt{\mathbb{E}\left\{\|\bar{\mathbf{w}}_k\|^2\right\}}} \tag{6.14}$$

where $\rho_k \geq 0$ is the total transmit power assigned to UE k from all the serving APs and $\bar{\mathbf{w}}_k \in \mathbb{C}^{LN}$ is an arbitrarily scaled vector that points out the direction of the precoding vector. Note that the normalization in (6.14) guarantees that

$$\mathbb{E}\{\|\mathbf{w}_k\|^2\} = \rho_k. \tag{6.15}$$

Any choice of $\bar{\mathbf{w}}_k$ will lead to a precoding scheme but scalability remains to be an important issue in the downlink operation. If we select $\bar{\mathbf{w}}_k$ based on a scalable uplink combining scheme, then the scalability property carries over to the downlink precoding.

MMSE Precoding

The optimal centralized uplink operation is using MMSE combining as defined in (5.11). Based on the uplink-downlink duality, the downlink counterpart is *MMSE precoding*, which is obtained from (6.14) using

$$\bar{\mathbf{w}}_k^{\mathrm{MMSE}} = p_k \left(\sum_{i=1}^{K} p_i \mathbf{D}_k (\widehat{\mathbf{h}}_i \widehat{\mathbf{h}}_i^{\mathrm{H}} + \mathbf{C}_i) \mathbf{D}_k + \sigma_{\mathrm{ul}}^2 \mathbf{I}_{LN} \right)^{-1} \mathbf{D}_k \widehat{\mathbf{h}}_k. \tag{6.16}$$

While MMSE combining is provably optimal in the uplink, MMSE precoding is only optimal if we want to operate the downlink transmission in the way specified by the duality in Theorem 6.2. One can then show that MMSE precoding is minimizing a sum MSE between the vector of data signals and the received signal vector of all UEs [**Vojcic1998a**]. In general, we want to make use of the flexibility to allocate the downlink power differently, thus we call MMSE precoding *nearly optimal* rather than optimal. Intuitively, MMSE precoding will balance between transmitting a strong signal to the desired UE and

limiting the interference caused to other UEs. We will present several simplified precoding methods below and while we expect MMSE precoding to outperform them in terms of SE, there are only theoretical guarantees of this when using the power allocation specified by the duality theorem.

If MMSE combining is used in the uplink, then MMSE precoding can be used in the downlink without incurring any additional computational complexity (except for the scaling in (6.14) which can be absorbed into the generation of the data signals). However, we know from Table 5.1 on p. 306 that the number of complex multiplications, regarding channel estimation and combining vector computation, grows with K when using MMSE combining, thus MMSE precoding is not scalable for networks with a large number of UEs.

P-MMSE and P-RZF Precoding

Two scalable combining schemes were presented in Section 5.1.4 on p. 305 as an approximation of the optimal MMSE combining. These were called P-MMSE combining in (5.16) and P-RZF combining in (5.19). Motivated by the uplink-downlink duality, we can use these to develop two scalable precoding schemes. The first one is *P-MMSE precoding*, which is obtained from (6.14) using

$$\bar{\mathbf{w}}_k^{\mathrm{P-MMSE}} = p_k \left(\sum_{i \in \mathcal{S}_k} p_i \mathbf{D}_k \widehat{\mathbf{h}}_i \widehat{\mathbf{h}}_i^{\mathrm{H}} \mathbf{D}_k + \mathbf{Z}_{\mathcal{S}_k} + \sigma_{\mathrm{ul}}^2 \mathbf{I}_{LN} \right)^{-1} \mathbf{D}_k \widehat{\mathbf{h}}_k \quad (6.17)$$

where $\mathcal{S}_k$ from (5.15) is the set of UEs served by partially the same APs as UE k and $\mathbf{Z}_{\mathcal{S}_k}$ was defined in (5.17) as the total estimation error correlation matrix related to those UEs. This precoding scheme is expected to behave similarly to MMSE precoding since it has the same structure, except that it neglects UEs that are only served by other APs, in order to reduce the computational complexity. If the DCC is properly designed, then $\mathcal{S}_k$ should include all the UEs that are within the range of influence of the APs that serve UE k, making the performance difference between MMSE and P-MMSE precoding small.

The inverse of an $LN \times LN$ matrix must be computed in (6.17) but, since only the signals transmitted from the $|\mathcal{M}_k|$ APs that serve

UE k matter, we can implement P-MMSE precoding by only inverting an $N|\mathcal{M}_k| \times N|\mathcal{M}_k|$ matrix (see Figure 5.3 on p. 304 for details). If P-MMSE combining is used in the uplink, then we can use P-MMSE precoding in the downlink without any extra computation.

As in the uplink, we can reduce the computational complexity by using *P-RZF precoding* instead, which is obtained by neglecting the estimation error correlation matrix $\mathbf{Z}_{\mathcal{S}_k}$ in (6.17). This enables us to rewrite the precoding expression in a more efficient form. P-RZF precoding is obtained from (6.14) using

$$\bar{\mathbf{w}}_k^{\mathrm{P-RZF}} = \left[\mathbf{D}_k \widehat{\mathbf{H}}_{\mathcal{S}_k} \left(\widehat{\mathbf{H}}_{\mathcal{S}_k}^{\mathrm{H}} \mathbf{D}_k \widehat{\mathbf{H}}_{\mathcal{S}_k} + \sigma_{\mathrm{ul}}^2 \mathbf{P}_{\mathcal{S}_k}^{-1} \right)^{-1} \right]_{:,1} \tag{6.18}$$

where we recall that $\widehat{\mathbf{H}}_{\mathcal{S}_k} \in \mathbb{C}^{LN \times |\mathcal{S}_k|}$ is obtained by stacking all the vectors $\widehat{\mathbf{h}}_i$ with indices $i \in \mathcal{S}_k$ with the first column being $\widehat{\mathbf{h}}_k$. Moreover, $\mathbf{P}_{\mathcal{S}_k} \in \mathbb{R}^{|\mathcal{S}_k| \times |\mathcal{S}_k|}$ is a diagonal matrix containing the transmit powers p_i for $i \in \mathcal{S}_k$, listed in the same order as the columns of $\widehat{\mathbf{H}}_{\mathcal{S}_k}$. If P-RZF combining is used in the uplink, then we can use P-RZF precoding in the downlink, without requiring any additional computational complexity. This makes it a scalable precoding scheme.

Note that the names "P-MMSE" and "P-RZF" are referring to the properties of the combining vector counterparts in the dual uplink. Intuitively, both methods will balance between transmitting a strong signal to the desired UE and limiting the interference that is caused to other UEs. The difference is that the former scheme takes the channel estimation uncertainty into account, while the latter is neglecting it to reduce the computational complexity.

6.1.3 Fronthaul Signaling Load for Centralized Operation

The centralized operation is associated with a fronthaul signaling load since all the computations are made at the CPU. The signaling related to sending the received pilot signals from the APs to the CPU is the same as for fully centralized operation in the uplink; see Section 5.1.5 on p. 310. Hence, we will not list them in this section to avoid counting them twice and instead focus on the signaling from the CPU to the APs that is unique to the downlink.

Table 6.1: Number of complex scalars to be shared over the fronthaul per coherence block in the centralized downlink operation. These scalars are sent from the CPU to the APs.

Scheme	Each coherence block
Any precoding	$\tau_d NL$

The only additional fronthaul signaling in the downlink is to provide each AP l with the transmit signal $\mathbf{x}_l$ in (6.2), which is a superposition of the precoded downlink data signals. There are τ_d such vectors to transmit per coherence block. Hence, the fronthaul signaling per AP is $\tau_d N$ complex scalars. This value is the same irrespective of the choice of precoding scheme and is summarized in Table 6.1. Interestingly, the APs can be totally unaware of what precoding scheme is used and how many UEs are being served since the CPU is performing all the signal processing in the centralized operation. Note that the fronthaul signaling is independent of K and, thus, it is scalable.

6.2 Distributed Downlink Operation

A distributed operation is also possible in the downlink where almost all the processing is carried out locally at each AP. The CPU encodes the downlink data signals $\{\varsigma_i : i = 1, \ldots, K\}$ and send them to the serving APs, which carry out the remaining signal processing, as illustrated in Figure 6.2(b). We stress that the CPU is a logical entity; the task of encoding the data signal to a given UE can be carried out anywhere in the network, thus there is no need for having a physical centralized unit. A major benefit of the distributed operation is that we can deploy new APs without having to upgrade the computational power of the CPU since each AP contains a local processor that can carry out its associated baseband processing tasks. In the distributed operation, each AP l locally designs its transmitted signal

$$\mathbf{x}_l = \sum_{i=1}^{K} \mathbf{D}_{il} \mathbf{w}_{il} \varsigma_i \tag{6.19}$$

that was previously stated in (6.2). To this end, AP l computes its local channel estimates as discussed in Section 4.2 on p. 266 and selects the local precoding vectors $\{\mathbf{w}_{il} : i \in \mathcal{D}_l\}$ based on those estimates.

Recall that, in the distributed uplink operation, we considered a two-stage operation where each AP is first processing its signals locally, followed by a second step where the CPU weighs the information obtained from the different APs together. We then considered how to optimize the weight vector. A similar operation exists in the downlink but only implicitly. The second stage is namely represented by the power allocation between the different APs, which determines which power each AP should assign to a given UE. This power allocation task is analyzed in Section 7.1.2 on p. 399, while we will consider a heuristic power allocation in this section.

6.2.1 Spectral Efficiency With Distributed Operation

The received signal at UE k is

$$\begin{aligned} y_k^{\mathrm{dl}} &= \sum_{l=1}^{L} \mathbf{h}_{kl}^{\mathrm{H}} \mathbf{x}_l + n_k \\ &= \underbrace{\left(\sum_{l=1}^{L} \mathbf{h}_{kl}^{\mathrm{H}} \mathbf{D}_{kl} \mathbf{w}_{kl} \right) \varsigma_k}_{\text{Desired signal}} + \underbrace{\sum_{\substack{i=1 \\ i \neq k}}^{K} \left(\sum_{l=1}^{L} \mathbf{h}_{kl}^{\mathrm{H}} \mathbf{D}_{il} \mathbf{w}_{il} \right) \varsigma_i}_{\text{Inter-user interference}} + \underbrace{n_k}_{\text{Noise}} \end{aligned} \tag{6.20}$$

which is identical to the case of centralized downlink operation in (6.1). The key difference lies in the precoding selection, which now must be carried out locally at each AP using the locally available channel estimates. The information available at the UE for signal detection is the same in both cases. Hence, we can use Theorem 6.1 to also compute the SE achieved by the distributed operation. We will restate the result using the distributed notation containing summations over the APs.

Corollary 6.3. An achievable SE of UE k in the distributed operation is

$$\mathsf{SE}_k^{(\mathrm{dl},\mathsf{d})} = \frac{\tau_d}{\tau_c} \log_2 \left(1 + \mathsf{SINR}_k^{(\mathrm{dl},\mathsf{d})} \right) \quad \text{bit/s/Hz} \tag{6.21}$$

with the effective SINR given by

$$\mathsf{SINR}_k^{(\mathrm{dl,d})} = \frac{\left| \sum_{l=1}^{L} \mathbb{E}\left\{ \mathbf{h}_{kl}^{\mathrm{H}} \mathbf{D}_{kl} \mathbf{w}_{kl} \right\} \right|^2}{\sum_{i=1}^{K} \mathbb{E}\left\{ \left| \sum_{l=1}^{L} \mathbf{h}_{kl}^{\mathrm{H}} \mathbf{D}_{il} \mathbf{w}_{il} \right|^2 \right\} - \left| \sum_{l=1}^{L} \mathbb{E}\left\{ \mathbf{h}_{kl}^{\mathrm{H}} \mathbf{D}_{kl} \mathbf{w}_{kl} \right\} \right|^2 + \sigma_{\mathrm{dl}}^2}. \tag{6.22}$$

The SE of UE k and its effective SINR can be interpreted in the same way as in the centralized case so we will not repeat the discussion here, but instead focus on the local precoding selection.

6.2.2 Local Transmit Precoding

Only a subset of the APs are transmitting a downlink data signal to a particular UE k. Hence, the effective precoding vectors $\mathbf{D}_{kl}\mathbf{w}_{kl}$ only need to be selected for AP $l \in \mathcal{M}_k$. For an AP l that serves UE k, we can express the precoding vector as

$$\mathbf{w}_{kl} = \sqrt{\rho_{kl}} \frac{\bar{\mathbf{w}}_{kl}}{\sqrt{\mathbb{E}\left\{ \left\| \bar{\mathbf{w}}_{kl} \right\|^2 \right\}}} \tag{6.23}$$

where $\rho_{kl} \geq 0$ is the transmit power that AP l assigns to UE k and $\bar{\mathbf{w}}_{kl} \in \mathbb{C}^N$ is an arbitrarily scaled vector pointing out the direction of the precoding vector. Note that the normalization in (6.23) makes

$$\mathbb{E}\left\{ \left\| \mathbf{w}_{kl} \right\|^2 \right\} = \rho_{kl}. \tag{6.24}$$

Using this notation, we can effectively divide the precoding selection at AP l into the following two subtasks:

1. Selecting the directivity of the transmission represented by $\{\bar{\mathbf{w}}_{kl} : k \in \mathcal{D}_l\}$;
2. Selecting the power allocation represented by $\{\rho_{kl} : k \in \mathcal{D}_l\}$.

The first task must be carried out in every coherence block, based on the local channel estimates, which change at this pace. In contrast, the power allocation is dominated by large-scale effects (such as pathloss)

and we will, therefore, assume that the power allocation is maintained constant over many coherence blocks and computed based on the channel statistics. We will optimize the power allocation in Section 7.1.2 on p. 399 and consider scalable heuristic methods in Section 7.2 on p. 410.

In the centralized downlink operation, we used the uplink-downlink duality in Theorem 6.2 to motivate that each centralized precoding vector should be selected to be parallel to the corresponding centralized combining vector. The duality result is not unique to the centralized operation: for any choice of the receive combining vectors and their resulting uplink SINRs, we can achieve the same SINR in the downlink by using the same vectors for precoding. Hence, we can utilize the local combining schemes from Section 5.2 on p. 311 and normalize them to obtain local precoding vectors that should serve as reasonable heuristics. The complexity of computing these vectors is the same as in the uplink. Hence, if the same vectors are used in both directions then there is no extra complexity incurred in the downlink. We will now present the downlink counterparts of the L-MMSE, LP-MMSE, and MR combining schemes that we considered in the uplink.

L-MMSE Precoding

The locally optimal operation in the uplink is L-MMSE combining, which was defined in (5.29). The downlink counterpart is *L-MMSE precoding* and is obtained from (6.23) using

$$\bar{\mathbf{w}}_{kl}^{\mathrm{L-MMSE}} = p_k \left(\sum_{i=1}^{K} p_i \left(\widehat{\mathbf{h}}_{il} \widehat{\mathbf{h}}_{il}^{\mathrm{H}} + \mathbf{C}_{il} \right) + \sigma_{\mathrm{ul}}^2 \mathbf{I}_N \right)^{-1} \mathbf{D}_{kl} \widehat{\mathbf{h}}_{kl}. \tag{6.25}$$

We expect this precoding method to provide the highest SE among the distributed alternatives, due to its tight connection to the L-MMSE combining method, but this cannot be formally proved. Roughly the same power gain from coherent precoding can be achieved by centralized MMSE precoding and L-MMSE precoding, if the same APs are utilized. However, there is a major difference when it comes to interference suppression. As illustrated in Figure 6.1, in the centralized operation, the serving APs can cooperate in suppressing interference; for example, by sending interfering signals with opposite phases so they cancel each

other at an undesired receiver. Hence, the spatial degrees-of-freedom available for interference suppression is equal to the total number of antennas that is transmitting the signal and given by $N|\mathcal{M}_k|$ for UE k. In contrast, in the distributed operation, each AP can only control the interference that itself is causing. Each AP has N spatial degrees-of-freedom available for interference suppression, which is expected to be a rather small number since each AP in a Cell-free Massive MIMO system is envisioned to be equipped with few antennas. In fact, each AP might serve more UEs than it has antennas. With the distributed operation, the interference suppression capability does not increase as more APs are assigned to serving the UE.

L-MMSE precoding in (6.25) is not a scalable scheme since it contains a summation of all UEs in the network (see Table 5.3 on p. 321 for the exact complexity). Two scalable alternatives are considered next.

MR Precoding

The first scalable option is *MR precoding*, which is obtained from (6.23) using

$$\bar{\mathbf{w}}_{kl}^{\mathrm{MR}} = \mathbf{D}_{kl}\widehat{\mathbf{h}}_{kl}. \tag{6.26}$$

This scheme maximizes the fraction of the transmitted power from AP l that is received at the desired UE (i.e., the numerator of the effective SINR). However, MR precoding ignores the interference that the AP is causing, particularly among the UEs that it is serving. The early works on Cell-free Massive MIMO considered MR precoding with $N = 1$ [**Nayebi2017a**], [**Ngo2017b**], in which case each AP has too few antennas to suppress interference. One key benefit of this scheme is that the effective SINR in Corollary 6.3 can be computed in closed form.

Corollary 6.4. Suppose MR precoding, as defined in (6.26), is used. The expectations in (6.22) then become

$$\sum_{l=1}^{L} \mathbb{E}\left\{\mathbf{h}_{kl}^{\mathrm{H}}\mathbf{D}_{kl}\mathbf{w}_{kl}\right\} = \sum_{l=1}^{L} \sqrt{\rho_{kl}\eta_k\tau_p \mathrm{tr}(\mathbf{D}_{kl}\mathbf{R}_{kl}\boldsymbol{\Psi}_{t_kl}^{-1}\mathbf{R}_{kl})} \tag{6.27}$$

$$\mathbb{E}\left\{\left|\sum_{l=1}^{L}\mathbf{h}_{kl}^{\mathrm{H}}\mathbf{D}_{il}\mathbf{w}_{il}\right|^2\right\} = \sum_{l=1}^{L}\rho_{il}\frac{\mathrm{tr}\left(\mathbf{D}_{il}\mathbf{R}_{il}\boldsymbol{\Psi}_{t_il}^{-1}\mathbf{R}_{il}\mathbf{R}_{kl}\right)}{\mathrm{tr}(\mathbf{R}_{il}\boldsymbol{\Psi}_{t_il}^{-1}\mathbf{R}_{il})}$$
$$+\begin{cases}\left|\sum\limits_{l=1}^{L}\sqrt{\rho_{il}\eta_k\tau_p}\dfrac{\mathrm{tr}\left(\mathbf{D}_{il}\mathbf{R}_{il}\boldsymbol{\Psi}_{t_il}^{-1}\mathbf{R}_{kl}\right)}{\sqrt{\mathrm{tr}(\mathbf{R}_{il}\boldsymbol{\Psi}_{t_il}^{-1}\mathbf{R}_{il})}}\right|^2 & i\in\mathcal{P}_k\\ 0 & i\notin\mathcal{P}_k.\end{cases} \tag{6.28}$$

Proof. By inserting the MR precoding expression into (6.22), we can observe that the expectations that appear are the same as in Corollary 5.6 on p. 318 (except that some indices must be interchanged). □

The expressions in Corollary 6.4 can be inserted into the effective SINR in (6.22) to obtain a closed-form SE expression. However, the final expression is lengthy. Hence, we will focus on discussing its properties instead of presenting it in a single equation.

The signal term in the numerator of the effective SINR is the square of (6.27). We can recognize $\eta_k\tau_p\mathbf{R}_{kl}\boldsymbol{\Psi}_{t_kl}^{-1}\mathbf{R}_{kl}$ from Corollary 4.1 on p. 267 as the correlation matrix of the MMSE channel estimate $\widehat{\mathbf{h}}_{kl}$. Hence, the signal term can be equivalently expressed as

$$\left(\sum_{l=1}^{L}\sqrt{\rho_{kl}\mathbb{E}\{\|\mathbf{D}_{kl}\widehat{\mathbf{h}}_{kl}\|^2\}}\right)^2 \tag{6.29}$$

and grows when the channel estimation quality improves. To make the signal term large, it is important to maximize the estimation quality, which can be achieved by limiting the pilot contamination effect in the ways discussed in Section 4.3 on p. 274. The contributions from the serving APs are coherently combined, which is represented by the fact that the square roots of their contributions are added up followed by taking the square of the sum. This operation leads to a power gain as compared to the case when the contributions are added up directly in the power domain. As a simple example of this, suppose a and b are the power of the contributions from two different APs. We can then notice that

$$(\sqrt{a}+\sqrt{b})^2 = a+b+2\sqrt{ab} > a+b \tag{6.30}$$

for any strictly positive values of a and b.

The interference term caused by the signal transmission to UE i is given in (6.28). This expression is more complicated to interpret than the signal term but we can recognize terms of the kind $\mathbf{R}_{il}\boldsymbol{\Psi}_{t_i l}^{-1}\mathbf{R}_{il}$ which are proportional to the correlation matrix of the channel estimate $\widehat{\mathbf{h}}_{il}$. This is natural since this channel estimate is used for MR precoding to UE i. The first interference term in (6.28) contains a product $\mathbf{R}_{il}\mathbf{R}_{kl}$ between the correlation matrices of the interfering UE and the desired UE. If these UEs have very different spatial correlation properties (e.g., if the dominant eigenvectors are orthogonal), then the interference will be smaller than if they have matching correlation matrices. There is also an additional interference term, which is only included if UE i uses the same pilot as UE k. This term is called coherent interference since it also contains a summation of square roots followed by a squaring. This pilot-contaminated term becomes equal to the signal term if $i = k$.

The effective SINR expression with MR precoding can be substantially simplified in the case of $N = 1$ antenna per AP.

Corollary 6.5. If each AP has $N = 1$ antenna, then the MMSE estimate of the scalar channel $\mathbf{h}_{kl} \in \mathbb{C}$ has variance

$$\gamma_{kl} = \frac{\eta_k \tau_p \beta_{kl}^2}{\sum_{i \in \mathcal{P}_k} \eta_i \tau_p \beta_{il} + \sigma_{\text{ul}}^2}. \tag{6.31}$$

The effective SINR in (6.22) with MR precoding then becomes

$$\mathsf{SINR}_k^{(\text{dl},\mathsf{d})} = \frac{\left(\sum_{l \in \mathcal{M}_k} \sqrt{\rho_{kl}\gamma_{kl}} \right)^2}{\sum_{i=1}^{K} \sum_{l \in \mathcal{M}_i} \rho_{il}\beta_{kl} + \sum_{i \in \mathcal{P}_k \setminus \{k\}} \left(\sum_{l \in \mathcal{M}_i} \sqrt{\rho_{il}\gamma_{kl}} \right)^2 + \sigma_{\text{dl}}^2}. \tag{6.32}$$

Proof. Using the notation in (6.31), the expectations in Corollary 6.4 can be rewritten as $\sum_{l=1}^{L} \mathbb{E}\left\{\mathbf{h}_{kl}^{\text{H}}\mathbf{D}_{kl}\mathbf{w}_{kl}\right\} = \sum_{l \in \mathcal{M}_k} \sqrt{\rho_{kl}\gamma_{kl}}$ and

$$\mathbb{E}\left\{ \left| \sum_{l=1}^{L} \mathbf{h}_{kl}^{\text{H}}\mathbf{D}_{il}\mathbf{w}_{il} \right|^2 \right\} = \sum_{l \in \mathcal{M}_i} \rho_{il}\beta_{kl} + \begin{cases} \left| \sum_{l \in \mathcal{M}_i} \sqrt{\rho_{il}\gamma_{kl}} \right|^2 & i \in \mathcal{P}_k \\ 0 & i \notin \mathcal{P}_k. \end{cases}$$

Inserting these values into (6.22) yields the SINR expression in (6.32), where absolute values of positive terms have been replaced by parenthesis. □

The closed-form effective SINR in (6.32) shows the key behaviors of the cell-free operation even clearer than for $N > 1$. The signal term $(\sum_{l \in \mathcal{M}_k} \sqrt{\rho_{kl}\gamma_{kl}})^2$ in the numerator contains a coherent combination of the signal contributions from all the serving APs, which leads to a power gain. The first term in the denominator contains non-coherent interference, which means that it is a summation of the powers of the individual signals transmitted to all the UEs from their respective serving APs. The second term is the additional coherent interference caused by pilot contamination. The third term is the noise power.

MR precoding is expected to work well if there is a high degree of favorable propagation, which is not guaranteed to be the case in cell-free networks (see Section 2.6.2 on p. 234). If each AP is equipped with multiple antennas, then we can use a local precoding scheme that is also capable of suppressing interference. The aforementioned unscalable L-MMSE precoding is an example of this but there are scalable alternatives that are suitable for cell-free networks.

LP-MMSE Precoding

We can achieve a scalable approximation of L-MMSE precoding by only considering the UEs that the AP serves, as was done for the uplink in (5.39). We call this *LP-MMSE precoding* and obtain it from (6.23) using

$$\bar{\mathbf{w}}_{kl}^{\mathrm{LP-MMSE}} = p_k \left(\sum_{i \in \mathcal{D}_l} p_i \left(\widehat{\mathbf{h}}_{il} \widehat{\mathbf{h}}_{il}^{\mathrm{H}} + \mathbf{C}_{il} \right) + \sigma_{\mathrm{ul}}^2 \mathbf{I}_N \right)^{-1} \mathbf{D}_{kl} \widehat{\mathbf{h}}_{kl}. \quad (6.33)$$

If AP l is serving all the UEs in its area of influence, LP-MMSE precoding will be approximately equal to L-MMSE precoding. However, the major benefit is that LP-MMSE was previously shown to be a scalable scheme. If it is used in both uplink and downlink, then no additional computations are required for creating the downlink precoding vectors.

Table 6.2: Number of complex scalars to be shared over the fronthaul per coherence block in the distributed downlink operation. These scalars are sent from the CPU to the APs.

Scheme	Each coherence block
Any precoding	$\tau_d \sum_{l=1}^{L} \lvert\mathcal{D}_l\rvert$

Remark 6.2 (Other precoding schemes)**.** The literature contains other precoding schemes than those considered in this monograph, in particular, different variations on MMSE and L-MMSE precoding where the regularization terms are selected differently or are removed, as in the case of ZF precoding [**Interdonato2020a**], [**Nguyen2017a**]. Moreover, different APs can use different schemes, to reduce the overall computational complexity or protect certain "sensitive" UEs from interference. Interestingly, in the special case when each AP has more antennas than there are pilots (i.e., $N > \tau_p$) and the channels are subject to uncorrelated Rayleigh fading, there are several local ZF schemes for which the SE can be computed in closed form [**Interdonato2020a**].

6.2.3 Fronthaul Signaling Load for Distributed Operation

The only downlink processing that is carried out at the CPU in the distributed operation is the encoding of the data signals. The CPU needs to send a portion of the data signals $\{\varsigma_k : k = 1, \ldots, K\}$ to each AP, corresponding to the UEs that the AP is serving. AP l needs to receive $\tau_d|\mathcal{D}_l|$ complex scalars per coherence block. This number is independent of the choice of precoding scheme. A key difference from the centralized operation is that the fronthaul signaling requirement of an AP is proportional to the number of UEs that it serves rather than the number of antennas. The total number of complex scalars is summarized in Table 6.2.

6.3 Numerical Performance Evaluation

We will now compare the downlink SE achieved by the centralized and distributed precoding schemes described earlier in this section. We will

continue the running example defined in Section 5.3 on p. 325. There are many power allocation parameters that need to be selected to run simulations. We will consider power allocation optimization in detail in Section 7.1.2 on p. 399, while only heuristic power allocation schemes are utilized in this section. We will elaborate more on heuristic power allocation in Section 7.2 on p. 410.

In the centralized operation, we use a specific version of the scalable centralized power allocation in Section 7.2 on p. 410 with the transmit power allocated to each UE as

$$\rho_k = \rho_{\max} \frac{\left(\sqrt{\sum_{l \in \mathcal{M}_k} \beta_{kl}} \right)^{-1} (\sqrt{\omega_k})^{-1}}{\max\limits_{\ell \in \mathcal{M}_k} \sum\limits_{i \in \mathcal{D}_\ell} \left(\sqrt{\sum_{l \in \mathcal{M}_i} \beta_{il}} \right)^{-1} \sqrt{\omega_i}} \tag{6.34}$$

with

$$\omega_k = \max_{\ell \in \mathcal{M}_k} \mathbb{E} \left\{ \left\| \bar{\mathbf{w}}'_{k\ell} \right\|^2 \right\} \tag{6.35}$$

where $\bar{\mathbf{w}}'_{k\ell} \in \mathbb{C}^N$ is the portion of the normalized centralized precoding vector $\frac{\bar{\mathbf{w}}_k}{\sqrt{\mathbb{E}\{\|\bar{\mathbf{w}}_k\|^2\}}}$ from (6.14) that corresponds to AP ℓ. The precoding scheme will determine how this power is distributed between the different APs, but it is typically the closest APs that contribute with the majority of the power. The normalization factor in the denominator is selected to make sure that none of the APs will transmit with more power than the maximum value $\rho_{\max}$. With this scheme, the transmit powers of the UEs are selected inversely proportional to the square-root of their total channel gains from their serving APs. The details of the general version of this scalable power allocation scheme and how each per-AP power constraint is satisfied are shown in Section 7.2 on p. 410.

In the distributed operation, a specific version of the distributed power allocation scheme that was proposed in [**Interdonato2019a**] will be used:

$$\rho_{kl} = \begin{cases} \rho_{\max} \frac{\sqrt{\beta_{kl}}}{\sum_{i \in \mathcal{D}_l} \sqrt{\beta_{il}}} & k \in \mathcal{D}_l \\ 0 & k \notin \mathcal{D}_l \end{cases} \tag{6.36}$$

where the per-AP transmit power constraints are satisfied by construction. With this scheme, each AP will assign more power to the UEs that it has good channels to, than to UEs with worse channels. The general version of this heuristic scheme is discussed in more detail in Section 7.2 on p. 410.

Unless otherwise stated, the main simulation parameters are as follows. There are $K = 40$ UEs and the pilot sequence length is $\tau_p = 10$. The remaining $\tau_d = \tau_c - \tau_p = 190$ transmission symbols of each coherence block are used for downlink data only. The Gaussian local scattering model is used to generate the spatial correlation matrices with ASD $\sigma_\varphi = \sigma_\theta = 15°$. We use the same Monte Carlo simulation methodology as described in Section 5.4 on p. 331 to generate the numerical results. In each figure, the legend will indicate the schemes that are being used. We use "(DCC)" to denote the DCC implementation with Algorithm 4.1 on p. 289 and "(All)" to denote the case where all APs serve all UEs.

6.3.1 Benchmark Schemes

The SE expressions provided in Theorem 6.1 and Corollary 6.3 are computed under the assumption that the receiving UE is applying a simple receiver structure that treats the average effective channel $\mathbb{E}\{\mathbf{h}_k^{\mathrm{H}}\mathbf{D}_k\mathbf{w}_k\}$ as the true channel. To evaluate the tightness of this capacity bound, we will compare it with a benchmark where the UE has perfect channel knowledge, which has been obtained in some genie-aided manner. If the gap between the expressions is small, then the previously provided SEs are good performance metrics. We note that the effective downlink SINRs in the centralized and distributed operations, given in (6.10) and (6.22), respectively, have the same structure. The difference is in the selection of the transmit precoding vectors $\{\mathbf{w}_{il} : i \in \mathcal{D}_l, l = 1, \ldots, L\}$. Hence, we will obtain a common genie-aided SE expression, which can be used in both cases.

If the channels $\{\mathbf{h}_{kl} : k \in \mathcal{D}_l, l = 1, \ldots, L\}$ are known at UE k, we

can rewrite the received signal at UE k in (6.1) as

$$y_k^{\mathrm{dl}} = \underbrace{\sum_{l=1}^{L} \mathbf{h}_{kl}^{\mathrm{H}} \mathbf{D}_{kl} \mathbf{w}_{kl} \varsigma_k}_{\text{Desired signal}} + \underbrace{\sum_{\substack{i=1 \\ i \neq k}}^{K} \sum_{l=1}^{L} \mathbf{h}_{kl}^{\mathrm{H}} \mathbf{D}_{il} \mathbf{w}_{il} \varsigma_i}_{\text{Interference}} + \underbrace{n_k}_{\text{Noise}} . \tag{6.37}$$

The first term contains the desired signal, while we treat the remaining terms as noise in line with the capacity bounds considered previously in this section. We then have the following genie-aided SE expression.

Corollary 6.6. A genie-aided SE of UE k is

$$\mathsf{SE}_k^{(\mathrm{gen-dl})} = \frac{\tau_d}{\tau_c} \mathbb{E} \left\{ \log_2 \left(1 + \mathsf{SINR}_k^{(\mathrm{gen-dl})} \right) \right\} \tag{6.38}$$

where

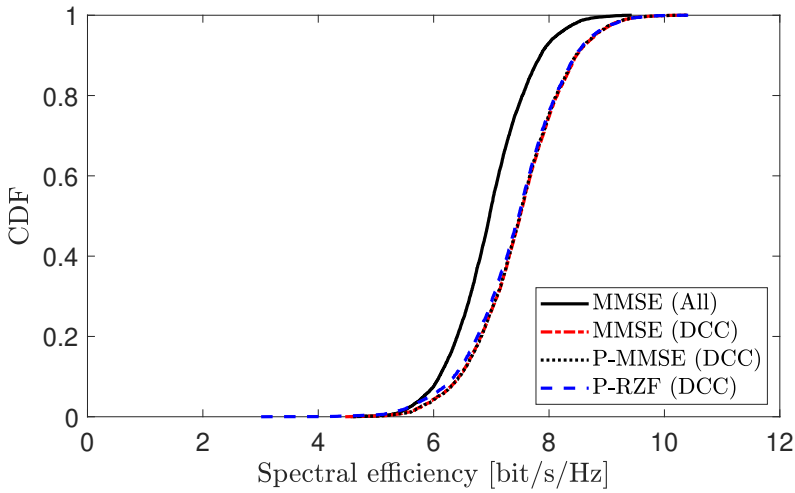
$$\mathsf{SINR}_k^{(\mathrm{gen-dl})} = \frac{\left| \sum_{l=1}^{L} \mathbf{h}_{kl}^{\mathrm{H}} \mathbf{D}_{kl} \mathbf{w}_{kl} \right|^2}{\sum_{\substack{i=1 \\ i \neq k}}^{K} \left| \sum_{l=1}^{L} \mathbf{h}_{kl}^{\mathrm{H}} \mathbf{D}_{il} \mathbf{w}_{il} \right|^2 + \sigma_{\mathrm{dl}}^2}. \tag{6.39}$$

Proof. This follows from utilizing Lemma 3.5 on p. 248. □

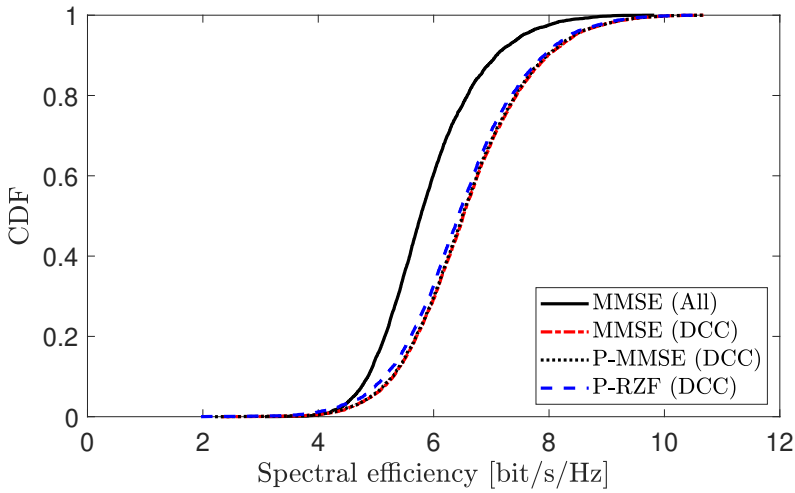
6.3.2 Performance With Centralized Operation

In Figure 6.3, we show the CDF of the downlink SE per UE in the centralized operation. The randomness that gives rise to the CDF is due to the AP and UE locations, as well as to the shadow fading realizations. The SE is computed using (6.9) in Theorem 6.1 for the different centralized precoding schemes described in Section 6.1: MMSE, P-MMSE, and P-RZF. We stress that these are three heuristic precoding schemes that are motivated, via the uplink-downlink duality in Theorem 6.2, by the good performance of their uplink counterparts. Although the strict duality is only valid for specific transmit power coefficients, these precoding schemes perform satisfactorily also for other power allocations. In this part, we use the heuristic power allocation scheme in (6.34).

Figure 6.3(a) considers the scenario with $L = 400$ APs and $N = 1$ antenna per AP and Figure 6.3(b) considers the scenario with $L = 100$ APs with $N = 4$ antennas per AP. For both scenarios, the key



(a) $L = 400$, $N = 1$.



(b) $L = 100$, $N = 4$.

Figure 6.3: CDF of the downlink SE per UE in the centralized operation. We consider $K = 40$, $\tau_p = 10$, and spatially correlated Rayleigh fading with ASD $\sigma_\varphi = \sigma_\theta = 15°$. Different precoding schemes are compared.

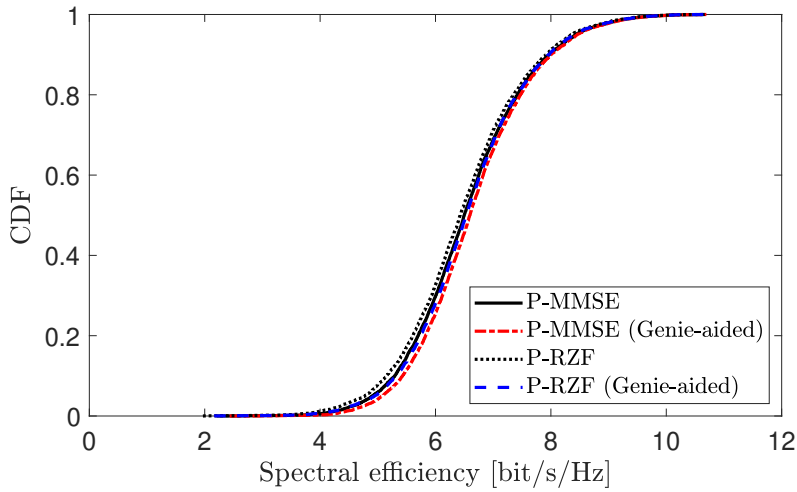


Figure 6.4: CDF of the downlink SE per UE in the centralized operation of the same scenario as in Figure 6.3(b). We consider $L = 100$, $N = 4$, $K = 40$, $\tau_p = 10$, and spatially correlated Rayleigh fading with ASD $\sigma_\varphi = \sigma_\theta = 15°$. The SE expression from (6.9) is compared with genie-aided SE in (6.38).

observation is that MMSE precoding using all the APs provides smaller SE than the schemes where the DCC framework is used to restrict how many UEs that each AP is serving. Transmitting from all the APs is not an efficient way when it comes to the downlink operation. Instead, selecting the best APs is a more efficient way of achieving a power allocation that suppresses interference. The difference between MMSE, P-MMSE, and P-RZF is almost negligible. Recall that MMSE precoding is not a scalable scheme but, fortunately, the fully scalable P-MMSE and P-RZF precoding schemes provide almost the same SE.

When we switch to the second scenario in Figure 6.3(b) with less densely deployed APs, we see a SE drop for all UEs, except in the upper tail. Hence, we conclude that the former scenario with many single-antenna APs is more desirable for the centralized downlink operation. This is in line with our previous observations in Section 5.4.1 on p. 333 for the centralized uplink operation.

Tightness of the SE Expression

The SE expression for centralized operation in (6.9) is derived under the assumption that the receiver treats the average effective channel as the true channel. This results in a valid lower bound on the capacity but there is a risk that an improved receiver will perform better, particularly, if there is a low degree of channel hardening. To quantify the tightness of the SE expression in (6.9), Figure 6.4 considers the same simulation setup as in Figure 6.3(b) but also includes the genie-aided SEs from (6.38), which are obtained when the UEs have perfect CSI. We consider the scalable P-MMSE and P-RZF precoding schemes. In both cases, the performance gap to the genie-aided curve is virtually non-existing. Hence, similar to the uplink, we conclude that a centralized operation leads to a high degree of channel hardening in the running example, so that it is sufficient for the receiver to only know the average effective channel when detecting data. However, one can create simulation setups where the gap is larger (cf. [**Chen2018b**]) and then the downlink channel estimation methods discussed in Remark 6.1 are needed to close the gap.

6.3.3 Performance With Distributed Operation

We will now consider the distributed operation. Figure 6.5 shows the CDF of the SE per UE in the same scenarios as in Figure 6.3 with the heuristic power allocation method stated in (6.36). Figure 6.5(a) assumes $L = 400$ and $N = 1$, while Figure 6.5(b) considers $L = 100$ and $N = 4$. In both scenarios, the L-MMSE precoding scheme that uses all the APs results in significantly lower SE compared to the DCC implementation with L-MMSE and its scalable version LP-MMSE. This unexpected result is caused by the assumed power allocation scheme and can be explained as follows. In general, most of the UEs have negligibly small channel gains to a particular AP and, hence, transmitting to all the UEs is like screaming and anyway only being heard by the closest UEs. The AP is essentially taking power that could have been used to serve the nearby UEs and assign it to faraway UEs, which mainly results in extra interference to the nearby UEs. In this case, the DCC essentially

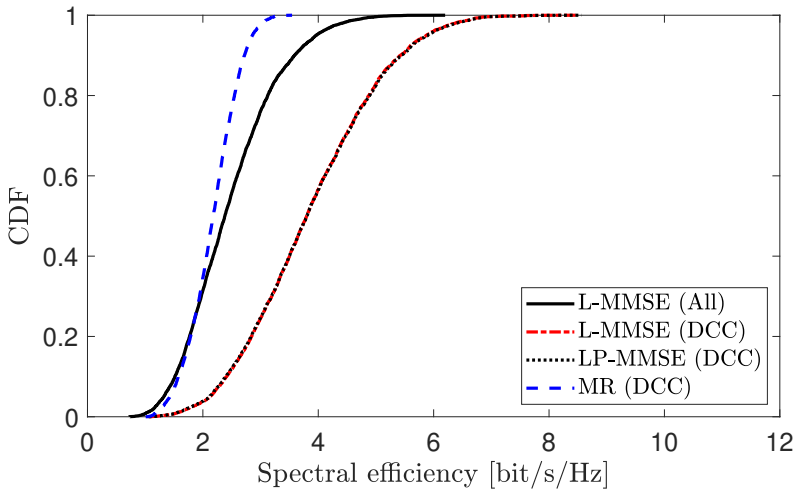
provides an improved heuristic power allocation scheme where the power is allocated only to the UEs that are assigned to an AP with good channel conditions [**Buzzi2017b**].

Similar to the uplink, both figures demonstrate that we can achieve roughly the same SE by using the scalable LP-MMSE precoder instead of the unscalable L-MMSE precoder. When comparing the two scenarios, we notice that the performance is greatly improved when considering fewer APs with multiple antennas. This is an expected result since both L-MMSE and LP-MMSE are taking the interference and estimation errors into account and they can suppress interference better with $N = 4$ antennas per AP. The improvements are largest in the upper tail of the CDF curves, where interference is the limiting factor, and this also means that the SE curves are more spread out with $N = 4$. These conclusions are, qualitatively speaking, in line with those made for the uplink in Section 5.4 on p. 331. Finally, we note that the aforementioned properties do not apply to MR, which performs almost equally bad in both scenarios.

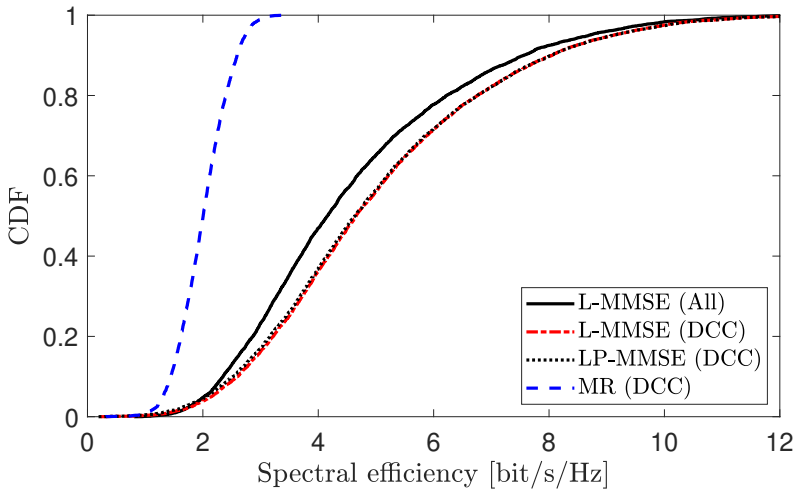
Tightness of the SE Expression

As in the centralized case, the SE expression used in the distributed operation assumes that the UEs have only access to the mean of the effective channels. To quantify the tightness of the provided SE results, compared to what could be achieved with the refined downlink channel estimation schemes described in Remark 6.1, we will compare with the genie-aided SE in (6.38), which assumes the same precoding schemes are used at the APs but the UEs are having access to perfect CSI when detecting the downlink signals.

Figure 6.6 shows the CDF of the SE with the scalable precoding schemes from Figure 6.5(b) with $L = 100$ and $N = 4$. The genie-aided results match rather well with the achievable SE values provided by (6.21) when using LP-MMSE precoding. The gap is slightly larger than in the centralized case (see Figure 6.4) but nevertheless small. Hence, for LP-MMSE precoding, we can conclude that the provided SE expression is both practically achievable and fairly close to what could be achieved with perfect CSI at the UEs.



(a) $L = 400$, $N = 1$.



(b) $L = 100$, $N = 4$.

Figure 6.5: CDF of the downlink SE per UE in the distributed operation. We consider $K = 40$, $\tau_p = 10$, and spatially correlated Rayleigh fading with ASD $\sigma_\varphi = \sigma_\theta = 15°$. Different precoding schemes are compared.

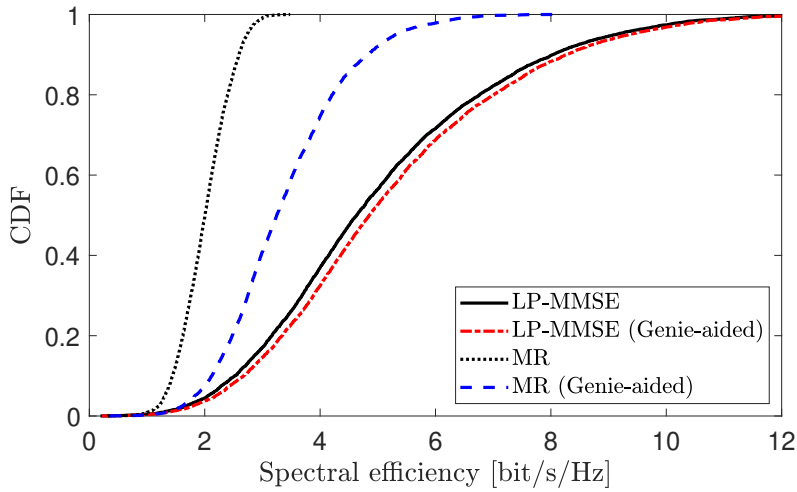
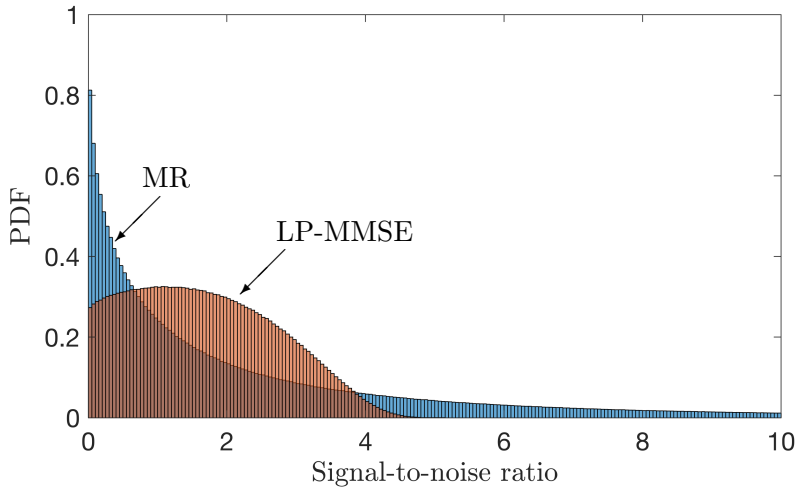
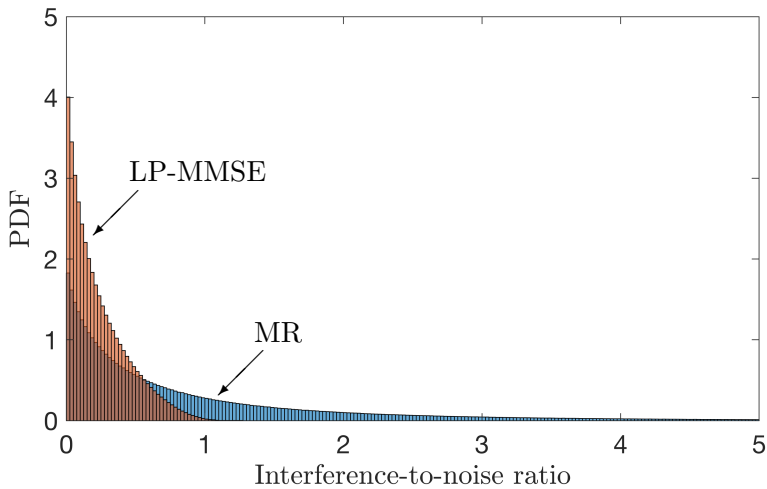


Figure 6.6: CDF of the downlink SE per UE in the distributed operation in the same scenario of Figure 6.5(b). We consider $L = 100$, $N = 4$, $K = 40$, $\tau_p = 10$, and spatially correlated Rayleigh fading with ASD $\sigma_\varphi = \sigma_\theta = 15°$. The SE expression from (6.21) is compared with genie-aided SE in (6.38).

There is a substantial gap between the achievable and genie-aided SE expressions when using MR precoding. To explain the intuition behind this result, Figure 6.7 shows the variations in signal and interference power in a toy example with $L = 1$ AP, $N = 2$ antennas, $K = 2$ UEs, perfect CSI, and uncorrelated Rayleigh fading channels. Figure 6.7(a) shows the PDF of the signal power (normalized by the noise power) when considering different channel realizations and either MR or LP-MMSE precoding. MR provides almost twice the average signal power as compared with LP-MMSE, but even when accounting for that, it is clear that LP-MMSE has bounded support while MR has a distribution with a long tail. Similar behaviors can be observed in Figure 6.7(b), where the interference power caused to the other UE is considered. LP-MMSE gives substantially smaller values, due to its interference suppression, and also a distribution with bounded support, while MR gives rise to large variations. It is the large variations in the signal and interference power that cause the gap between the achievable and genie-aided SE expressions, which implies that more refined downlink



(a) Variations in the received signal power.



(b) Variations in the received interference power.

Figure 6.7: Distribution of the signal and interference powers when using MR or LP-MMSE precoding in a simple setup with $L = 1$ AP, $N = 2$ antennas, $K = 2$ UEs, perfect CSI, and uncorrelated Rayleigh fading channels. The distributions are widely different: LP-MMSE gives rise to bounded support while MR provides large variations, which makes it harder to find tight capacity bounds.

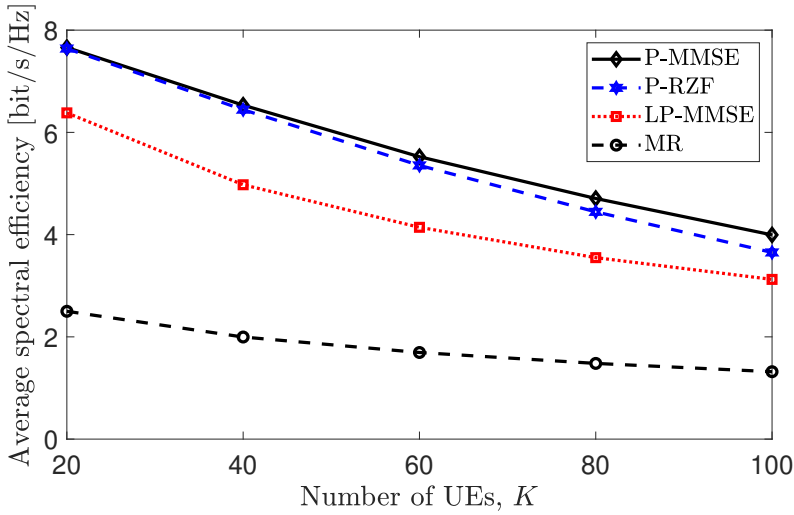
estimation schemes or capacity bounds are needed when using MR; see [**Caire2017a**], [**Chen2018b**], [**Interdonato2019b**]. Note that these results are in line with the uplink counterpart from Section 5.4 on p. 331.

6.3.4 Impact of the Number of UEs

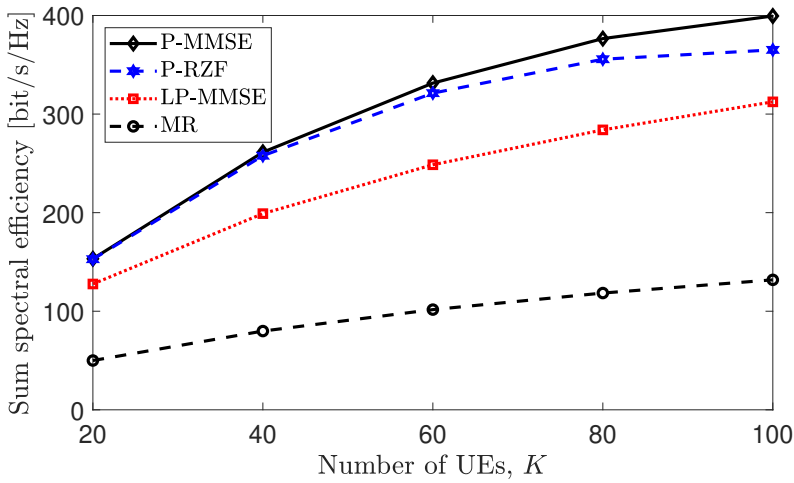
We will now analyze the impact of the number of UEs on network performance. The setup with $L = 100$ APs with $N = 4$ antennas per AP is considered. We plot the average SE per UE in Figure 6.8(a) whereas the average sum SE is reported in Figure 6.8(b). Centralized operation with P-MMSE or P-RZF precoding is compared with distributed operation with LP-MMSE or MR precoding. The centralized schemes outperform the distributed ones, in accordance with the previous results. P-MMSE provides a slightly larger SE than P-RZF, as also observed in Figure 6.3. In addition, LP-MMSE provides substantially higher SE than MR, since it can suppress interference. Note that the pilot length is $\tau_p = 10$ irrespective of the number of UEs. Hence, the pilot contamination increases with K, but despite the extra interference, we observe no more than a two-fold reduction in SE per UE when we increase the number of UEs from $K = 20$ and $K = 100$ in Figure 6.8(a). The benefit of multiplexing more UEs is that the sum SE increases almost linearly with K in Figure 6.8(b). This shows how a cell-free network is capable of serving a huge number of UEs.

6.3.5 Impact of Spatial Correlation

We conclude this section by analyzing the impact that spatial channel correlation has on the setup with $L = 100$, $N = 4$, and $K = 40$. The ASD is the same in the azimuth and elevation domains: $\sigma_\varphi = \sigma_\theta$. Figure 6.9 shows the average SE per UE as a function of the ASD, for different precoding schemes. For each scheme, the corresponding straight dotted line shows the performance achieved when having uncorrelated Rayleigh fading (i.e., no spatial correlation). The lines corresponding to the uncorrelated Rayleigh fading follow the same order as the other lines for each scheme. We notice that spatial correlation degrades the



(a) The average SE per UE.



(b) The average sum SE.

Figure 6.8: The average downlink SE per UE and the sum SE as a function of the number of UEs K for different operations of Cell-free Massive MIMO. We consider $L = 100$, $N = 4$, $\tau_p = 10$, and spatially correlated Rayleigh fading with ASD $\sigma_\varphi = \sigma_\theta = 15^\circ$.

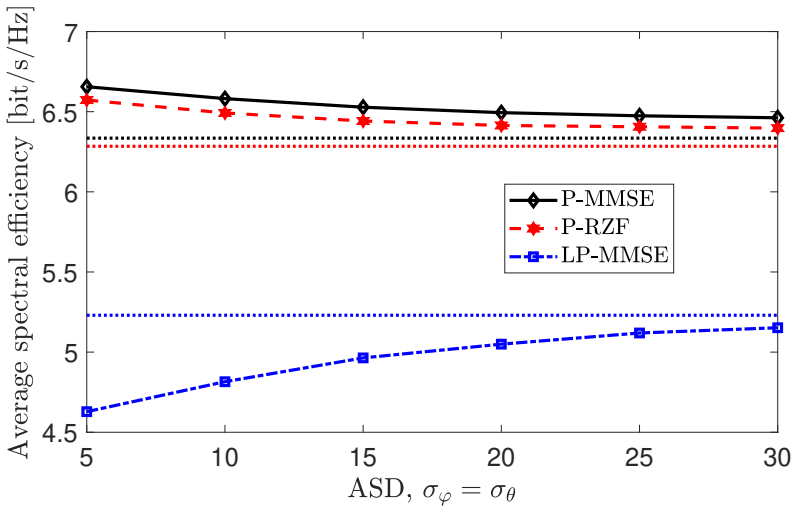


Figure 6.9: The average downlink SE per UE as a function of ASD for azimuth and elevation angles, $\sigma_\varphi = \sigma_\theta$ for different operations of Cell-free Massive MIMO. We consider $L = 100$, $N = 4$, $K = 40$, and $\tau_p = 10$. The results for uncorrelated Rayleigh fading are shown as reference by the dotted lines.

average SE in the distributed operation while it is advantageous in the centralized case, since a smaller ASD corresponds to a more highly correlated channel. This is consistent with previous observations in the uplink; see Figure 5.10 on p. 345. When the ASD increases, the average SE approaches the reference case with no spatial correlation.

Recall that there are several tradeoffs in connection with spatial correlation. In Section 2.6 on p. 228, we showed that the degree of channel hardening decreases with an increasing spatial correlation, which is a negative impact of correlation. There are also some benefits from correlation. The level of favorable propagation between two UEs is maximized when the UEs' channels are highly spatially correlated but have very different dominant eigenspaces. Furthermore, the channel estimation quality is improved with increased spatial correlation, as analyzed in Section 4.3.3 on p. 281. When putting these properties together, it is clear that the negative effects of spatial correlation dominate over the positive effects when using a distributed operation where each AP selects a low-dimensional precoding vector that cannot exploit the spatial correlation to a sufficient extent.

Finally, we note that the gap between P-MMSE and P-RZF is slightly larger under spatial correlation than with uncorrelated fading. Recall that the difference between these schemes is that P-MMSE is utilizing the estimation error correlation matrices $\mathbf{C}_k$, which affects the precoding direction in the correlated fading case. However, under uncorrelated fading, they are scaled identity matrices that can be lumped together with the noise term, thereby reducing their impact.

We compared the fronthaul signaling load with the centralized and distributed uplink operations in Fig. 5.12(a) on p. 348. We notice that the fronthaul signaling loads for data and pilots are the same in the downlink as in the uplink when $\tau_d = \tau_u$. This is seen by comparing Tables 6.1 and 6.2 with Table 5.2 on p. 311. Hence, we will not provide any additional numerical results related to this.

6.4 Summary of the Key Points in Section 6

- The downlink of a cell-free network can be implemented with either centralized or distributed operation.
- In the centralized operation, the CPU is selecting the precoding vectors and computes the signals to be transmitted, while the APs are only taking care of the physical transmission. This operation enables centralized precoding where the signals transmitted from multiple APs can be coherently received at desired UEs while also suppressing each other at the undesired UEs.
- The downlink SE with the centralized operation is given in Theorem 6.1. The optimal precoding is computationally intractable to obtain, but the uplink-downlink duality in Theorem 6.2 establishes that a scaled version of the uplink combining vector is a good heuristic for the corresponding downlink precoding vector. The MMSE, P-MMSE, and P-RZF precoding schemes are defined in this way, whereof the latter two admit scalable implementations.
- In the distributed operation, each AP is receiving the downlink data from the CPU and designs the locally transmitted signal using local precoding based on the locally available channel estimates. The signals transmitted from the serving APs are coherently received at the desired UE, but the interference suppression capability is reduced compared to the centralized operation since each AP can only suppress the interference that itself is generating. This is a limiting factor since each AP is envisioned to have few antennas.

- The downlink SE with the distributed operation is given in Corollary 6.3. The uplink-downlink duality motivates using the local combining vectors from the uplink as precoding vectors in the downlink. L-MMSE, LP-MMSE, and MR precoding are defined in this way, whereof the latter two admit scalable implementations.
- We compared the performance of different centralized and distributed implementations of cell-free networks using the running example. In line with the uplink, scalable precoding schemes in both centralized and distributed operations can achieve almost the same SE as their unscalable counterparts, but with much lower complexity.
- The precoding schemes are motivated by the uplink-downlink duality, thus there is no guarantee that the optimal combining scheme (from the virtual uplink system) provides the highest performance when used for precoding in the downlink.
- In the centralized operation, it is preferable to have many single-antenna APs when there is a fixed total number of antennas, just as in the uplink. In the distributed operation, UEs with good channel conditions benefit from having a smaller number of multi-antenna APs, which can use local precoding schemes such as LP-MMSE to suppress interference.
- Increasing the number of UEs in the network improves the sum SE in both the centralized and distributed operations, although the average SE per UE decreases due to the extra interference. Scalable cell-free networks have the capability of serving a large number of UEs with a fixed complexity.

7

Spatial Resource Allocation

This section describes some important considerations for the optimization, operation, and implementation of User-centric Cell-free Massive MIMO in practical networks. While previous sections have presented the foundations, this section is more related to the latest trends and developments. New algorithms and insights are likely to arise in the future, beyond what is presented here. The general theme is spatial resource allocation, which refers to the allocation of transmit powers and pilot signals to the UEs, the design of cooperation clusters, and the provisioning of fronthaul resources among the spatially distributed UEs and APs. Section 7.1 describes power optimization algorithms for maximizing the sum SE and max-min fairness utility functions in both uplink and downlink. Since these algorithms solve network-wide optimization problems, they are not scalable but serve as theoretical benchmarks. Next, Section 7.2 presents heuristic power allocation alternatives, which are designed to be scalable and efficient. The optimized and scalable power allocation methods are compared in Section 7.3. In Section 7.4 and Section 7.5, we provide brief overviews of algorithms for pilot assignment and DCC selection, respectively. Section 7.6 discusses important implementation constraints that appear in practical deployments, including having limited fronthaul capacity. Finally, a summary of the key points is provided in Section 7.7.

7.1 Transmit Power Optimization

The transmit power coefficients have so far been treated as heuristically selected constants, but these can be optimized to maximize a network-wide utility function. Section 3.4 on p. 251 provided the theoretical foundations for maximizing the max-min fairness and sum SE utilities. These represent two structured ways of selecting an operating point at the outer boundary of the SE region, as was illustrated in Figure 3.2 on p. 254. In this section, we will consider these utilities in the cell-free context and maximize them with respect to the uplink and downlink power coefficients under their respective power constraints. We will mainly use fixed-point and weighted MMSE-based algorithms that take the channel statistics and choice of combining/precoding scheme as inputs. These algorithms can be implemented at the CPU and the same solution can be applied as long as the channel statistics remain the same. Hence, there is no need to adapt the transmit power on a coherence block basis. We will consider the uplink and downlink separately.

7.1.1 Uplink Power Optimization

We begin by considering the uplink power optimization. UE k transmits its uplink data with a power p_k, which can be selected as any number between 0 and the maximum power $p_{\max}$. The procedure of selecting an appropriate uplink power is often called *power control* since it can be viewed as controlling how much different UEs cut down on their powers, compared to $p_{\max}$, to maximize a particular utility function.

Several different SE expressions for the centralized and distributed operation were derived in Section 5 on p. 294. In this section, we consider the SE for the centralized case in Theorem 5.2 on p. 300 and the SE for the distributed case in Theorem 5.4 on p. 314. Both results were derived using the UatF bounding technique [**massivemimobook**] and, therefore, share a common structure that enables us to optimize the transmit powers using the same algorithms.[1]

[1] We have previously used the SE expression in Theorem 5.1 on p. 298 for analyzing the uplink performance in the centralized operation. This expression gives slightly larger SE values that better represent the practically achievable communication performance, but since the transmit powers only appear inside of expectations, the

We can gather all the uplink powers in a vector $\mathbf{p} = [p_1 \ \ldots \ p_K]^{\mathrm{T}}$ and notice that they affect all the UEs. The uplink SE of UE k depends on $\mathbf{p}$ via its effective SINR. The numerator of the SINR depends on the power p_k of its desired signal and the interference term in the denominator depends on all the power coefficients in $\mathbf{p}$. The effective SINR for UE k in the centralized and distributed uplink operation can be jointly expressed in the generic form

$$\mathrm{SINR}_k(\mathbf{p}) = \frac{b_k p_k}{\mathbf{c}_k^{\mathrm{T}} \mathbf{p} + \sigma_k^2} \tag{7.1}$$

as a function of the vector $\mathbf{p}$ with all transmit power coefficients. The difference between the two types of operation lies in the values of the parameters $b_k \geq 0$, $\mathbf{c}_k = [c_{k1} \ \ldots \ c_{kK}]^{\mathrm{T}} \in \mathbb{R}_{\geq 0}^K$, and $\sigma_k^2 \geq 0$. They represent the average channel gain of the desired signal, the vector with the average channel gains for the respective interfering signals, and the effective noise variance, which are given as

$$b_k = \begin{cases} \left|\mathbb{E}\left\{\mathbf{v}_k^{\mathrm{H}} \mathbf{D}_k \mathbf{h}_k\right\}\right|^2 & \text{for centralized operation} \\ \left|\mathbf{a}_k^{\mathrm{H}} \mathbb{E}\left\{\mathbf{g}_{kk}\right\}\right|^2 & \text{for distributed operation,} \end{cases} \quad \forall k \tag{7.2}$$

$$c_{kk} = \begin{cases} \mathbb{E}\left\{\left|\mathbf{v}_k^{\mathrm{H}} \mathbf{D}_k \mathbf{h}_k\right|^2\right\} - b_k & \text{for centralized operation} \\ \mathbb{E}\left\{\left|\mathbf{a}_k^{\mathrm{H}} \mathbf{g}_{kk}\right|^2\right\} - b_k & \text{for distributed operation,} \end{cases} \quad \forall k \tag{7.3}$$

$$c_{ki} = \begin{cases} \mathbb{E}\left\{\left|\mathbf{v}_k^{\mathrm{H}} \mathbf{D}_k \mathbf{h}_i\right|^2\right\} & \text{for centralized operation} \\ \mathbb{E}\left\{\left|\mathbf{a}_k^{\mathrm{H}} \mathbf{g}_{ki}\right|^2\right\} & \text{for distributed operation,} \end{cases} \quad \forall k, \forall i \neq k \tag{7.4}$$

$$\sigma_k^2 = \begin{cases} \sigma_{\mathrm{ul}}^2 \mathbb{E}\left\{\left\|\mathbf{D}_k \mathbf{v}_k\right\|^2\right\} & \text{for centralized operation} \\ \mathbf{a}_k^{\mathrm{H}} \mathbf{F}_k \mathbf{a}_k & \text{for distributed operation,} \end{cases} \quad \forall k. \tag{7.5}$$

The generic SINR expression in (7.1) has a structure that we recognize from Lemma 3.8 on p. 252 with $\left|\mathrm{DS}_k(\mathbf{p})\right|^2 = b_k p_k$ being the signal term in the numerator, $\sum_{i=1}^K \mathbb{E}\left\{\left|\mathrm{I}_{ki}(\mathbf{p})\right|\right\}^2 = \mathbf{c}_k^{\mathrm{T}} \mathbf{p}$ being the total interference power, and σ_k^2 is the noise power. Hence, we can use the optimization

expression is not amendable for power optimization. The power allocation algorithms that we develop in this section are optimally designed for the more conservative capacity bounds in Theorem 5.2 on p. 300 and in Theorem 5.4 on p. 314, but can also be used along with the bound in Theorem 5.1.

algorithms presented in Section 3.4 on p. 251, which apply for arbitrary instances of the SINR expression from Lemma 3.8, to solve two power optimization problems: max-min SE fairness and sum SE maximization. We stress that the algorithms presented below are applying to centralized operation with arbitrary receive combining and distributed operation with arbitrary local receive combining and LSFD weights. This does not mean that the same transmit powers are optimal in all situations. Since different network operations lead to different values of b_k, $\mathbf{c}_k$, and σ_k^2, the optimal transmit powers will naturally be different even if they are obtained using the same algorithms.

Uplink Max-Min SE Fairness

We first consider the max-min SE fairness problem, which was formulated for the general case in (3.33). In the considered uplink scenario, there are K individual transmit power constraints and the max-min fairness optimization problem can be particularized as

$$\begin{array}{lll} \underset{\mathbf{p} \geq \mathbf{0}_K}{\text{maximize}} & \underset{k \in \{1,\ldots,K\}}{\min} & \dfrac{b_k p_k}{\mathbf{c}_k^{\mathrm{T}} \mathbf{p} + \sigma_k^2} \\ \text{subject to} & p_k \leq p_{\max}, & k = 1, \ldots, K. \end{array} \tag{7.6}$$

This problem has the same structure as the generic problem in (3.33) with $R = K$ and $\mathbf{a}_r$ having a one in the rth entry and zeros elsewhere for $r = 1, \ldots, K$. Under the following conditions, the solution to (7.6) can be computed by using the fixed-point algorithm in Algorithm 3.2 on p. 259.

Lemma 7.1. The three conditions in Lemma 3.9 on p. 258 are satisfied for $\mathrm{SINR}_k(\mathbf{p})$ in (7.1) if the coefficients b_k, c_{ki} for all $i \neq k$, and σ_k^2 are strictly positive. Hence, we can use Algorithm 3.2 to solve (7.6).

The requirements stated in the lemma are mild and basically say that the processed signal for every UE includes a non-zero fraction of the transmitted power from all other UEs and that the noise power is non-zero. By particularizing Algorithm 3.2 for the problem at hand, we obtain Algorithm 7.1, which converges to the optimal solution of the max-min fairness problem in (7.6). All UEs will have equal SE at

the optimum and at least one UE will use the maximum power $p_{\max}$. Algorithm 7.1 converges quickly and has relatively low computational complexity since it only involves iterative closed-form updates of the variables. However, its complexity grows with K, thus it is not scalable.

Algorithm 7.1 Fixed-point algorithm for solving the uplink max-min fairness problem in (7.6).

1: **Initialization:** Set arbitrary initial power $\mathbf{p} > \mathbf{0}_K$ and the solution accuracy $\epsilon > 0$
2: **while** $\max\limits_{k\in\{1,\ldots,K\}} \mathrm{SINR}_k(\mathbf{p}) - \min\limits_{k\in\{1,\ldots,K\}} \mathrm{SINR}_k(\mathbf{p}) > \epsilon$ **do**
3: $\quad p_k \leftarrow \frac{p_k}{\mathrm{SINR}_k(\mathbf{p})}$, $k = 1, \ldots, K$
4: $\quad \mathbf{p} \leftarrow \frac{p_{\max}}{\max\limits_{k\in\{1,\ldots,K\}} p_k}\mathbf{p}$
5: **end while**
6: **Output:** Optimal transmit powers $\mathbf{p}$
7: Max-min SE $\min\limits_{k\in\{1,\ldots,K\}} \frac{\tau_u}{\tau_c} \log_2(1 + \mathrm{SINR}_k(\mathbf{p}))$

Uplink Sum SE Maximization

We will now consider the sum SE maximization problem, which is formulated in the uplink as

$$\begin{aligned} \underset{\mathbf{p}\geq\mathbf{0}_K}{\text{maximize}} \quad & \sum_{k=1}^{K} \log_2\left(1 + \frac{b_k p_k}{\mathbf{c}_k^{\mathrm{T}}\mathbf{p} + \sigma_k^2}\right) \\ \text{subject to} \quad & p_k \leq p_{\max}, \quad k = 1, \ldots, K. \end{aligned} \tag{7.7}$$

This problem has the same structure as the generic one in (3.38). Hence, we can conclude that (7.7) is not convex, however, a local optimum can be attained by the block coordinate descent algorithm given in Algorithm 3.3 on p. 261. That algorithm was developed based on a weighted MMSE reformulation of the sum SE maximization problem, where the MSE in the data detection was stated in (3.39). For the particular SINR expression considered in this section, the corresponding MSE can be particularized as

$$e_k(\mathbf{p}, u_k) = u_k^2\left(b_k p_k + \mathbf{c}_k^{\mathrm{T}}\mathbf{p} + \sigma_k^2\right) - 2u_k\sqrt{b_k p_k} + 1 \tag{7.8}$$

where we have used the fact that b_k is real-valued and positive and u_k is also real-valued since all the coefficients in the SINR expression are real-valued. The solution to the optimization problem in Step 6 of Algorithm 3.3 on p. 261 can be obtained in closed-form as

$$p_k = \min \left(p_{\max}, \frac{b_k d_k^2 u_k^2}{\left(d_k u_k^2 b_k + \sum_{i=1}^{K} d_i u_i^2 c_{ik} \right)^2} \right) \tag{7.9}$$

by treating $\sqrt{p_1}, \ldots, \sqrt{p_K}$ as the optimization variables and decomposing the problem into K independent subproblems each of which is a quadratic minimization under a bound constraint. The steps of the block coordinate descent algorithm for the uplink sum SE maximization can then be simplified to Algorithm 7.2, which is an iterative algorithm where all the updates are based on closed-form expressions. The algorithm converges quickly and has relatively low computational complexity, but the complexity grows with K so the algorithm is not scalable. Note that Algorithm 7.2 only guarantees to find a local optimum to (7.7), thus different initializations may lead to different solutions.

A typical property of sum SE maximization problems is that the optimal solution might assign zero power to some UEs, which naturally leads to zero SE. This can occur for UEs that have weak channels to all APs, however, we will show in Section 7.3 that this is not an issue that appears when applying the algorithm to the running example. We can be sure that at least one UE will use the maximum power $p_{\max}$ at the optimum solution.

Remark 7.1 (Pilot power optimization)**.** The optimization problems in (7.6) and (7.7) optimize the transmit powers $p_1, \ldots, p_K$ that are used for uplink data transmission. The uplink transmission also involves the pilot powers $\eta_1, \ldots, \eta_K$, which were implicitly assumed to be constant (they appear at the inside of b_k, $\mathbf{c}_k$, and σ_k^2). If the pilot assignment is carried out properly, then pilot transmission with maximum power $\eta_1 = \ldots = \eta_K = p_{\max}$ is expected to be a nearly optimal solution. There are a few papers that also consider pilot power optimization. For example, [**Mai2018**] optimizes the pilot powers by minimizing the

Algorithm 7.2 Block coordinate descent algorithm for solving the sum SE maximization problem in (7.7).

1: **Initialization:** Set the solution accuracy $\epsilon > 0$
2: Set an arbitrary feasible power vector $\mathbf{p}$
3: **while** $\sum_{k=1}^{K} \left(d_k e_k \left(\mathbf{p}, u_k \right) - \ln \left(d_k \right) \right)$ is either improved more than ϵ or not improved at all **do**
4: $\quad u_k \leftarrow \frac{\sqrt{b_k p_k}}{b_k p_k + \mathbf{c}_k^{\mathrm{T}} \mathbf{p} + \sigma_k^2}, \quad k = 1, \ldots, K$
5: $\quad d_k \leftarrow \frac{1}{u_k^2 \left(b_k p_k + \mathbf{c}_k^{\mathrm{T}} \mathbf{p} + \sigma_k^2 \right) - 2u_k \sqrt{b_k p_k} + 1}, \quad k = 1, \ldots, K$
6: $\quad p_k \leftarrow \min \left(p_{\max}, \frac{b_k d_k^2 u_k^2}{\left(d_k u_k^2 b_k + \sum_{i=1}^{K} d_i u_i^2 c_{ik} \right)^2} \right), \quad k = 1, \ldots, K$
7: **end while**
8: **Output:** Optimal transmit powers $\mathbf{p}$
9: Sum SE $\frac{\tau_u}{\tau_c} \sum_{k=1}^{K} \log_2 \left(d_k \right)$

maximum NMSE in the channel estimation, but without considering the performance in the data transmission phase. The joint pilot and data power optimization is considered in [**Masoumi2018**] for single-antenna APs, but only an approximation of the max-min fairness problem is solved. Further research on the joint pilot and data power optimization problem is required.

7.1.2 Downlink Power Optimization

We will now consider the downlink power optimization. AP l transmits the signal $\mathbf{x}_l = \sum_{i=1}^{K} \mathbf{D}_{il} \mathbf{w}_{il} \varsigma_i$ defined in (6.2) and is assumed to have a maximum transmit power of $\rho_{\max}$. This power can be arbitrarily divided between the UEs and this procedure is often called *power allocation*. An AP can decide to not use all of its power to not cause unnecessary interference, similar to the situation in the uplink. The centralized and distributed operation will be handled differently in this section, even if the power constraints are the same in both cases. The reason is that the precoding vectors are coupled between the APs in the centralized

case, while they are not in the distributed case.

In the centralized operation, the centralized precoding vectors

$$\mathbf{w}_k = \sqrt{\rho_k} \frac{\bar{\mathbf{w}}_k}{\sqrt{\mathbb{E}\left\{\|\bar{\mathbf{w}}_k\|^2\right\}}} \tag{7.10}$$

from (6.14) are used for $k = 1, \ldots, K$. There are K power allocation coefficients $\{\rho_k : k = 1, \ldots, K\}$ to optimize, where $\rho_k \geq 0$ represents the total downlink power allocated to UE k from all the serving APs.

In the distributed operation, the local precoding vectors for UE k are defined in (6.23) as

$$\mathbf{w}_{kl} = \sqrt{\rho_{kl}} \frac{\bar{\mathbf{w}}_{kl}}{\sqrt{\mathbb{E}\left\{\|\bar{\mathbf{w}}_{kl}\|^2\right\}}} \tag{7.11}$$

for $l \in \mathcal{M}_k$. There are $\sum_{k=1}^{K} |\mathcal{M}_k|$ power allocation coefficients $\{\rho_{kl} : l \in \mathcal{M}_k, k = 1, \ldots, K\}$ to optimize, where $\rho_{kl} \geq 0$ denotes the downlink power that AP l is assigning to UE k. Since there are more coefficients in the distributed case, we can expect the optimization to be more complex.

Optimized Centralized Downlink Power Allocation

We begin with the centralized operation for which the downlink SE is given by Theorem 6.1 on p. 361. We notice that the effective SINR, $\mathsf{SINR}_k^{(\mathrm{dl,c})}$, achieved by UE k can be expressed as a function of the downlink power coefficients $\boldsymbol{\rho} = [\rho_1 \ \ldots \ \rho_K]^{\mathrm{T}}$ as

$$\mathsf{SINR}_k^{(\mathrm{dl,c})}(\boldsymbol{\rho}) = \frac{\tilde{b}_k \rho_k}{\tilde{\mathbf{c}}_k^{\mathrm{T}} \boldsymbol{\rho} + \sigma_{\mathrm{dl}}^2} \tag{7.12}$$

where $\tilde{b}_k$ is the average channel gain of the desired signal and $\tilde{\mathbf{c}}_k = [\tilde{c}_{k1} \ \ldots \ \tilde{c}_{kK}]^{\mathrm{T}} \in \mathbb{R}_{\geq 0}^{K}$ is a vector containing the average channel gains for the respective interfering signals. The specific values of these parameters

in the centralized downlink operation are

$$\tilde{b}_k = \frac{\left|\mathbb{E}\left\{\mathbf{h}_k^{\mathrm{H}}\mathbf{D}_k\bar{\mathbf{w}}_k\right\}\right|^2}{\mathbb{E}\left\{\|\bar{\mathbf{w}}_k\|^2\right\}}, \quad \forall k \tag{7.13}$$

$$\tilde{c}_{kk} = \frac{\mathbb{E}\left\{\left|\mathbf{h}_k^{\mathrm{H}}\mathbf{D}_k\bar{\mathbf{w}}_k\right|^2\right\}}{\mathbb{E}\left\{\|\bar{\mathbf{w}}_k\|^2\right\}} - \tilde{b}_k, \quad \forall k \tag{7.14}$$

$$\tilde{c}_{ki} = \frac{\mathbb{E}\left\{\left|\mathbf{h}_k^{\mathrm{H}}\mathbf{D}_i\bar{\mathbf{w}}_i\right|^2\right\}}{\mathbb{E}\left\{\|\bar{\mathbf{w}}_i\|^2\right\}}, \quad \forall k, \forall i \neq k. \tag{7.15}$$

Recall that $\{\bar{\mathbf{w}}_k : k = 1, \ldots, K\}$ are vectors that point out the directions of the centralized precoding vectors. MMSE, P-MMSE, and P-RZF precoding were considered in Section 6.1.2 on p. 362 and represent different ways of selecting $\bar{\mathbf{w}}_k$. The choice of precoding scheme will affect the values of $\tilde{b}_k$ and $\tilde{\mathbf{c}}_k$, but the same optimization algorithms can be utilized for any precoding since the SINRs are always given by (7.12).

We note that the effective downlink SINR in (7.12) has the same structure as the effective uplink SINR in (7.1), except that the K transmit power coefficients are now contained in a vector that we denote $\boldsymbol{\rho}$. Hence, we can apply almost the same optimization algorithms to solve the max-min fairness and sum SE maximization problems, with the only structural difference that we have one power constraint per AP. Let $\bar{\mathbf{w}}'_{kl} \in \mathbb{C}^N$ denote the portion of the normalized centralized precoding vector $\bar{\mathbf{w}}_k / \sqrt{\mathbb{E}\{\|\bar{\mathbf{w}}_k\|^2\}}$ corresponding to AP l, so we have

$$\frac{\bar{\mathbf{w}}_k}{\sqrt{\mathbb{E}\left\{\|\bar{\mathbf{w}}_k\|^2\right\}}} = \begin{bmatrix} \bar{\mathbf{w}}'_{k1} \\ \vdots \\ \bar{\mathbf{w}}'_{kL} \end{bmatrix}. \tag{7.16}$$

Note that $\sum_{l=1}^{L} \mathbb{E}\{\|\bar{\mathbf{w}}'_{kl}\|^2\} = 1$ by construction, while the individual vectors $\{\bar{\mathbf{w}}'_{kl} : l \in \mathcal{M}_k\}$ can have arbitrary average squared norms between 0 and 1. The value $\mathbb{E}\{\|\bar{\mathbf{w}}'_{kl}\|^2\} \in [0, 1]$ represents the fraction of the total transmit power assigned to UE k that will be sent from AP l. Hence, the power constraint for AP l can be formulated as

$$\sum_{k \in \mathcal{D}_l} \rho_k \mathbb{E}\left\{\left\|\bar{\mathbf{w}}'_{kl}\right\|^2\right\} \leq \rho_{\max}. \tag{7.17}$$

We first consider the max-min SE fairness problem that is formulated for the general case in (3.33). In the special case at hand, there are L per-AP transmit power constraints of the kind in (7.17). The max-min fairness optimization problem can then be particularized as

$$\begin{aligned} \underset{\boldsymbol{\rho} \geq \mathbf{0}_K}{\text{maximize}} \quad & \min_{k \in \{1,\ldots,K\}} \frac{\tilde{b}_k \rho_k}{\tilde{\mathbf{c}}_k^{\mathrm{T}} \boldsymbol{\rho} + \sigma_{\mathrm{dl}}^2} \\ \text{subject to} \quad & \sum_{k \in \mathcal{D}_l} \rho_k \mathbb{E}\left\{ \left\| \bar{\mathbf{w}}'_{kl} \right\|^2 \right\} \leq \rho_{\max}, \quad l = 1, \ldots, L. \end{aligned} \tag{7.18}$$

The above problem has the same structure as the generic problem in (3.33) with $R = L$ and $\mathbf{a}_r$ is the vector whose elements are $\mathbb{E}\{\|\bar{\mathbf{w}}'_{kr}\|^2\}$ for UEs that are served by AP r and zero elsewhere. Note that the three conditions in Lemma 3.9 on p. 258 are satisfied, just as in uplink, and we can thus design a similar fixed-point algorithm that converges to the optimal solution of (7.18). The specific steps are provided in Algorithm 7.3. This algorithm will converge quickly but the complexity grows with K. This means that it is not scalable to optimize the powers in this way in a large cell-free network. As usual for max-min fairness problems, all UEs will obtain the same SE at the optimum.

Algorithm 7.3 Fixed-point algorithm for solving the centralized downlink max-min fairness problem in (7.18).

1: **Initialization:** Set arbitrary initial power $\boldsymbol{\rho} > \mathbf{0}_K$ and the solution accuracy $\epsilon > 0$
2: **while** $\max_{k \in \{1,\ldots,K\}} \mathrm{SINR}_k(\boldsymbol{\rho}) - \min_{k \in \{1,\ldots,K\}} \mathrm{SINR}_k(\boldsymbol{\rho}) > \epsilon$ **do**
3: $\quad \rho_k \leftarrow \frac{\rho_k}{\mathrm{SINR}_k(\boldsymbol{\rho})}$, $k = 1, \ldots, K$
4: $\quad \boldsymbol{\rho} \leftarrow \frac{\rho_{\max}}{\max_{l \in \{1,\ldots,L\}} \sum_{k \in \mathcal{D}_l} \rho_k \mathbb{E}\{\|\bar{\mathbf{w}}'_{kl}\|^2\}} \boldsymbol{\rho}$
5: **end while**
6: **Output:** Optimal transmit powers $\boldsymbol{\rho}$
7: Max-min SE $\min_{k \in \{1,\ldots,K\}} \frac{\tau_d}{\tau_c} \log_2(1 + \mathrm{SINR}_k(\boldsymbol{\rho}))$

We now consider the downlink sum SE maximization problem in

the centralized operation, which is formulated as

$$\begin{aligned} \underset{\boldsymbol{\rho}\geq\mathbf{0}_K}{\text{maximize}} \quad & \sum_{k=1}^{K} \log_2\left(1+\frac{\tilde{b}_k\rho_k}{\tilde{\mathbf{c}}_k^{\mathrm{T}}\boldsymbol{\rho}+\sigma_{\mathrm{dl}}^2}\right) \\ \text{subject to} \quad & \sum_{k\in\mathcal{D}_l} \rho_k \mathbb{E}\left\{\left\|\bar{\mathbf{w}}_{kl}'\right\|^2\right\} \leq \rho_{\max}, \quad l=1,\ldots,L. \end{aligned} \tag{7.19}$$

This problem also resembles the uplink counterpart with the main structural difference being the power constraints that are coupled between the optimization variables. We notice that (7.19) is a problem of the kind in (3.38) with $R = L$ and $\mathbf{a}_r$ defined in the same way as for the max-min fairness problem above. This implies that the sum SE maximization problem is not convex but a local optimal solution can be obtained by particularizing the block coordinate descent algorithm in Algorithm 3.3 on p. 261 for the problem at hand. This algorithm utilizes a weighted MMSE reformulation of the sum SE maximization problem and the corresponding MSE in (3.39) becomes

$$e_k\left(\boldsymbol{\rho}, u_k\right) = u_k^2\left(\tilde{b}_k\rho_k + \tilde{\mathbf{c}}_k^{\mathrm{T}}\boldsymbol{\rho} + \sigma_{\mathrm{dl}}^2\right) - 2u_k\sqrt{\tilde{b}_k\rho_k} + 1 \tag{7.20}$$

in the considered downlink problem. Algorithm 7.4 provides the resulting algorithm and (7.21) in Step 6 contains a subproblem that needs to be solved in every iteration. If we treat $\{\sqrt{\rho_k} : k = 1,\ldots,K\}$ as the optimization variables, (7.21) becomes a convex quadratically-constrained quadratic programming problem. Although a closed-form solution does not exist in general due to the multiple power constraints involving the same variables, any convex solver can be utilized to solve this subproblem [**Boyd2004a**].

Remark 7.2 (General-purpose convex solvers)**.** A convex optimization problem that lacks a closed-form solution can be solved either by a dedicated solver, which is developed to exploit the special structure of the problem at hand, or by a general-purpose solver that has been optimized for a wide class of problems. The former approach is preferred from a runtime efficiency perspective, while the latter approach is preferred for faster code development since the finer implementation details are abstracted away. We took the second approach when comparing power

allocation schemes in Section 7.3. More precisely, we made use of the solver SDPT3 [**SDPT3**] and wrote the code using CVX [**cvx2014**], an interface for specifying convex problems and connect them to a solver. An example of the first approach is [**Chakraborty2020a**].

Algorithm 7.4 Block coordinate descent algorithm for solving the sum SE maximization problem in (7.19).

1: **Initialization:** Set the solution accuracy $\epsilon > 0$
2: Set an arbitrary feasible initial power vector $\boldsymbol{\rho}$
3: **while** $\sum_{k=1}^{K} \left(d_k e_k \left(\boldsymbol{\rho}, u_k \right) - \ln \left(d_k \right) \right)$ is either improved more than ϵ or not improved at all **do**
4: $\quad u_k \leftarrow \frac{\sqrt{\tilde{b}_k \rho_k}}{\tilde{b}_k \rho_k + \tilde{\mathbf{c}}_k^{\mathrm{T}} \boldsymbol{\rho} + \sigma_{\mathrm{dl}}^2}, \quad k = 1, \ldots, K$
5: $\quad d_k \leftarrow 1 \Big/ \left(u_k^2 \left(\tilde{b}_k \rho_k + \tilde{\mathbf{c}}_k^{\mathrm{T}} \boldsymbol{\rho} + \sigma_{\mathrm{dl}}^2 \right) - 2 u_k \sqrt{\tilde{b}_k \rho_k} + 1 \right), \quad k = 1, \ldots, K$
6: $\quad$ Solve the following convex problem for the current values of u_k and d_k:

$$\begin{aligned} \underset{\boldsymbol{\rho} \geq \mathbf{0}_K}{\text{minimize}} \quad & \sum_{k=1}^{K} d_k e_k \left(\boldsymbol{\rho}, u_k \right) \\ \text{subject to} \quad & \sum_{k \in \mathcal{D}_l} \rho_k \mathbb{E} \left\{ \left\| \bar{\mathbf{w}}_{kl}' \right\|^2 \right\} \leq \rho_{\max}, \quad l = 1, \ldots, L \end{aligned} \tag{7.21}$$

7: $\quad$ Update $\boldsymbol{\rho}$ by the obtained solution to (7.21).
8: **end while**
9: **Output:** Optimal transmit powers $\boldsymbol{\rho}$
10: Sum SE $\frac{\tau_d}{\tau_c} \sum_{k=1}^{K} \log_2 \left(d_k \right)$

Optimized Distributed Downlink Power Allocation

In the distributed downlink operation, there are $\sum_{k=1}^{K} |\mathcal{M}_k|$ power allocation coefficients to optimize: $\{\rho_{kl} : l \in \mathcal{M}_k, k = 1, \ldots, K\}$. This number is larger than K (as in the centralized operation), except in the extreme case when each UE is only served by one AP. For example, if each AP serves τ_p UEs (one per pilot), then there will be $L\tau_p$ power coefficients to optimize in the distributed operation.

The same SE expression is used as in the centralized operation, but the effective SINRs have a different dependence on the optimization variables than in the centralized case. To obtain tractable formulations, we first introduce a new set of optimization variables:

$$\tilde{\rho}_{kl} = \sqrt{\rho_{kl}} \geq 0 \tag{7.22}$$

for $k \in \mathcal{D}_l$ and $l = 1, \ldots, L$. These are the square roots of the transmit power coefficients. There is a one-to-one correspondence in (7.22) between the power variables, thus optimization with respect to the new variables $\{\tilde{\rho}_{kl} : l \in \mathcal{M}_k, k = 1, \ldots, K\}$ is equivalent to the optimization with respect to the original ones. The benefit of the new variables is that the effective downlink SINR for UE k in (6.22) can be expressed as

$$\mathsf{SINR}_k^{(\mathrm{dl},\mathsf{d})}(\{\tilde{\boldsymbol{\rho}}_i\}) = \frac{\left|\tilde{\mathbf{b}}_k^{\mathrm{T}} \tilde{\boldsymbol{\rho}}_k\right|^2}{\sum\limits_{i=1}^{K} \tilde{\boldsymbol{\rho}}_i^{\mathrm{T}} \tilde{\mathbf{C}}_{ki} \tilde{\boldsymbol{\rho}}_i - \left|\tilde{\mathbf{b}}_k^{\mathrm{T}} \tilde{\boldsymbol{\rho}}_k\right|^2 + \sigma_{\mathrm{dl}}^2} \tag{7.23}$$

where

$$\tilde{\boldsymbol{\rho}}_k = [\tilde{\rho}_{k1} \ \ldots \ \tilde{\rho}_{kL}]^{\mathrm{T}} \in \mathbb{R}_{\geq 0}^{L}, \quad \forall k \tag{7.24}$$

$$\tilde{\mathbf{b}}_k \in \mathbb{R}_{\geq 0}^{L},$$

$$\left[\tilde{\mathbf{b}}_k\right]_l = \begin{cases} \frac{\mathbb{E}\{\mathbf{h}_{kl}^{\mathrm{H}} \bar{\mathbf{w}}_{kl}\}}{\sqrt{\mathbb{E}\{\|\bar{\mathbf{w}}_{kl}\|^2\}}} & l \in \mathcal{M}_k \\ 0 & l \notin \mathcal{M}_k, \end{cases} \quad \forall k \tag{7.25}$$

$$\tilde{\mathbf{C}}_{ki} \in \mathbb{C}^{L \times L},$$

$$\left[\tilde{\mathbf{C}}_{ki}\right]_{lr} = \begin{cases} \frac{\mathbb{E}\{\mathbf{h}_{kl}^{\mathrm{H}} \bar{\mathbf{w}}_{il} \bar{\mathbf{w}}_{ir}^{\mathrm{H}} \mathbf{h}_{kr}\}}{\sqrt{\mathbb{E}\{\|\bar{\mathbf{w}}_{il}\|^2\}} \sqrt{\mathbb{E}\{\|\bar{\mathbf{w}}_{ir}\|^2\}}} & l \in \mathcal{M}_i, r \in \mathcal{M}_i \\ 0 & l \notin \mathcal{M}_i \text{ or } r \notin \mathcal{M}_i, \end{cases} \quad \forall k, i. \tag{7.26}$$

The elements of $\tilde{\mathbf{b}}_k$ are assumed to be real and non-negative, which is satisfied when using L-MMSE, LP-MMSE, and MR precoding. This condition can also be satisfied without loss of generality for any other type of precoding by rotating the phase of the precoding vectors [**Bengtsson2001a**].

The max-min fairness problem can now be formulated for the distributed downlink operation as

$$\begin{aligned} \underset{\tilde{\boldsymbol{\rho}}_k \geq \mathbf{0}_L, \forall k, t \geq 0}{\text{maximize}} \quad & t \\ \text{subject to} \quad & \frac{\left|\tilde{\mathbf{b}}_k^{\mathrm{T}} \tilde{\boldsymbol{\rho}}_k\right|^2}{\sum\limits_{i=1}^{K} \tilde{\boldsymbol{\rho}}_i^{\mathrm{T}} \tilde{\mathbf{C}}_{ki} \tilde{\boldsymbol{\rho}}_i - \left|\tilde{\mathbf{b}}_k^{\mathrm{T}} \tilde{\boldsymbol{\rho}}_k\right|^2 + \sigma_{\mathrm{dl}}^2} \geq t, \quad k = 1, \ldots, K \\ & \sum_{k \in \mathcal{D}_l} \tilde{\rho}_{kl}^2 \leq \rho_{\max}, \quad l = 1, \ldots, L \end{aligned} \tag{7.27}$$

by utilizing the auxiliary variable t that represents the minimum SINR among all UEs. Note that the power constraint $\sum_{k \in \mathcal{D}_l} \tilde{\rho}_{kl}^2 \leq \rho_{\max}$ of AP l sums up all the squared coefficients related to this AP. We notice that (7.27) is an instance of the generic problem formulation in (3.34), thus an optimal solution to (7.27) can be obtained using Algorithm 3.1 on p. 257. This algorithm contains a bisection search over t where a solution to (3.36) is computed for each candidate value. By including a minimization of the total transmit power in the objective as in (3.36) to improve convergence rate, the subproblem for the problem at hand can be expressed in the second-order cone programming form:

$$\begin{aligned} \underset{\tilde{\boldsymbol{\rho}}_k \geq \mathbf{0}_L, \forall k}{\text{minimize}} \quad & \sum_{k=1}^{K} \|\tilde{\boldsymbol{\rho}}_k\|^2 \\ \text{subject to} \quad & \left\| \begin{bmatrix} \tilde{\mathbf{C}}_{k1}^{\frac{1}{2}} \tilde{\boldsymbol{\rho}}_1 \\ \vdots \\ \tilde{\mathbf{C}}_{kK}^{\frac{1}{2}} \tilde{\boldsymbol{\rho}}_K \\ \sqrt{\sigma_{\mathrm{dl}}^2} \end{bmatrix} \right\| \leq \sqrt{\frac{1 + t^{\text{candidate}}}{t^{\text{candidate}}}} \tilde{\mathbf{b}}_k^{\mathrm{T}} \tilde{\boldsymbol{\rho}}_k, \quad k = 1, \ldots, K \\ & \left\| \begin{bmatrix} \tilde{\rho}_{1l} & \ldots & \tilde{\rho}_{Kl} \end{bmatrix} \right\| \leq \sqrt{\rho_{\max}}, \quad l = 1, \ldots, L \\ & \tilde{\boldsymbol{\rho}}_k \geq \mathbf{0}_L, \quad k = 1, \ldots, K \end{aligned} \tag{7.28}$$

where we implicitly set $\tilde{\rho}_{kl}$ to zero for $k \notin \mathcal{D}_l$ in the second constraint.

This reformulation is achieved by noticing that the SINR constraint $\mathsf{SINR}_k^{(\mathsf{dl},\mathsf{d})}(\{\tilde{\boldsymbol{\rho}}_i\}) \geq t^{\text{candidate}}$ can be equivalently expressed as

$$\frac{(\tilde{\mathbf{b}}_k^{\mathrm{T}}\tilde{\boldsymbol{\rho}}_k)^2}{\left\| \left[\tilde{\mathbf{C}}_{k1}^{\frac{1}{2}}\tilde{\boldsymbol{\rho}}_1 \quad \ldots \quad \tilde{\mathbf{C}}_{kK}^{\frac{1}{2}}\tilde{\boldsymbol{\rho}}_K \quad \sqrt{\sigma_{\text{dl}}^2} \right] \right\|^2 - (\tilde{\mathbf{b}}_k^{\mathrm{T}}\tilde{\boldsymbol{\rho}}_k)^2} \geq t^{\text{candidate}} \tag{7.29}$$

and then rearranged as in the first constraint of (7.28). This is a trick that first appeared in [**Bengtsson2001a**] and then was used for Cell-free Massive MIMO in [**Ngo2017b**]. By utilizing the subproblem in (7.28), we obtain Algorithm 7.5.

It remains to solve the downlink sum SE maximization problem for the distributed operation, which can be formulated as

$$\begin{aligned} \underset{\tilde{\boldsymbol{\rho}}_k \geq \mathbf{0}_L, \forall k}{\text{maximize}} \quad & \sum_{k=1}^{K} \log_2 \left(1 + \frac{\left| \tilde{\mathbf{b}}_k^{\mathrm{T}}\tilde{\boldsymbol{\rho}}_k \right|^2}{\sum\limits_{i=1}^{K} \tilde{\boldsymbol{\rho}}_i^{\mathrm{T}}\tilde{\mathbf{C}}_{ki}\tilde{\boldsymbol{\rho}}_i - \left| \tilde{\mathbf{b}}_k^{\mathrm{T}}\tilde{\boldsymbol{\rho}}_k \right|^2 + \sigma_{\text{dl}}^2} \right) \\ \text{subject to} \quad & \sum_{k \in \mathcal{D}_l} \tilde{\rho}_{kl}^2 \leq \rho_{\max}, \quad l = 1, \ldots, L. \end{aligned} \tag{7.31}$$

We recognize (7.31) as an instance of the generic sum SE maximization problem in (3.38). Due to Lemma 3.10 on p. 260, a local optimal solution to (7.31) can be obtained by the block coordinate descent algorithm given in Algorithm 7.6. This algorithm is designed based on a weighted MMSE reformulation of the sum SE maximization problem, where the MSE in (3.39) becomes

$$e_k(\{\tilde{\boldsymbol{\rho}}_i\}, u_k) = u_k^2 \left(\sum_{i=1}^{K} \tilde{\boldsymbol{\rho}}_i^{\mathrm{T}}\tilde{\mathbf{C}}_{ki}\tilde{\boldsymbol{\rho}}_i + \sigma_{\text{dl}}^2 \right) - 2u_k \tilde{\mathbf{b}}_k^{\mathrm{T}}\tilde{\boldsymbol{\rho}}_k + 1 \tag{7.32}$$

for the considered downlink problem. Step 6 in Algorithm 7.6 contains a subproblem that needs to be solved at each iteration. It is a convex quadratically-constrained quadratic programming problem, which can be solved using any numerical solver for convex problems. See Remark 7.2 for some suggested tools.

Remark 7.3 (Other utility functions and scheduling)**.** There are other utility functions than max-min fairness and sum SE that can be maximized. For example, one can introduce user-specific weights in the utility

Algorithm 7.5 Bisection search algorithm for solving the max-min fairness problem in (7.27).

Initialization: Set the solution accuracy $\epsilon > 0$
Set the initial lower and upper bounds for the max-min SINR:
$t^{\text{lower}} \leftarrow 0$
$t^{\text{upper}} \leftarrow \min\limits_{k\in\{1,\ldots,K\}} \quad |\mathcal{M}_k|\rho_{\max}\frac{\tilde{\mathbf{b}}_k^{\mathrm{T}}\tilde{\mathbf{b}}_k}{\sigma_{\text{dl}}^2}$
Initialize solution variables: $\tilde{\boldsymbol{\rho}}_k^{\text{opt}} = \mathbf{0}_L$ for $k = 1, \ldots, K$, $t^{\text{opt}} = 0$
while $t^{\text{upper}} - t^{\text{lower}} > \epsilon$ **do**
    $t^{\text{candidate}} \leftarrow \frac{t^{\text{lower}}+t^{\text{upper}}}{2}$
    Solve the following convex problem:

$$\begin{aligned} \underset{\tilde{\boldsymbol{\rho}}_k \geq \mathbf{0}_L, \forall k}{\text{minimize}} \quad & \sum_{k=1}^{K} \|\tilde{\boldsymbol{\rho}}_k\|^2 \\ \text{subject to} \quad & \left\| \begin{bmatrix} \tilde{\mathbf{C}}_{k1}^{\frac{1}{2}}\tilde{\boldsymbol{\rho}}_1 \\ \vdots \\ \tilde{\mathbf{C}}_{kK}^{\frac{1}{2}}\tilde{\boldsymbol{\rho}}_K \\ \sqrt{\sigma_{\text{dl}}^2} \end{bmatrix} \right\| \leq \sqrt{\frac{1+t^{\text{candidate}}}{t^{\text{candidate}}}}\tilde{\mathbf{b}}_k^{\mathrm{T}}\tilde{\boldsymbol{\rho}}_k, \quad k = 1, \ldots, K \\ & \left\| \begin{bmatrix} \tilde{\rho}_{1l} & \ldots & \tilde{\rho}_{Kl} \end{bmatrix} \right\| \leq \sqrt{\rho_{\max}}, \quad l = 1, \ldots, L \\ & \tilde{\boldsymbol{\rho}}_k \geq \mathbf{0}_L, \quad k = 1, \ldots, K \end{aligned} \tag{7.30}$$

    if (7.30) is feasible **then**
        $t^{\text{lower}} \leftarrow t^{\text{candidate}}$
        $\tilde{\boldsymbol{\rho}}_k^{\text{opt}} \leftarrow \tilde{\boldsymbol{\rho}}_k$, which is the solution to (7.30) for $k = 1, \ldots, K$
    else
        $t^{\text{upper}} \leftarrow t^{\text{candidate}}$
    end if
end while
Output: Optimal square-roots of the transmit powers $\tilde{\boldsymbol{\rho}}_1^{\text{opt}}, \ldots, \tilde{\boldsymbol{\rho}}_K^{\text{opt}}$, $t^{\text{opt}} = \min\limits_{k\in\{1,\ldots,K\}} \mathsf{SINR}_k\left(\left\{\tilde{\boldsymbol{\rho}}_i^{\text{opt}}\right\}\right)$
Max-min SE $\frac{\tau_d}{\tau_c}\log_2\left(1+t^{\text{opt}}\right)$

Algorithm 7.6 Block coordinate descent algorithm for solving the sum SE maximization problem in (7.31).

Initialization: Set the solution accuracy $\epsilon > 0$

Set arbitrary feasible initial powers $\{\tilde{\boldsymbol{\rho}}_k\}$

while $\sum_{k=1}^{K} \left(d_k e_k \left(\{\tilde{\boldsymbol{\rho}}_i\}, u_k \right) - \ln\left(d_k\right) \right)$ is either improved more than ϵ or not improved at all **do**

$$u_k \leftarrow \frac{\tilde{\mathbf{b}}_k^{\mathrm{T}} \tilde{\boldsymbol{\rho}}_k}{\sum\limits_{i=1}^{K} \tilde{\boldsymbol{\rho}}_i^{\mathrm{T}} \tilde{\mathbf{C}}_{ki} \tilde{\boldsymbol{\rho}}_i + \sigma_{\mathrm{dl}}^2}, \quad k = 1, \ldots, K$$

$$d_k \leftarrow \frac{1}{u_k^2 \left(\sum\limits_{i=1}^{K} \tilde{\boldsymbol{\rho}}_i^{\mathrm{T}} \tilde{\mathbf{C}}_{ki} \tilde{\boldsymbol{\rho}}_i + \sigma_{\mathrm{dl}}^2 \right) - 2u_k \tilde{\mathbf{b}}_k^{\mathrm{T}} \tilde{\boldsymbol{\rho}}_k + 1}, \quad k = 1, \ldots, K$$

Solve the following problem for the current values of u_k and d_k:

$$\begin{aligned} \underset{\tilde{\boldsymbol{\rho}}_k \geq \mathbf{0}_L, \forall k}{\text{minimize}} \quad & \sum_{k=1}^{K} d_k \left(u_k^2 \left(\sum_{i=1}^{K} \tilde{\boldsymbol{\rho}}_i^{\mathrm{T}} \tilde{\mathbf{C}}_{ki} \tilde{\boldsymbol{\rho}}_i + \sigma_{\mathrm{dl}}^2 \right) - 2u_k \tilde{\mathbf{b}}_k^{\mathrm{T}} \tilde{\boldsymbol{\rho}}_k \right) \\ \text{subject to} \quad & \sum_{k \in \mathcal{D}_l} \tilde{\rho}_{kl}^2 \leq \rho_{\max}, \quad l = 1, \ldots, L \end{aligned} \tag{7.33}$$

Update $\{\tilde{\boldsymbol{\rho}}_k\}$ by the obtained solution to (7.33).

end while

Output: Optimal square-roots of the transmit powers $\tilde{\boldsymbol{\rho}}_1, \ldots, \tilde{\boldsymbol{\rho}}_K$

Sum SE $\frac{\tau_d}{\tau_c} \sum_{k=1}^{K} \log_2\left(d_k\right)$

functions to get the weighted max-min fairness and weighted sum SE utilities. The algorithms presented in this monograph can be extended to handle these cases as well. Moreover, there are utility functions that have a totally different structure, such as the geometric mean of the SEs (also known as proportional fairness) and the harmonic mean of the SEs [**Bjornson2013d**], [**Farooq2020a**], [**Luo2008a**]. However, apart from limiting the length of this monograph, there are two strong reasons for why we only covered the max-min fairness and sum SE utilities. The first reason is that the early works [**Nayebi2017a**], [**Ngo2017b**] on Cell-free Massive MIMO emphasized the importance of the max-min fairness utility, and the resulting ability of cell-free networks to deliver uniformly good service over large coverage areas. This is why we considered that metric. The second reason is that the weighted sum SE maximization problem is of main importance in practical networks, where the UEs have data queues of limited lengths and packets are arriving at random to these queues. One can then formulate dynamic resource allocation problems that take the queue dynamics into account [**Georgiadis2006a**], [**Shirani2010a**] and irrespective of what the long-term utility function is, the problems that need to be solved regularly are weighted sum SE maximization problem where the weights are computed based on the lengths of the data queues and the long-term utility. The first step into dynamic resource allocation for cell-free networks was taken in [**Chen2020b**], which only considers the uplink. There appear to be many open research questions related to the interplay between scheduling and power allocation in Cell-free Massive MIMO.

7.2 Scalable Distributed Power Optimization

The algorithms presented in Section 7.1 jointly optimize the transmit powers for all UEs to maximize a network-wide utility function. Since there are K SINRs to optimize and at least K optimization variables, it is unavoidable that the computational complexity becomes unscalable as $K \to \infty$. In fact, the complexity of any network-wide power allocation optimization grows unboundedly with K, which makes them unscalable according to Definition 2.2 on p. 217. Hence, to obtain a scalable power allocation scheme that is practically implementable in large cell-free

networks, we need to devise distributed and heuristic schemes, where each AP or UE makes a local decision with limited involvement of the other APs/UEs. For example, a device can make decisions based on the CSI that it can acquire locally.

It is easy to design a scalable scheme; for example, we can let every UE transmit with full power in the uplink and let every AP allocate its power equally among the UEs it serves in the downlink. The challenge is to identify heuristic schemes that also provide reasonably good SEs, according to network-wide utility functions. This basically boils down to numerically evaluating different heuristic schemes against the optimized baselines of the kind presented in Section 7.1. In this section, we will present some selected recent approaches to distributed power allocation. We refer to [**Bjornson2011a**], [**Bjornson2010c**], [**Interdonato2019a**], [**Nayebi2017a**], [**Bjornson2013d**], for further examples and stress that further research is needed in this direction.

7.2.1 Scalable Uplink Power Control

Fractional power control is a classical heuristic in the uplink of multi-user systems, particularly in cellular networks [**Simonsson2008a**], [**Whitehead1993a**]. It builds on the principle of using power control to compensate for a fraction of the pathloss differences within each cell. Power control is a balancing act between utilizing the pathloss differences to provide the cell-center UEs with high SEs (at the expense of causing interference to other UEs and cells) and compensating for the pathloss differences to improve the SE for cell-edge UEs (at the expense of forcing cell-center UEs to reduce their power and thereby their SE). The characteristic feature of fractional power control is that each UE computes its transmit power as a function of only its channel gain to the serving AP, but since the same function is utilized by all UEs, the interference statistics can also be implicitly tuned by an appropriate function selection [**Whitehead1993a**].

The power control situation is different in cell-free networks compared to cellular networks since each UE has multiple serving APs. A UE can have a strong channel to one serving AP but a weak channel to another serving AP, while another UE experiences the opposite situation. *Should any of the UEs reduce its power in this situation to limit the mu-*

tual interference and, if yes, how much? There are no simple answers but there are algorithms that work fairly well and these are essentially adding up the channel gains from the serving APs as if there was only one AP.

A fractional power control algorithm for cell-free networks was proposed and motivated in [**Nikbakht2019a**], [**Nikbakht2020a**] for a setup where all APs serve all UEs using MR combining, but it can be easily adapted to the case when each AP serves a subset of the UEs using an arbitrary combining scheme [**Chen2020a**]. UE k selects its uplink transmit power as

$$p_k = p_{\max} \frac{\left(\sum_{l \in \mathcal{M}_k} \beta_{kl} \right)^{\upsilon}}{\max_{i \in \{1,\ldots,K\}} \left(\sum_{l \in \mathcal{M}_i} \beta_{il} \right)^{\upsilon}} \tag{7.34}$$

where the exponent υ dictates the power control behavior. The denominator in (7.34) makes sure that $p_k \in [0, p_{\max}]$. Note that $\sum_{l \in \mathcal{M}_i} \beta_{il}$ is the total channel gain from UE i to the APs that serve it. If $\upsilon = 0$, all UEs transmit with maximum power as assumed in the simulations in Section 5.4 on p. 331. If $\upsilon = -1$, each UE is fully compensating for the variations in the total channel gain among the UEs so that $p_i \sum_{l \in \mathcal{M}_i} \beta_{il}$ becomes the same for $i = 1, \ldots, K$. This can be viewed as an approximation of max-min fairness power control. Fractional power control traditionally consists of finding a tradeoff between these extremes by selecting $\upsilon \in [-1, 0]$ so this is the range considered in [**Nikbakht2019a**], [**Nikbakht2020a**]. To identify an approximation of sum SE maximization, we also need to consider $\upsilon > 0$ so more power is used by the UEs with good channel conditions. Another possible generalization is to replace β_{kl} with $\beta_{kl} - \mathrm{tr}(\mathbf{C}_{kl})/N$ to take the performance loss due to channel estimation errors into account.

One problem with the fractional power control scheme described above is that it is not scalable since we need to compute the maximum of K terms in the denominator of (7.34). However, we can easily modify (7.34) such that the number of the terms in the denominator does not grow with K and thereby achieve scalability. In cellular networks, this is normally done by replacing the denominator with a constant representing the worst-case cell-edge conditions (e.g., what is the lowest channel gain that a connected UE can have). This makes good sense

in symmetric cellular deployments, where this value is roughly the same in every cell, while it can be an overly pessimistic approximation in asymmetric cell-free networks. Inspired by the scalable combining schemes from Section 5.1.4 on p. 305, one option is to only compute the maximum with respect to the indices of the UEs that are partially served by the same APs. This makes intuitive sense in large networks where each UE is mostly affected by the closest UEs. Using the set $\mathcal{S}_k$ defined in (5.15), for which $|\mathcal{S}_k|$ does not grow unboundedly when K goes to infinity, a scalable version of the fractional power control in (7.34) is obtained by selecting the transmit power of UE k as

$$p_k = p_{\max} \frac{\left(\sum_{l \in \mathcal{M}_k} \beta_{kl} \right)^{\upsilon}}{\max_{i \in \mathcal{S}_k} \left(\sum_{l \in \mathcal{M}_i} \beta_{il} \right)^{\upsilon}}. \tag{7.35}$$

Depending on the value of υ, we can heuristically aim at optimizing different utility functions. To find the desired value, the network designer can simulate the CDF of the per-UE SE for different values of υ and then determine which value gives the most desirable tradeoff between high fairness and high sum SE.

Remark 7.4 (Minimizing the signal-to-interference ratio)**.** Fractional power control is a heuristic scheme in the sense that it is not directly connected to maximizing a utility function of the individual SEs. However, under certain conditions, such power control can be shown to minimize the variations in the signal-to-interference ratio (SIR) experienced by UEs at random locations. For example, [**Whitehead1993a**] proved that if the channel gains are Gaussian distributed in the decibel scale and have equal variance, then fractional power control with $\upsilon = -1/2$ minimizes the variance of the SIR in a two-UE setup. A generalization of this result to cell-free networks is provided in [**Nikbakht2020a**] and serves as a motivation for why fractional power control makes practical sense to shape the CDF curve of the SIR experienced at random locations, even if the connections to the CDF of the per-UE SE or to utility functions such as the sum SE and max-min fairness are heuristic.

7.2.2 Scalable Centralized Downlink Power Allocation

In the centralized downlink operation, the transmit powers that different APs are transmitting with to a given UE are coupled through the centralized precoding vectors

$$\frac{\bar{\mathbf{w}}_k}{\sqrt{\mathbb{E}\left\{\|\bar{\mathbf{w}}_k\|^2\right\}}} = \begin{bmatrix} \bar{\mathbf{w}}'_{k1} \\ \vdots \\ \bar{\mathbf{w}}'_{kL} \end{bmatrix}, \quad k = 1, \ldots, K. \tag{7.36}$$

If one AP increases its power to UE k, then all the other APs that serve this UE must do the same to keep the same direction of the precoding vector. Otherwise, the ability for the APs to cancel each others' interference (as illustrated in Figure 6.1 on p. 358) is lost. This ability is one of the key benefits of the centralization operation and should not be lost if the performance gain over the distributed operation should be retained. Recall from (7.17) that the power constraint at AP l is

$$\sum_{k \in \mathcal{D}_l} \rho_k \mathbb{E}\left\{\left\|\bar{\mathbf{w}}'_{kl}\right\|^2\right\} \leq \rho_{\max}. \tag{7.37}$$

Since $\sum_{l \in \mathcal{M}_k} \mathbb{E}\{\|\bar{\mathbf{w}}'_{kl}\|^2\} = 1$, the fraction $\mathbb{E}\{\|\bar{\mathbf{w}}'_{kl}\|^2\} \in [0, 1]$ that an arbitrary AP $l \in \mathcal{M}_k$ allocates to UE k is generally much smaller than ρ_k. This must be accounted for so that all the serving APs can select a common power value that satisfies all of their power constraints. A simple heuristic solution is to assign the same power

$$\rho_k = \frac{\rho_{\max}}{\tau_p} \tag{7.38}$$

to all UEs, as proposed in [**Bjornson2020a**]. In this case, the normalized precoding vector in (7.36) determines how this power is distributed between the APs, and all APs are guaranteed to satisfy their power constraints since they serve at most τ_p UEs and will at most allocate $\rho_{\max}/\tau_p$ to each of them. Since the computation of (7.38) is independent of K, this network-wide equal power allocation scheme is scalable.

A more general heuristic can be obtained by taking inspiration from the fractional uplink power control in (7.35) by selecting the downlink power allocation coefficients proportionally to the total channel gain

from the serving APs:

$$\rho_k \propto \left(\sum_{l \in \mathcal{M}_k} \beta_{kl} \right)^{\upsilon}. \tag{7.39}$$

The exponent $\upsilon \in [-1, 1]$ in (7.39) determines the power allocation behavior. The proportionality constant must be selected so that all the power constraints are satisfied, which can be done as follows:

$$\rho_k = \rho_{\max} \frac{\left(\sum_{l \in \mathcal{M}_k} \beta_{kl} \right)^{\upsilon}}{\max_{\ell \in \mathcal{M}_k} \sum_{i \in \mathcal{D}_\ell} \left(\sum_{l \in \mathcal{M}_i} \beta_{il} \right)^{\upsilon}}. \tag{7.40}$$

If $\upsilon = 0$ and $|\mathcal{D}_\ell| = \tau_p$, then (7.40) turns into the network-wide equal power allocation scheme in (7.38). More power is allocated to the UEs with higher total channel gains if $\upsilon > 0$, while it is the other way around if $\upsilon < 0$. The former is resembling the main characteristics of power allocation for maximum sum SE, while the latter is resembling max-min fairness.

The denominator in (7.40) makes sure that all the power constraints are satisfied since $\sum_{i \in \mathcal{D}_l} \rho_i \leq \rho_{\max}$ for $l = 1, \ldots, L$. However, this is a conservative design for the worst-case situation that each UE is only served by only one AP, which is generally not the case in cell-free networks. The consequence is that all APs will operate far below their maximum power. Suppose we know the largest fraction of ρ_k that any of the serving APs can be assigned to transmit:

$$\omega_k = \max_{\ell \in \mathcal{M}_k} \mathbb{E} \left\{ \left\| \bar{\mathbf{w}}'_{k\ell} \right\|^2 \right\}. \tag{7.41}$$

We can then use it as an additional tuning parameter and change (7.39) into

$$\rho_k \propto \frac{\left(\sum_{l \in \mathcal{M}_k} \beta_{kl} \right)^{\upsilon}}{\omega_k^{\kappa}} \tag{7.42}$$

where the exponent $0 \leq \kappa \leq 1$ is a parameter that reshapes the ratio of power allocation between different UEs. The motivation for such

a scaling factor is as follows. When the power allocation in (7.40) is used and $\omega_k \approx 1/|\mathcal{M}_k|$ (i.e., the smallest value it can take), then all the serving APs will approximately transmit with power $\rho_k/|\mathcal{M}_k|$ to UE k but manage their power constraints as if they transmitted with power ρ_k. To prevent this situation to some extent, we can scale the original power coefficient of each UE inversely proportional to ω_k^κ as in (7.42). The intuition is that each AP should instead manage its power constraint as if it transmits with power $\rho_k \omega_k^\kappa$ (note that $\rho_k \omega_k^\kappa \geq \rho_k \omega_k$). To satisfy the power constraint at each AP, we select the proportionality constant in (7.42) to obtain

$$\rho_k = \rho_{\max} \frac{\left(\sum_{l \in \mathcal{M}_k} \beta_{kl} \right)^{\upsilon} \omega_k^{-\kappa}}{\max\limits_{\ell \in \mathcal{M}_k} \sum\limits_{i \in \mathcal{D}_\ell} \left(\sum\limits_{l \in \mathcal{M}_i} \beta_{il} \right)^{\upsilon} \omega_i^{1-\kappa}}. \tag{7.43}$$

The following chain of inequalities demonstrates that the power constraint is satisfied at each AP l':

$$\begin{aligned} \sum_{k \in \mathcal{D}_{l'}} \rho_k \mathbb{E}\left\{ \left\| \bar{\mathbf{w}}'_{kl'} \right\|^2 \right\} &\leq \sum_{k \in \mathcal{D}_{l'}} \rho_k \omega_k = \sum_{k \in \mathcal{D}_{l'}} \frac{\rho_{\max} \left(\sum_{l \in \mathcal{M}_k} \beta_{kl} \right)^{\upsilon} \omega_k^{1-\kappa}}{\max\limits_{\ell \in \mathcal{M}_k} \sum\limits_{i \in \mathcal{D}_\ell} \left(\sum\limits_{l \in \mathcal{M}_i} \beta_{il} \right)^{\upsilon} \omega_i^{1-\kappa}} \\ &\leq \rho_{\max} \frac{\sum\limits_{k \in \mathcal{D}_{l'}} \left(\sum\limits_{l \in \mathcal{M}_k} \beta_{kl} \right)^{\upsilon} \omega_k^{1-\kappa}}{\max\limits_{\ell \in \bigcap\limits_{k \in \mathcal{D}_{l'}} \mathcal{M}_k} \sum\limits_{i \in \mathcal{D}_\ell} \left(\sum\limits_{l \in \mathcal{M}_i} \beta_{il} \right)^{\upsilon} \omega_i^{1-\kappa}} \\ &\leq \rho_{\max} \end{aligned} \tag{7.44}$$

where the last inequality follows from the fact the set $\bigcap_{k \in \mathcal{D}_{l'}} \mathcal{M}_k$ includes l', thus the numerator is smaller or equal to the denominator.

We stress that (7.43) is a scalable power allocation scheme since the computational complexities associated with the terms in the numerator and denominator do not grow with K. When evaluating the downlink performance in the running example in Section 6.3 on p. 376, we used

(7.43) with $\upsilon = -0.5$ and $\kappa = 0.5$. In Section 7.3, we will investigate numerically how to select the parameters $\upsilon \in [-1, 1]$ and $\kappa \in [0, 1]$.

7.2.3 Scalable Distributed Downlink Power Allocation

In the distributed downlink operation, the power allocation contains multiple power coefficients per UE. On the one hand, this increases the complexity compared to uplink power control, but on the other hand, it becomes easier to find suitable tradeoffs. If a UE is served by one AP with a strong channel and by one AP with a weak channel, then it can be assigned widely different powers from these APs in the downlink, while it is less obvious if it should transmit with high or low power in the uplink. From a UE's perspective, it is preferable to be allocated more downlink power from APs with strong channels than APs with weak channels, since this leads to a higher SNR than the opposite allocation. From an AP's perspective, it is natural to prioritize transmitting to UEs that it has good channels to, compared to UEs with weak channels, because the opposite strategy would cause large interference to the UEs with good channels. These basic principles can be utilized to determine a distributed heuristic power allocation scheme where each AP determines its own transmit power allocation independently of the other APs. Since the number of UEs that are served by a particular AP does not grow with K for a scalable cell-free network (e.g., it serves at most τ_p UEs when we use Algorithm 4.1 on p. 289), this kind of distributed power allocation is scalable.

A distributed power allocation scheme was proposed in [**Interdonato2019a**] for a system with $N = 1$. We will present a generalization for arbitrary N using the notation of this monograph. With this power allocation policy, AP l selects its downlink powers $\rho_{1l}, \ldots, \rho_{Kl}$ as

$$\rho_{kl} = \begin{cases} \rho_{\max} \frac{f(\mathcal{G}_{kl})}{\sum_{i \in \mathcal{D}_l} f(\mathcal{G}_{il})} & k \in \mathcal{D}_l \\ 0 & k \notin \mathcal{D}_l \end{cases} \tag{7.45}$$

where $f(\cdot)$ is a pre-determined function and the input is given by the channel statistics which, for the channel between UE i and AP l, is

$$\mathcal{G}_{il} = \{\mathbf{R}_{il}, \mathbf{C}_{il}\}. \tag{7.46}$$

The intuition is that $f(\mathcal{G}_{il})$ should somehow determine the relative importance of transmitting to UE i, while the term in the denominator of (7.45) normalizes the transmit powers so that the AP is always using its maximum power: $\sum_{k\in\mathcal{D}_l} \rho_{kl} = \rho_{\max}$.

Many different power allocation policies can be formulated according to (7.45). If we select $f(\mathcal{G}_{il}) = 1$, every UE is equally important and we obtain per-AP equal power allocation with $\rho_{kl} = \frac{\rho_{\max}}{|\mathcal{D}_l|}$. Another option is $f(\mathcal{G}_{il}) = (\beta_{il})^{\upsilon}$ [**Interdonato2019a**], which turns (7.45) into

$$\rho_{kl} = \begin{cases} \rho_{\max} \frac{(\beta_{kl})^{\upsilon}}{\sum_{i\in\mathcal{D}_l} (\beta_{il})^{\upsilon}} & k \in \mathcal{D}_l \\ 0 & k \notin \mathcal{D}_l \end{cases} \tag{7.47}$$

where the exponent υ dictates the power allocation behavior. If $\upsilon = 0$, we obtain the per-AP equal power allocation. If $\upsilon = 1$, we give higher emphasis to the UEs according to their respective channel gains. This leads to allocating more power to the UEs with better channel qualities. If $\upsilon = -1$, the power allocation is inversely proportional to the channel gain, so that each of the served UEs will obtain the same received power. This might seem like a good feature from a fairness perspective, but it is generally not since the UEs with good channel conditions will then be subject to high interference. A more opportunistic power allocation with $\upsilon \in [0, 1]$ seems to be preferred to make efficient use of the fact that each UE typically has a good channel from some APs and a worse channel from other APs [**Interdonato2019a**]. Hence, even if (7.47) has an expression that resembles the uplink fractional power control expression in (7.34), the intended operation is very different: we want to emphasize SNR differences instead of mitigating them. When evaluating the downlink performance in the running example in Section 6.3 on p. 376, we used (7.47) with $\upsilon = 0.5$.

Power allocation policies of the kind in (7.45) are scalable by design since only the UEs in $\mathcal{D}_l$ are considered. Another good feature is that every UE will be allocated a non-zero power from all its serving APs, leading to a non-zero SE. Estimation errors can be taken into account by selecting $f(\mathcal{G}_{il}) = (\beta_{il} - \mathrm{tr}(\mathbf{C}_{il})/N)^{\upsilon}$ instead of $f(\mathcal{G}_{il}) = (\beta_{il})^{\upsilon}$, with the consequence of moving power from UEs with lower channel quality

to UEs with better channel quality. For any choice of the function $f(\cdot)$, we can identify the desired value of v in a given network by simulating CDF curves of the per-UE SE for UEs at different locations. By plotting different curves for different values of v, we can determine which value gives the most desirable tradeoff between high fairness and high sum SE.

7.3 Comparison of Power Optimization Schemes

We will now compare the scalable power control/allocation schemes from Section 7.2 with the optimized benchmarks from Section 7.1. To this end, we continue the running example that was defined in Section 5.3 on p. 325. The specific system parameters are: $L = 100$ APs, $N = 4$ antennas per AP, and $K = 40$ UEs. We consider the user-centric implementation based on the DCC formation algorithm presented in Section 4.4 on p. 286 with different scalable processing schemes. The channels follow spatially correlated Rayleigh fading with ASDs $\sigma_\varphi = \sigma_\theta = 15°$.

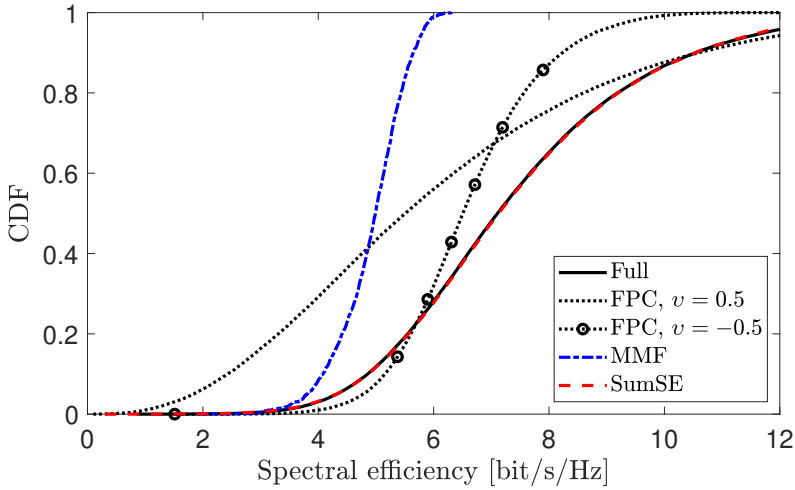
When we compare the performance of different schemes, we will show CDF curves of the SEs that the UEs achieve at different random locations. When we did the same thing in previous chapters, the curves achieved by the different schemes that we compared were often non-intersecting so that we could easily conclude that the rightmost curve was the preferable one. To achieve an efficient scalable implementation, we would pick the rightmost scheme among the scalable alternatives. The situation will be different when we compare different power optimization schemes in this section because we will get curves that intersect, which demonstrates that different schemes are preferable for different types of UEs. Note that the lower tail of a CDF curve represents the performance achieved by the UEs with the worst channel conditions, while the upper tail represents the performance achieved by the UEs with the best channel conditions. One way to measure fairness is by looking at the steepness of the CDF curve; if the curve is steep, then the SE difference between UEs with the worst and best channel conditions is small. However, one should also bear in mind that a network that gives zero SE to everyone features great fairness (everyone gets the same

performance) but it is not practically desirable. Hence, the network designer will eventually have to select a scheme that provides the right tradeoff between fairness and sum SE.

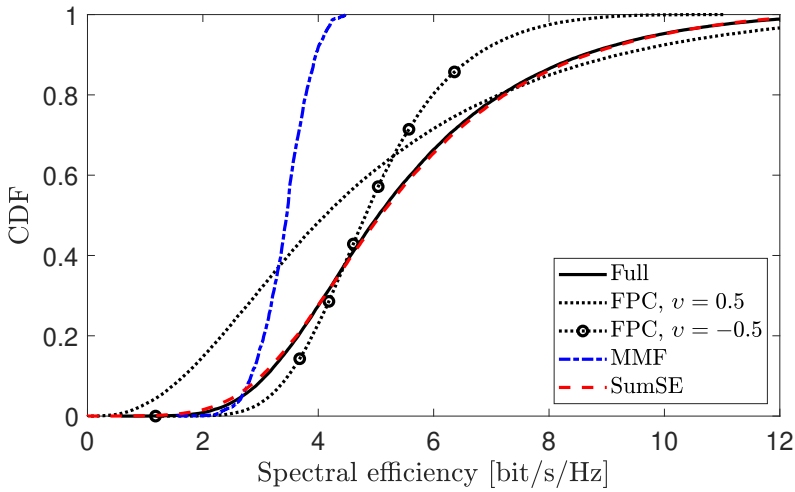
7.3.1 Comparison of Uplink Power Control Schemes

We first consider the uplink and compare both optimized and scalable schemes. As optimized power control schemes, we consider the max-min fairness and sum SE maximization algorithms from Section 7.1.1, namely Algorithm 7.1 and Algorithm 7.2, respectively. The corresponding results are respectively denoted by "MMF" (max-min fairness) and "SumSE". In addition, we consider three heuristic but scalable schemes. The first one is that all UEs transmit with full power, which is denoted by "Full". Moreover, we consider the fractional power control scheme in (7.35) with exponents $\upsilon = 0.5$ or $\upsilon = -0.5$ and denote it by "FPC". Note that the case with $\upsilon = 0.5$ allows UEs with better channel conditions to transmit with higher power whereas the case with $\upsilon = -0.5$ does the opposite.

Figure 7.1 compares the uplink SE of the power control methods described above. We consider a scalable centralized operation with P-MMSE combining in Figure 7.1(a) and a scalable distributed operation with n-opt LSFD and LP-MMSE combining in Figure 7.1(b). In both cases, the sum SE maximizing power control provides the same performance as with full-power transmission. In fact, when the initial point to Algorithm 7.2 is that everyone transmits with maximum power, then the objective function is not improved in the next iteration. Although Algorithm 7.2 attains only a local optimum, we tried several different random initializations and observed that the algorithm always converged to the solution where everyone transmits with full power. This does not mean that the full-power transmission will maximize the sum SE in any Cell-free Massive MIMO system, but it demonstrates that the considered network setup is very capable of suppressing interference so that everyone can transmit at maximum power. If we, hypothetically speaking, would add a UE with extremely bad channel conditions to the network, then the sum SE maximization might allocate zero power to it since this utility function does not provide any performance guarantees.



(a) Centralized uplink operation with P-MMSE combining.



(b) Distributed uplink operation with n-opt LSFD and LP-MMSE combining.

Figure 7.1: CDF of the uplink SE per UE in the centralized and distributed operation with different optimized and heuristic power control methods. We consider $L = 100$, $N = 4$, $K = 40$, $\tau_p = 10$, and spatially correlated Rayleigh fading with ASD $\sigma_\varphi = \sigma_\theta = 15^\circ$.

When it comes to max-min fairness power control, we notice that the lower tail of the corresponding CDF curve begins at the highest number among all the studied schemes. This is expected since the performance of the most unfortunate UE in the entire network is maximized. However, by zooming in at the lower tail of the CDF curves, we can see that max-min fairness only results in a marginal improvement compared to the sum SE curve, and the price to pay is substantially smaller SEs for the vast majority of UEs. In fact, the fractional power control method with $\upsilon = -0.5$ achieves nearly the same SE for the most unfortunate UEs, while not sacrificing the SE for the other UEs to the same extent. When we switch to fractional power control with $\upsilon = 0.5$, we notice that its behavior is similar to sum SE maximization for the UEs with good channel conditions, but not as good for the UEs with weaker channel conditions.

In summary, maximizing the sum SE will result in a CDF curve that is almost entirely to the right of the competing schemes. Hence, it is generally the preferred choice and this power optimization can be implemented in a scalable manner by letting all UEs transmit with full power. This is why we utilized this scheme for performance comparisons in Section 5.4 on p. 331. If we want a higher level of fairness, in terms of increasing the SE for the most unfortunate UEs, then a fractional power control method with $\upsilon = -0.5$ is a good scalable option. It must be thus clear that the max-min fairness problem focuses on an extreme type of fairness that only helps one UE in the network, at the expense of everyone else. This effect is particularly evident in a large cell-free network where most UEs are barely affecting the most unfortunate UE but anyway are forced to cut down on their transmit powers.

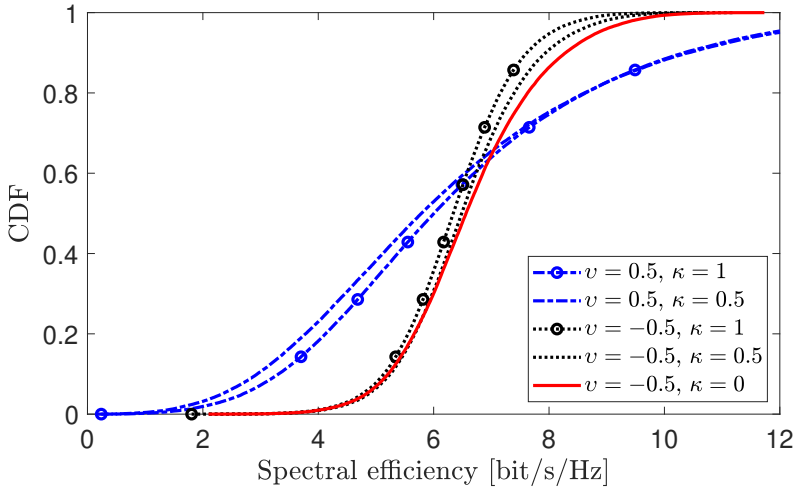
7.3.2 Comparison of Downlink Power Allocation Schemes

We continue with the comparison of optimized and scalable power allocation schemes by considering the downlink. We begin by studying the centralized operation where we use Algorithm 7.3 for max-min fairness power allocation and Algorithm 7.4 for sum SE maximization. We also consider the network-wide equal power allocation scheme in (7.38) and the fractional power allocation method in (7.43) as scalable benchmarks. Recall that the latter method was considered in the simulations in Sec-

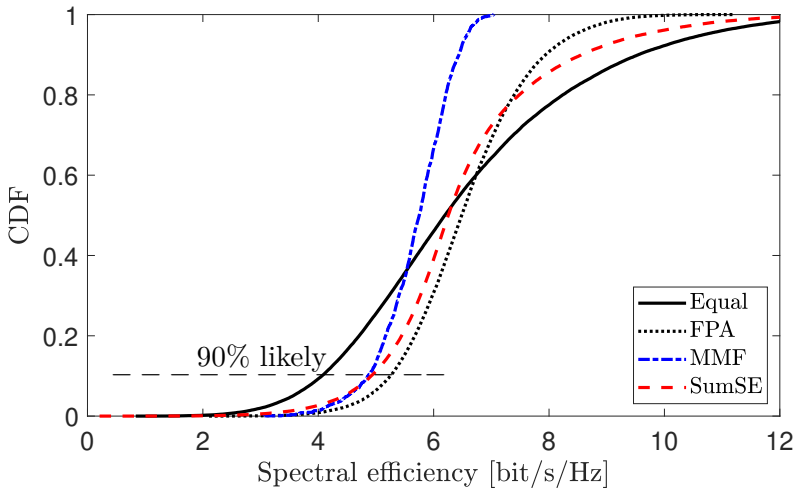
tion 6.3 on p. 376. These methods are denoted by "MMF", "SumSE", "Equal", and "FPA" respectively.

Figure 7.2 shows the CDFs of the SE per UE for the P-MMSE precoding scheme, which is scalable. In Figure 7.2(a), we consider the effect of tuning parameters in the FPA scheme, i.e., the exponents υ and κ in (7.43). When υ is positive, more emphasis is put on the UEs with higher total channel gains whereas it is the other way around for a negative υ. Hence, $\upsilon < 0$ mimics the max-min fairness optimization. There is an additional tuning parameter κ that reshapes the power allocation to account for the unequal contributions that the serving APs provide to the centralized precoding vector. From the figure, we notice that the effect of υ on the SE spread is more substantial compared to κ. When $\upsilon = 0.5$, the more fortunate UEs attain higher SE compared to the case of $\upsilon = -0.5$. On the other hand, $\upsilon = -0.5$ is more preferable to provide more uniform service quality. Another observation is that the selection of κ affects the SE differently for the different values of υ. When $\upsilon = 0.5$, it is better to use a higher κ while the reverse is true when $\upsilon = -0.5$.

In Figure 7.2(b), we compare the CDFs of the SE per UE for the optimized and scalable power allocation methods described above. We use the fractional power allocation from Figure 7.2(a) with $\upsilon = -0.5$ and $\kappa = 0.5$, as we did in the results of Section 6.3 on p. 376. This is an option that gives the best performance in the lower tail. We notice that the max-min fairness-based power allocation begins at the largest value and has the smallest difference between its lower and upper tails, which are two properties that are expected from this type of power allocation. However, as in the uplink case studied in Figure 7.1, this type of fairness results in a significant performance drop for all UEs except the most unfortunate ones. In fact, with sum SE maximization or fractional power allocation, it is possible to serve nearly all of the UEs with higher SE and only a few with reduced SE in comparison with the max-min fairness solution. Hence, we conclude that the sum SE metric is preferable over the max-min fairness metric both when it comes to average performance and fairness. Fractional power allocation provides a good heuristic tradeoff. A key observation is that most of the UEs attain higher SE with the scalable fractional power allocation



(a) Fractional power allocation with different parameter values.



(b) Centralized downlink operation with several power allocation schemes.

Figure 7.2: CDF of the downlink SE per UE in the centralized operation with different optimized and heuristic power allocation methods. We consider P-MMSE precoding, $L = 100$, $N = 4$, $K = 40$, $\tau_p = 10$, and spatially correlated Rayleigh fading with ASD $\sigma_\varphi = \sigma_\theta = 15°$.

than with sum SE maximization and the 90% likely SE (where the CDF curve is 0.1) is better than with max-min fairness. Hence, this scheme strikes a good tradeoff between fairness and sum SE.

Different from the uplink case, sum SE maximization generally performs better than the network-wide equal power allocation. The CDF for the network-wide equal power allocation is the rightmost one in the upper tails, but there is a substantial gap to the other curves for all other UEs. For example, the 90% likely SE is around 1.5 bit/s/Hz higher when considering fractional power allocation.

In Figure 7.3, we consider the distributed operation with the scalable LP-MMSE precoding. When it comes to the optimized power allocation schemes, we consider the max-min fairness and sum SE maximization algorithms in Algorithm 7.5 and Algorithm 7.6, respectively. As scalable alternatives, we consider per-AP equal power allocation with $\rho_{kl} = \frac{\rho_{\max}}{|\mathcal{D}_l|}$ and the heuristic scheme in (7.47). We denote the latter one by "FPA" in analogy with the conceptually similar fractional uplink power control approach in (7.35) and we consider two different exponents: $\upsilon = 0.5$ and $\upsilon = -0.5$. Note the case with $\upsilon = 0.5$ leads to that each AP allocates more power to the UEs with better channel conditions whereas $\upsilon = -0.5$ leads to the opposite effect.

Most of the UEs benefits from sum SE maximization in Figure 7.3, however, max-min fairness can significantly improve the SE achieved by the most unfortunate UEs in the network. The UEs with the 15% worst channel conditions (below 85% likely SE) achieve higher SE with max-min fairness than with sum SE maximization, which is a much higher percentage than we have observed for uplink power control and centralized downlink power allocation. We further notice that using the exponent $\upsilon = -0.5$ in the heuristic scheme in (7.47) is not an efficient approach, since the CDF curve is to the left of all the other curves. On the other hand, the case with $\upsilon = 0.5$, which we utilized for performance comparison in Section 6.3 on p. 376, is much better than the per-AP equal power allocation. The reason is that less interference is created when each UE obtains most of its power from the AP that it has the best channel to, compared to when all the serving APs are transmitting with equal power. The gap between the case of $\upsilon = 0.5$ and the sum SE maximization scheme is relatively small.

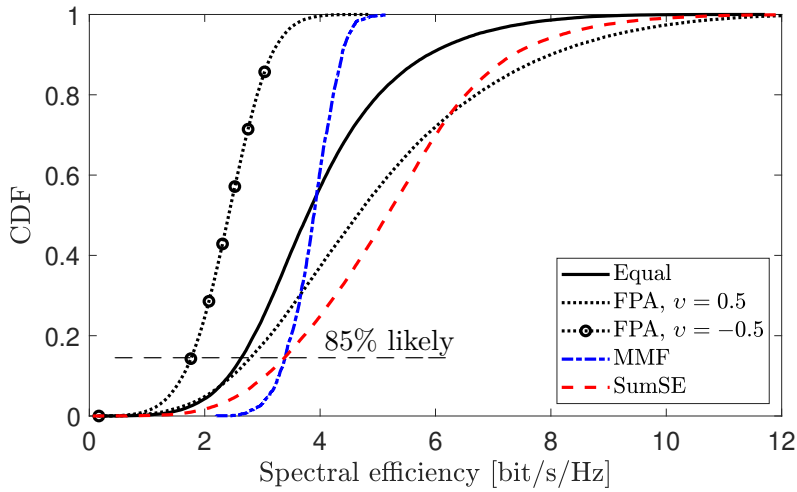


Figure 7.3: CDF of the downlink SE per UE in the distributed operation with different optimized and heuristic power allocation methods. Scalable LP-MMSE precoding is used. We consider $L = 100$, $N = 4$, $K = 40$, $\tau_p = 10$, and spatially correlated Rayleigh fading with ASD $\sigma_\varphi = \sigma_\theta = 15°$.

We conclude that we can choose between sum SE maximization and max-min fairness in the distributed downlink operation depending on the tradeoff between average SE and fairness that we want to obtain. The heuristic scheme in (7.47) is a reasonable approximation of the sum SE maximization case but further research on scalable power allocation schemes is needed.

Remark 7.5 (Machine learning for efficient power optimization)**.** Most of the signal processing problems that have been considered in previous sections of this monograph have known optimal solutions, as well as computationally scalable approximations that seem to perform very well. The presented algorithms for channel estimation, precoding, and combining in centralized and distributed operations are examples of successful man-made algorithms. In contrast, the downlink power allocation and uplink power control problems are still not solved from a practical perspective. The optimization algorithms presented earlier in this section have computational complexities that grow polynomially with the number of UEs, which is an algorithmic deficiency [**Bjornson2020b**].

The man-made heuristic schemes presented earlier are not bad but there is a substantial gap to the optimal benchmark algorithms, partially because the heuristic schemes have access to less information and partially because it is inherently hard to design heuristic schemes that work well everywhere in the network. Machine learning might be the tool that is needed to overcome these deficiencies and some first steps are taken in [**Bashar2020c**], [**Chakraborty2019a**], [**Nikbakht2020b**], [**Yan2020a**], [**Zhao2020a**]. A deep learning approach can potentially find shortcuts in existing centralized optimization algorithms, thereby lowering the computational complexity while only suffering from small performance penalties. Other utility functions than those considered in this monograph can also be considered. When it comes to distributed power allocation schemes (or other schemes that have access to less information than the benchmark algorithms), there is no need to have the same policy at every AP but the specific characteristics of the local propagation environment around each AP can be learned to enhance the performance. The benefit from this can hopefully outweigh the drawback of only having access to local information at the AP. This research direction is only in its infancy.

7.4 Pilot Assignment

When multiple UEs are simultaneously using the same pilot signal, there will be pilot contamination that reduces the estimation quality of the pilot-sharing UEs' channels and increases the mutual interference in the data transmission phase. This issue is not unique to cell-free networks but appears also in cellular networks. Pilot contamination has received particular attention in the literature of Cellular Massive MIMO [**Adhikary2017a**], [**ashikhmin2012pilot**], [**BjornsonHS17**], [**Jose2011b**], [**Marzetta2010a**], [**Mueller2014b**], [**Rusek2013a**], [**Sanguinetti**
[**Yin2016a**], mainly because the resulting additional interference grows with the number of AP antennas. The easiest approach to limit the pilot contamination issue is by selecting the pilot-sharing UEs in a judicious manner. To this end, the standard approach in cellular networks is to associate each cell with a predetermined subset of the pilots. This subset can, for instance, be selected so that neighboring cells are using different subsets. Each AP can then assign the pilots arbitrarily to the

UEs that reside in its cell. Hence, the cellular structure is exploited to make the pilot assignment relatively straightforward to implement in a distributed fashion. Such methods cannot be used in a cell-free network and, therefore, there is a need for developing new algorithms for pilot assignment.

The general behaviors of pilot contamination in cell-free networks were exemplified in Section 4.3 on p. 274. A key insight was that we want to avoid that two UEs that are close to the same set of APs are assigned to the same pilot. A naive approach to pilot assignment is random allocation [**Interdonato2018**], where we generate a random integer from 1 to τ_p for every UE and assign this UE to the pilot with the matching index. A benefit of this approach is that it requires no coordination between different UEs or APs, while the main drawback is that the probability that a UE will use the same pilot as its geographically closest neighbor is $1/\tau_p$. This is the worst-case situation that should be avoided by using a more structured pilot assignment algorithm.

7.4.1 Utility-Based Pilot Assignment

The key to structured pilot assignment is to define a utility function $f(t_1, \ldots, t_K)$ that represents the goal of the pilot assignment and takes the indices $t_1, \ldots, t_K \in \{1, \ldots, \tau_p\}$ of the pilots assigned to the different UEs as input. The utility can, for example, be defined so that it is maximized when the extra interference caused by pilot contamination is minimized [**Ngo2017b**] or when the sum SE is maximized [**Liu2020a**]. In any case, we have the following general problem formulation:

$$\begin{aligned} \underset{t_1,\ldots,t_K}{\text{maximize}} \quad & f(t_1, \ldots, t_K) \\ \text{subject to} \quad & t_k \in \{1, \ldots, \tau_p\}, \quad k = 1, \ldots, K. \end{aligned} \tag{7.48}$$

This is a combinatorial problem and we can, thus, find the optimal pilot assignment for a given utility function by an exhaustive search over all τ_p^K possible pilot assignment.[2] The complexity grows exponentially with

[2] This number can be slightly reduced by exploiting that it only matters which UEs use the same pilot and not what the index of their pilot is.

K, which makes it challenging to implement an exhaustive search in a practical network with many UEs, even if it would only be used for offline benchmarking. We need to be able to solve the pilot assignment problem regularly since the utility will depend on which UEs are currently active in the network.

Instead of finding the global optimum to (7.48), we can design an algorithm that finds a decent suboptimal solution. One approach is to design a greedy algorithm that optimizes the utility with respect to one UE at a time. The algorithm can either consider each UE once or iterate until convergence. In small networks where most UEs can be assigned to unique pilots, the algorithm in [**Sabbagh2018a**] can be used to find suitable UE pairs that can reuse pilots. The greedy algorithm in [**Ngo2017b**] is first assigning pilots randomly to the UEs, followed by an iterative procedure where each UE determines if the extra interference caused by pilot contamination can be reduced by switching to another pilot. A variation of that greedy algorithm is proposed in [**Zhang2018b**] by also making use of the geographical locations of the UEs. A UE clustering algorithm is proposed in [**Attarifar2018a**] to dynamically divide the network into geographical clusters in which each pilot is only used once. This principle resembles the cellular approach to the pilot assignment problem but makes use of the actual UE locations. A refined clustering algorithm that utilizes the physical distances between APs and UEs is proposed in [**Chen2020a**]. Yet another clustering algorithm is proposed in [**Femenias2019a**] but it is using the inner products between the vectors $[\beta_{k1} \ \ldots \ \beta_{kL}]^{\mathrm{T}}$ of different UEs as a similarity metric instead of location-based parameters.

Several of the aforementioned algorithms make direct use of the observation that two closely located UEs should not be assigned to the same pilot. However, it is important to bear in mind that it is the channel gains that matter when determining the pilot contamination and even if these are strongly correlated with the UE locations, there can be large variations, which are modeled by shadow fading.

The mathematical literature contains many combinatorial algorithms that can potentially be applied to pilot assignment. Tabu-search is an algorithm that was used in [**Liu2020a**] for SE-maximizing pilot

assignment. The main principle is to iteratively explore small variations on the current assignment and select a preferable one. A list of previously selected solutions is kept to avoid going back to them. The Hungarian algorithm was utilized in [**Buzzi2020a**] and tuned to optimize different utilities.

It is generally hard to make a fair comparison of pilot assignment algorithms because they might optimize different utilities, they might get stuck in different local optima in different setups, and their computational complexity can be widely different. However, one can conclude that network-wide algorithms are preferably implemented at the CPU and the complexity will grow at least linearly with K, which makes their implementation unsuitable for large networks. Our experimental experience is that the most important aspect of a pilot assignment in cell-free networks is to avoid the worst assignments where closely located UEs use the same pilot. This is fairly easy to achieve in a network with $L \gg K$ since each pilot will be reused quite sparsely in the network, as seen from the APs' perspective. The coherent interference that made pilot contamination a major concern in the Cellular Massive MIMO literature is likely not a major issue in cell-free networks, where there also are many antennas but each UE is only served by a small subset.

Scalable Pilot Assignment

If a pilot assignment algorithm should be scalable, it probably must be implemented by a local interaction between a UE and its neighboring APs. Any algorithm that attempts to maximize a network-wide utility and exploits network-wide information will require an immense complexity to evaluate the utility. The pilot assignment algorithm presented in Algorithm 4.1 on p. 289 is an attempt to design such a scalable algorithm. It originates from [**Bjornson2020a**] where the main idea was to connect the pilot assignment for a given UE to how it accesses the network (as it is usually done in cellular networks). When the UE is becoming active, it selects a neighboring AP and this AP locally determines which pilot is the most appropriate for the UE to use. More precisely, it computes/measures the amount of pilot interference at each of the pilots and assigns the UE to the pilot with the least interference. This likely corresponds to the pilot where the pilot-sharing UEs are furthest away from the AP, which makes intuitive sense: we want each

pilot to be reused as sparsely as possible in space. The algorithm is by no means optimal but has been used throughout this monograph and has provided good results.

Further research on scalable pilot assignment is certainly needed and since it is a type of clustering problem, machine learning might be a suitable tool for developing efficient algorithms.

7.5 Selection of Dynamic Cooperation Clusters

The DCC framework is restricting which APs are allowed to serve a given UE in the uplink and downlink. It is unavoidable that this leads to lower SE than in a network where any AP can serve any UE, since we have fewer degrees-of-freedom when optimizing the system. However, there are strong reasons for introducing these restrictions. As stressed earlier in this monograph, scalability in terms of computational complexity and fronthaul signaling is one reason. Another reason is to reduce the energy consumption [**Ngo2018a**] or to limit the delay spread of the downlink signals, which increases with the distance between the UE and the serving AP that is furthest away [**Bjornson2013d**], [**Zhang2008a**]. Intuitively, the SE loss can be kept small if each UE is served by all the APs within its area of influence, but the question is what this means in practice.

Some general guidelines for DCC selection were provided in [**Bjornson2013d**]. Firstly, each UE should have a "master AP" that it is anchored to. This AP is required to serve the UE, to guarantee a non-zero SE, while service from other APs is provisioned based on availability (e.g., if these APs are not overloaded with serving other UEs). Secondly, the UEs should be selected from a user-centric perspective, as emphasized throughout this monograph. Thirdly, different UEs can be assigned to different numbers of APs depending on the local propagation conditions. For example, a UE that has a very good channel to one AP might only need to be served by that AP, or a few more, while a UE that is in between many APs or is subject to much interference will require multiple APs to boost the SNR and/or suppress interference. Finally, [**Bjornson2013d**] suggests that one should use the channel quality rather than the geographical locations when measuring proximity to

different APs and determining which APs should serve the UE.

There are many possible user-centric clustering algorithms. While it is possible to design network-wide algorithms where all UEs are jointly considered, this will be an unscalable solution where the complexity grows with K. Hence, we will concentrate on approaches that consider one UE at a time.

Recall that $\mathcal{M}_k \subset \{1, \ldots, L\}$ is the subset of APs that serves UE k. One approach for selecting the subset is

$$\mathcal{M}_k = \{l = 1, \ldots, L : \beta_{kl} \geq \Delta\} \tag{7.49}$$

where $\Delta > 0$ is a constant parameter [**Bjornson2011a**]. This means that the UE is served by all APs that have a large-scale fading coefficient β_{kl} that is above a threshold specified by Δ. A variation of this approach is to predetermine that each UE k should be served by a certain number of APs. These APs are then selected as the ones with the largest values on β_{kl} [**Buzzi2017a**]. This corresponds to tuning the threshold Δ in (7.49) for each UE to get the predetermined number of serving APs. Alternatively, the number of serving APs can be selected on a per-UE basis to make sure that

$$\frac{\sum_{l \in \mathcal{M}_k} \beta_{kl}}{\sum_{l=1}^{L} \beta_{kl}} \geq \delta \tag{7.50}$$

where $\delta > 0$ is a threshold representing the fraction of the total received power that UE k would receive in the downlink [**Ngo2018a**]. For example, if $\delta = 0.95$, then the UE will be served by the subset of APs that has the strongest channels and which contributes to more than 95% of the total received power at the UE.

The aforementioned approaches are user-centric and make intuitive sense, but are not taking into account how many UEs a given AP can practically serve. There are two important types of limitations. The first limitation is scalability. Each AP has a limited processing power so it can only manage a limited number of UEs in a distributed operation and only transmit and receive data related to a limited number of UEs over its fronthaul link. The second limitation is pilot contamination. It is only reasonable for an AP to serve one UE per pilot [**Bjornson2020a**], [**Sabbagh2018a**], because otherwise the weaker of the pilot-sharing

UEs will be subject to strong interference. An exception to this rule is if the pilot-sharing UEs have very different spatial correlation matrices so that the AP can separate their channels in the estimation phase based on that information.

If we need to select the clusters $\mathcal{M}_1, \ldots, \mathcal{M}_K$ to comply with the per-AP constraint of the type $|\mathcal{D}_l| \leq \tau_p$ (where $\mathcal{D}_l$ is the set of UEs served by AP l), we cannot apply any of the approaches mentioned above. More precisely, we cannot make our selection from a purely user-centric perspective but must take the limitations caused by the network architecture into account. One scalable solution is to let the pilot assignment algorithm determine the clusters [**Bjornson2020a**]. For example, each AP can serve (up to) one UE per pilot, namely the UE that has the largest large-scale fading coefficient β_{kl} among those that use the pilot. This approach is scalable but the performance will rely on the fact that the pilot assignment problem has been solved appropriately. This was the approach taken in Algorithm 4.1 on p. 289 and followed throughout this monograph. We observed that the SE loss compared to serving all UEs by all APs is small, at least in the running example with $L \gg K$. However, further research on scalable cooperation cluster formation is needed, particularly for challenging cases with many UEs in certain areas of the network.

7.6 Implementation Constraints

Scalability has been the main focus when we developed the foundations of User-centric Cell-free Massive MIMO in this monograph. The basic definition of scalability, according to Definition 2.2 on p. 217, is that the signal processing complexity and fronthaul signaling per AP and UE remain finite as $K \to \infty$. This implies that once we have deployed an AP and connected it to its CPU, we will not have to update the local processors and fronthaul infrastructure as we are deploying additional APs, increase the coverage area, or serve a larger number of UEs. However, there are further implementation constraints to consider when building practical networks. We will briefly mention a few of these constraints in this section to stress that further research is needed in these directions.

7.6.1 Low-Cost Components

To enable a ubiquitous deployment of APs, it is important to use compact low-cost components, which might be based on UE-grade chipsets rather than conventional hardware for cellular infrastructure. Practical transceiver components are subject to hardware impairments of different kinds [**Schenk2008a**], including non-linearities in power amplifiers, mismatches in mixers, finite-resolution analog-to-digital and digital-to-analog conversion, and phase noise in local oscillators. These effects lead to signal distortion which in some cases can be modeled as a signal power loss plus an uncorrelated additive distortion term, using the Bussgang decomposition [**Bussgang1952a**], [**Demir2020a**]. The impact of such distortion on cell-free networks has been analyzed in [**Masoumi2020a**], [**Zhang2018a**], [**Zheng2020a**] for general hardware impairments, while the special case of quantization distortion in each analog-to-digital converter (ADC) was considered in [**Hu2019a**]. The obtained SE expressions can be utilized to get a better sense of the practically achievable SE and to optimize the transmit power based on these expressions. The analysis in the aforementioned papers follows the methodology that was developed in the Cellular Massive MIMO literature [**massivemimobook**] or approximations thereof. These models are relatively simple and future research should consider more detailed models.

7.6.2 Quantization of Fronthaul Signaling

We have focused on limiting the number of scalar numbers that need to be transmitted over the fronthaul links per coherence block. However, in practice, these signals must also be quantized before being transmitted over the fronthaul. While the quantization distortion caused by hardware impairments (such as finite-resolution ADCs) is applied on a sample-by-sample basis and largely uncontrollable, the fronthaul compression can be optimized using appropriately designed compression formats that are applied to a block of signal samples. The signal distortion can potentially be limited by making use of rate-distortion theory. Some papers in this direction are [**Bashar2020c**], [**Bashar2019b**], [**Boroujerdi2019a**],

[**Femenias2019a**], [**Maryopi2019a**], [**Masoumi2020a**], [**Parida2018a**]. An interesting consideration that appears when having fronthaul compression is that it matters where certain computations are carried out. If a processing task, such as channel estimation or signal detection, is carried out locally at the AP, the result will be more accurate than if the received signals are first compressed and sent to the CPU and then processed in the same way. For example, in a centralized operation, we can choose between estimate-and-quantify and quantify-and-estimate protocols [**Bashar2019b**], [**Maryopi2019a**]. Similar to the case of hardware impairments, SE expressions that are developed by taking the fronthaul compression into account can be used for optimizing the transmit power or other resource allocation tasks.

7.6.3 Radio Stripes & Other Constrained Fronthaul Architectures

The structure of the fronthaul might also be constrained in the sense that each AP might not have a dedicated fronthaul link but a shared connection with a subset of other APs. For example, in the radio stripes architecture [**Interdonato2018**], the APs are integrated into fronthaul cables that are connected to a CPU at one end, which also provides a power supply. A fronthaul signal from the outmost AP needs to travel via all the other APs before reaching the CPU. The benefit of this architecture is that amount of cabling can be greatly reduced. In a deployment with a particular fronthaul architecture, the structure can be utilized to optimize the fronthaul signaling and signal processing. For example, the APs can send signals to their neighboring APs which enables a co-processing and interference mitigation [**Shaik2020a**]. MMSE combining can be implemented in a sequential fashion [**Shaik2021a**].

7.6.4 Synchronization of APs

Proper synchronization of the distributed APs is necessary for coherent uplink and downlink transmission. The APs must not be phase-synchronized since this is momentarily achieved in every coherence block through the channel estimation, which estimates the combined effect of the propagation channels and the phase-shifts induced by the hardware. However, the cooperating APs must be synchronized in time

and frequency; see [**Jeong2020a**] for a recent review of how to achieve that in Cellular Massive MIMO systems and [**Etzlinger2018a**] for a more general review. Perfect synchronization has been assumed in this monograph, but the actual implementation can be challenging since the APs are distributed. Some initial algorithms for cell-free networks are described in [**Cheng2013a**], [**Interdonato2018**], but further work is needed on this topic. The user-centric cooperation clusters might relax the synchronization requirements since only the geographically closest APs must be well synchronized, while more distant APs that are serving non-overlapping sets of UEs do not require the same level of synchronization.

7.7 Summary of the Key Points in Section 7

- UEs in cell-free networks have conflicting performance goals due to the interference and shared power budgets. Power allocation can be used to find a tradeoff between them.
- Any algorithm that jointly optimizes the transmit powers in uplink or downlink to maximize a network-wide utility function becomes unscalable as $K \to \infty$ since there are K SINRs to optimize and at least K optimization variables.
- To obtain a practically implementable power allocation that is scalable, some kind of distributed heuristic scheme is needed, where each AP or UE makes a local decision based on the channel knowledge that it can acquire locally. Fractional power control is a classical heuristic that can be also used in cell-free networks.
- Sum SE maximization is often the preferable optimization criterion since it attains high SEs for the UEs with good channel conditions and satisfactorily SEs for UEs with bad channel conditions. An alternative metric is the max-min fairness, which provides (slightly) higher SE for the UE with the worst conditions but a substantial SE loss for most other UEs.
- In the uplink, the sum SE is maximized when all UEs transmit with full power. Hence, there is no need to run a network-wide optimization problem in the centralized and distributed uplink operation. To emphasize fairness, fractional power control can be utilized with a negative exponent.
- In the centralized downlink operation, fractional power allocation with a negative exponent provides higher SE to the most of the UEs than the sum SE maximization and it also achieves a good balance between fairness and sum SE.

- In the distributed downlink simulations, the fractional power allocation with a positive exponent performs much better than per-AP equal power allocation. There is a gap to the sum SE maximization-based scheme but it is relatively small.
- The pilot contamination effect can be reduced by assigning pilots to UEs to limit the interference. The optimal pilot assignment is a combinatorial problem whose complexity grows exponentially with K. However, greedy algorithms that operate at one UE at a time can be developed. They are scalable and guarantee reasonably good performance.
- An unscalable network where all APs serve all UEs provides the highest SEs, but the dynamic cooperation clusters can be selected to achieve a scalable operation with a small performance loss. Any user-centric clustering algorithm where all UEs are jointly considered is unscalable as $K \to \infty$. If properly designed, approaches that consider one UE at a time can provide good performance while being scalable.
- Scalability is not the only requirement for building practical networks. The need for enabling a ubiquitous deployment of APs makes it preferable to use compact low-cost components, whose hardware imperfections must be taken into account in the analysis and design. The fronthaul signal compression is another limiting factor. The structure of the fronthaul might also impose limitations since some APs might share a connection with a subset of other APs. Finally, proper synchronization of the distributed APs is necessary for coherent uplink and downlink transmission. These are topics (among many others) that require further investigation.

Acknowledgements

We would first like to thank our students and collaborators in the areas related to this monograph. Without the results, encouragements, and insights obtained through our joint research during the last decade, it wouldn't have been possible to write this monograph. We are grateful for the constructive feedback from the reviewers, which helped us to focus our final editing efforts at the right places. In particular, we would like to thank Angel Lozano, Jiayi Zhang, Mahmoud Zaher, and Yasaman Khorsandmanesh for giving detailed comments.

Özlem Tuğfe Demir and Emil Björnson have been supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. Emil Björnson has also been supported by the Excellence Center at Linköping – Lund in Information Technology (ELLIIT), the Center for Industrial Information Technology (CENIIT), the Swedish Research Council, and the Swedish Foundation for Strategic Research. Luca Sanguinetti has been partially supported by the University of Pisa under the PRA Research Project CONCEPT, and by the Italian Ministry of Education and Research (MIUR) in the framework of the CrossLab project (Departments of Excellence).

Appendices

A

Notation and Abbreviations

Mathematical Notation

Upper-case boldface letters are used to denote matrices (e.g., $\mathbf{X}, \mathbf{Y}$), while column vectors are denoted with lower-case boldface letters (e.g., $\mathbf{x}, \mathbf{y}$). Scalars are denoted by lower/upper-case italic letters (e.g., x, y, X, Y) and sets by calligraphic letters (e.g., $\mathcal{X}, \mathcal{Y}$).

The following mathematical notations are used:

$\mathbb{C}^{N\times M}$	The set of complex-valued $N \times M$ matrices
$\mathbb{R}^{N\times M}$	The set of real-valued $N \times M$ matrices
$\mathbb{C}^N, \mathbb{R}^N$	Short forms of $\mathbb{C}^{N\times 1}$ and $\mathbb{R}^{N\times 1}$ for vectors
$\mathbb{R}^N_{\geq 0}$	The set of non-negative members of $\mathbb{R}^N$
$x \in \mathcal{S}$	x is a member of the set $\mathcal{S}$
$x \notin \mathcal{S}$	x is not a member of the set $\mathcal{S}$
$\{x \in \mathcal{S} : P\}$	The subset of $\mathcal{S}$ containing all members that satisfy a property P
$[\mathbf{x}]_i$	The ith element of a vector $\mathbf{x}$
$[\mathbf{X}]_{ij}$	The (i,j)th element of a matrix $\mathbf{X}$
$[\mathbf{X}]_{:,1}$	The first column of a matrix $\mathbf{X}$
$\mathrm{diag}(\cdot)$	$\mathrm{diag}(x_1, \ldots, x_N)$ is a diagonal matrix with

	the scalars $x_1, \ldots, x_N$ on the diagonal,
	$\mathrm{diag}(\mathbf{X}_1, \ldots, \mathbf{X}_N)$ is a block diagonal matrix with
	the matrices $\mathbf{X}_1, \ldots, \mathbf{X}_N$ on the diagonal
$\mathbf{X}^\star$	The complex conjugate of $\mathbf{X}$
$\mathbf{X}^{\mathrm{T}}$	The transpose of $\mathbf{X}$
$\mathbf{X}^{\mathrm{H}}$	The conjugate transpose of $\mathbf{X}$
$\mathbf{X}^{-1}$	The inverse of a square matrix $\mathbf{X}$
$\mathbf{X}^{\frac{1}{2}}$	The square-root of a square matrix $\mathbf{X}$
$\Re(x)$	Real part of x
j	The imaginary unit
$\lvert x \rvert$	Absolute value (or magnitude) of a scalar variable x
e	Euler's number ($e \approx 2.718281828$)
$\log_a(x)$	Logarithm of x using the base $a > 0$
$\sin(x)$	The sine function of x
$\cos(x)$	The cosine function of x
$\mathrm{tr}(\mathbf{X})$	Trace of a square matrix $\mathbf{X}$
$\mathcal{N}(\mathbf{x}, \mathbf{R})$	The real Gaussian distribution
	with mean $\mathbf{x}$ and covariance matrix $\mathbf{R}$
$\mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathbf{R})$	The circularly symmetric complex Gaussian
	distribution with zero mean and correlation matrix $\mathbf{R}$,
	where circular symmetry means that if $\mathbf{y} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathbf{R})$
	then $e^{\mathrm{j}\phi}\mathbf{y} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathbf{R})$ for any given ϕ
$\mathbb{E}\{x\}$	The expectation of a random variable x
$\mathbb{V}\{x\}$	The variance of a random variable x
$\lVert \mathbf{x} \rVert$	The L_2-norm $\lVert \mathbf{x} \rVert = \sqrt{\sum_i \lvert [\mathbf{x}]_i \rvert^2}$ of a vector $\mathbf{x}$
$\mathbf{I}_M$	The $M \times M$ identity matrix
$\mathbf{1}_N$	The $N \times 1$ matrix (i.e., vector) with only ones
$\mathbf{1}_{N \times M}$	The $N \times M$ matrix with only ones
$\mathbf{0}_M$	The $M \times 1$ matrix (i.e., vector) with only zeros
$\mathbf{0}_{N \times M}$	The $N \times M$ matrix with only zeros

Abbreviations

The following acronyms and abbreviations are used in this monograph:

ADC	Analog-to-Digital Converter
AP	Access Point
ASD	Angular Standard Deviation
AWGN	Additive White Gaussian Noise
CDF	Cumulative Distribution Function
CDMA	Code-Division Multiple Access
CoMP	Coordinated Multipoint
C-RAN	Cloud Radio Access Network
CPU	Central Processing Unit
CSI	Channel State Information
DCC	Dynamic Cooperation Clustering
FDD	Frequency-Division Duplex
FIR	Finite Impulse Response
FPA	Fractional Power Allocation
FPC	Fractional Power Control
i.i.d.	Independent and Identically Distributed
JP	Joint Processing
L-MMSE	Local MMSE
LoS	Line-of-Sight
LP-MMSE	Local P-MMSE
LSFD	Large-Scale Fading Decoding
MIMO	Multiple-Input Multiple-Output
MMF	Max-Min Fairness
MMSE	Minimum Mean-Squared Error
MR	Maximum Ratio
MSE	Mean-Squared Error
n-opt	Nearly Optimal
NLoS	Non-Line-of-Sight
NMSE	Normalized MSE
OFDM	Orthogonal Frequency-Division Multiplexing
opt	Optimal
P-RZF	Partial Regularized Zero-Forcing
P-MMSE	Partial MMSE
PDF	Probability Density Function
RF	Radio Frequency
SE	Spectral Efficiency

SINR	Signal-to-Interference-plus-Noise Ratio
SIR	Signal-to-Interference Ratio
SISO	Single-Input Single-Output
SNR	Signal-to-Noise Ratio
TDD	Time-Division Duplex
UatF	Use-and-then-Forget
UE	User Equipment
ULA	Uniform Linear Array
ZF	Zero-Forcing

B

Useful Lemmas

This appendix contains a few classical results related to matrices, which are utilized to prove the results in other parts of the monograph.

Lemma B.1 (Matrix inversion lemma)**.** Consider the matrices $\mathbf{A} \in \mathbb{C}^{N_1 \times N_1}$, $\mathbf{B} \in \mathbb{C}^{N_1 \times N_2}$, $\mathbf{C} \in \mathbb{C}^{N_2 \times N_2}$, and $\mathbf{D} \in \mathbb{C}^{N_2 \times N_1}$. The following identity holds if all the involved inverses exist:

$$(\mathbf{A} + \mathbf{BCD})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{DA}^{-1}\mathbf{B} + \mathbf{C}^{-1})^{-1}\mathbf{DA}^{-1}. \tag{B.1}$$

Lemma B.2. For matrices $\mathbf{A} \in \mathbb{C}^{N_1 \times N_2}$ and $\mathbf{B} \in \mathbb{C}^{N_2 \times N_1}$, it holds that

$$(\mathbf{I}_{N_1} + \mathbf{AB})^{-1}\,\mathbf{A} = \mathbf{A}\,(\mathbf{I}_{N_2} + \mathbf{BA})^{-1} \tag{B.2}$$

$$\operatorname{tr}(\mathbf{AB}) = \operatorname{tr}(\mathbf{BA}). \tag{B.3}$$

Lemma B.3. For the non-zero positive semi-definite matrix $\mathbf{A} \in \mathbb{C}^{N \times N}$ and positive definite matrix $\mathbf{B} \in \mathbb{C}^{N \times N}$, their inner product is strictly positive:

$$\operatorname{tr}(\mathbf{AB}) > 0. \tag{B.4}$$

Proof. Consider the eigendecomposition of $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathrm{H}}$ where $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_N)$ with $\lambda_n \geq 0$, for $n = 1, \ldots, N$. Let $\mathbf{u}_n$ denote the nth column of the eigenvalue matrix $\mathbf{U}$. Then, we have

$$\operatorname{tr}(\mathbf{AB}) = \sum_{n=1}^{N} \lambda_n \mathbf{u}_n^{\mathrm{H}} \mathbf{B} \mathbf{u}_n \tag{B.5}$$

that is strictly greater than zero since $\mathbf{B}$ is a positive definite matrix and at least one eigenvalue of $\mathbf{A}$ is positive. □

Lemma B.4. For the positive semi-definite matrices $\mathbf{A} \in \mathbb{C}^{N\times N}$, $\mathbf{C} \in \mathbb{C}^{N\times N}$, and positive definite matrix $\mathbf{B} \in \mathbb{C}^{N\times N}$, the following inequality holds:

$$\operatorname{tr}\left(\mathbf{A}\left(\mathbf{B}+\mathbf{C}\right)^{-1}\right) \leq \operatorname{tr}\left(\mathbf{A}\mathbf{B}^{-1}\right) \tag{B.6}$$

where the equality holds only when $\mathbf{C}\mathbf{B}^{-1}\mathbf{A} = \mathbf{0}_{N\times N}$.

Proof. Consider the eigendecomposition of $\mathbf{C} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{H}} = \mathbf{U}_1\boldsymbol{\Lambda}_1\mathbf{U}_1^{\mathrm{H}}$, where $\mathbf{U} \in \mathbb{C}^{N\times N}$ is the unitary matrix of eigenvectors and $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_N)$ with the eigenvalues $\lambda_n \geq 0$, for $n = 1, \ldots, N$. $\mathbf{U}_1 \in \mathbb{C}^{N\times r}$ and $\boldsymbol{\Lambda}_1 \in \mathbb{C}^{r\times r}$ are the partitions of $\mathbf{U}$ and $\boldsymbol{\Lambda}$, respectively corresponding to the positive eigenvalues. Applying the matrix inversion lemma (Lemma B.1) to the inverse of $\mathbf{B} + \mathbf{U}_1\boldsymbol{\Lambda}_1\mathbf{U}_1^{\mathrm{H}}$, we can express the left side of the inequality in (B.6) as

$$\begin{aligned}&\operatorname{tr}\left(\mathbf{A}\left(\mathbf{B}+\mathbf{C}\right)^{-1}\right)\\ &= \operatorname{tr}\left(\mathbf{A}\mathbf{B}^{-1}\right) - \operatorname{tr}\left(\mathbf{A}\mathbf{B}^{-1}\mathbf{U}_1\left(\mathbf{U}_1^{\mathrm{H}}\mathbf{B}^{-1}\mathbf{U}_1 + \boldsymbol{\Lambda}_1^{-1}\right)^{-1}\mathbf{U}_1^{\mathrm{H}}\mathbf{B}^{-1}\right)\end{aligned} \tag{B.7}$$

that is strictly less than $\operatorname{tr}\left(\mathbf{A}\mathbf{B}^{-1}\right)$ if $\mathbf{U}_1^{\mathrm{H}}\mathbf{B}^{-1}\mathbf{A}\mathbf{B}^{-1}\mathbf{U}_1$ is non-zero by Lemma B.3 noting that the matrix $\left(\mathbf{U}_1^{\mathrm{H}}\mathbf{B}^{-1}\mathbf{U}_1 + \boldsymbol{\Lambda}_1^{-1}\right)^{-1}$ is positive definite. If $\mathbf{U}_1^{\mathrm{H}}\mathbf{B}^{-1}\mathbf{A}\mathbf{B}^{-1}\mathbf{U}_1 = \mathbf{0}_{r\times r}$ that is equivalent to $\mathbf{C}\mathbf{B}^{-1}\mathbf{A} = \mathbf{0}_{N\times N}$, then both sides of (B.6) are equal. □

Lemma B.5. Consider the vector $\mathbf{a} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_N, \mathbf{A})$, with covariance matrix $\mathbf{A} \in \mathbb{C}^{N\times N}$, and any deterministic matrix $\mathbf{B} \in \mathbb{C}^{N\times N}$. It holds that

$$\mathbb{E}\{|\mathbf{a}^{\mathrm{H}}\mathbf{B}\mathbf{a}|^2\} = |\operatorname{tr}(\mathbf{B}\mathbf{A})|^2 + \operatorname{tr}(\mathbf{B}\mathbf{A}\mathbf{B}^{\mathrm{H}}\mathbf{A}). \tag{B.8}$$

Proof. Note that $\mathbf{a} = \mathbf{A}^{\frac{1}{2}}\mathbf{w}$ for $\mathbf{w} = [w_1 \ \ldots \ w_N]^{\mathrm{T}} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_N, \mathbf{I}_N)$, thus we can write

$$\mathbb{E}\{|\mathbf{a}^{\mathrm{H}}\mathbf{B}\mathbf{a}|^2\} = \mathbb{E}\{|\mathbf{w}^{\mathrm{H}}(\mathbf{A}^{\mathrm{H}})^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}}\mathbf{w}|^2\} = \mathbb{E}\{|\mathbf{w}^{\mathrm{H}}\mathbf{C}\mathbf{w}|^2\} \tag{B.9}$$

where we defined $\mathbf{C} = (\mathbf{A}^{\mathrm{H}})^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}}$. Let c_{n_1,n_2} denote the element in $\mathbf{C}$ on row n_1 and column n_2. Using this notation, we can expand (B.9) as

$$\mathbb{E}\{|\mathbf{w}^{\mathrm{H}}\mathbf{C}\mathbf{w}|^2\} = \sum_{n_1=1}^{N}\sum_{n_2=1}^{N}\sum_{m_1=1}^{N}\sum_{m_2=1}^{N}\mathbb{E}\{w_{n_1}^{\star}c_{n_1,n_2}w_{n_2}w_{m_1}c_{m_1,m_2}^{\star}w_{m_2}^{\star}\}$$

$$\begin{aligned}
&\overset{(a)}{=} \sum_{n=1}^{N} \mathbb{E}\{|w_n|^4\}|c_{n,n}|^2 + \sum_{n=1}^{N}\sum_{\substack{m=1\\m\neq n}}^{N} \mathbb{E}\{|w_n|^2\}\mathbb{E}\{|w_m|^2\}c_{n,n}c_{m,m}^{\star} \\
&+ \sum_{n_1=1}^{N}\sum_{\substack{n_2=1\\n_2\neq n_1}}^{N} \mathbb{E}\{|w_{n_1}|^2\}\mathbb{E}\{|w_{n_2}|^2\}|c_{n_1,n_2}|^2 \\
&\overset{(b)}{=} \sum_{n=1}^{N} 2|c_{n,n}|^2 + \sum_{n=1}^{N}\sum_{\substack{m=1\\m\neq n}}^{N} c_{n,n}c_{m,m}^{\star} + \sum_{n_1=1}^{N}\sum_{\substack{n_2=1\\n_2\neq n_1}}^{N} |c_{n_1,n_2}|^2 \\
&= \sum_{n=1}^{N}\sum_{m=1}^{N} c_{n,n}c_{m,m}^{\star} + \sum_{n_1=1}^{N}\sum_{n_2=1}^{N} |c_{n_1,n_2}|^2 \\
&\overset{(c)}{=} |\mathrm{tr}(\mathbf{C})|^2 + \mathrm{tr}(\mathbf{C}\mathbf{C}^{\mathrm{H}})
\end{aligned} \tag{B.10}$$

where (a) utilizes that circular symmetry implies that $\mathbb{E}\{w_{n_1}^{\star} w_{n_2} w_{m_1} w_{m_2}^{\star}\}$ is only non-zero when the terms with conjugates have matching indices to the terms without conjugates. The first expression is given by $n_1 = n_2 = m_1 = m_2$, the second term is given by $n_1 = n_2$ and $m_1 = m_2$ with $n_1 \neq m_1$, and the third term is given by $n_1 = m_1$ and $n_2 = m_2$ with $n_1 \neq n_2$. In (b), we utilize that $\mathbb{E}\{|w_n|^2\} = 1$ and $\mathbb{E}\{|w_n|^4\} = 2$. In (c), we write the sums of elements in $\mathbf{C}$ using the trace. The resulting expression is equivalent to (B.8), which is shown by replacing $\mathbf{C}$ with $\mathbf{A}$ and $\mathbf{B}$ and utilizing the fact that $\mathrm{tr}(\mathbf{C}_1\mathbf{C}_2) = \mathrm{tr}(\mathbf{C}_2\mathbf{C}_1)$ for any matrices $\mathbf{C}_1, \mathbf{C}_2$ such that $\mathbf{C}_1$ and $\mathbf{C}_2^{\mathrm{T}}$ have the same dimensions. □

C

Collection of Proofs

This appendix contains proofs of lemmas, theorems, and corollaries that were deemed to long to be placed in the main body of the monograph.

C.1 Proofs from Section 4

We report below the proofs from Section 4.

C.1.1 Proof of Lemma 4.2

To prove this result, it is enough to show that the Hessian matrix $D^2\mathsf{NMSE}(\boldsymbol{\lambda})$ is negative definite for $\lambda_n \geq 0$. $D^2\mathsf{NMSE}(\boldsymbol{\lambda})$ is given as

$$D^2\mathsf{NMSE}(\boldsymbol{\lambda}) = \frac{1}{N\beta}\mathrm{diag}\left(\frac{-2\eta\tau_p\sigma_{\mathrm{ul}}^2}{\left(\eta\tau_p\lambda_1+\sigma_{\mathrm{ul}}^2\right)^3},\ldots,\frac{-2\eta\tau_p\sigma_{\mathrm{ul}}^2}{\left(\eta\tau_p\lambda_N+\sigma_{\mathrm{ul}}^2\right)^3}\right) \quad \text{(C.1)}$$

and is negative definite since it is diagonal with strictly negative entries.

C.1.2 Proof of Lemma 4.3

Note that all the elements of $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}'$ are identical except the $(r-1)$th and the rth ones. Hence, the difference between $\mathsf{NMSE}(\boldsymbol{\lambda})$ and $\mathsf{NMSE}(\boldsymbol{\lambda}')$ results from the summation terms in (4.29) for $n \in \{r-1, r\}$. Using this property, $\mathsf{NMSE}(\boldsymbol{\lambda}) - \mathsf{NMSE}(\boldsymbol{\lambda}')$ is

$$\begin{aligned}
&\frac{\eta\tau_p}{N\beta}\left(\frac{(\lambda_{r-1}+\lambda_r)^2}{\eta\tau_p(\lambda_{r-1}+\lambda_r)+\sigma_{\mathrm{ul}}^2}-\frac{\lambda_{r-1}^2}{\eta\tau_p\lambda_{r-1}+\sigma_{\mathrm{ul}}^2}-\frac{\lambda_r^2}{\eta\tau_p\lambda_r+\sigma_{\mathrm{ul}}^2}\right)\\
&\overset{(a)}{=}\frac{1}{N\beta}\left(\frac{(x+y)^2}{x+y+c}-\frac{x^2}{x+c}-\frac{y^2}{y+c}\right)\\
&=\frac{(x+y)^2\left(xy+c\left(x+y+c\right)\right)-\left(x^2\left(y+c\right)+y^2\left(x+c\right)\right)\left(x+y+c\right)}{N\beta\left(x+y+c\right)\left(x+c\right)\left(y+c\right)}
\end{aligned}$$

$$\overset{(b)}{=} \frac{(x+y)^2 xy + 2cxy(x+y+c) - xy(x+y)(x+y+c)}{N\beta(x+y+c)(x+c)(y+c)}$$
$$\overset{(c)}{=} \frac{2cxy(x+y+c) - cxy(x+y)}{N\beta(x+y+c)(x+c)(y+c)} \overset{(d)}{>} 0 \tag{C.2}$$

where we have introduced $x = \lambda_{r-1}$, $y = \lambda_r$, and the constant $c = \sigma_{\text{ul}}^2/(\eta\tau_p)$ in (a) for simplicity. In (b) and (c), we have canceled the common terms $(x^2+y^2)c(x+y+c)$ and $(x+y)^2 xy$, respectively, in the numerator. Finally, the result in (d) is obtained from the fact that $x > 0$, $y > 0$, and $c > 0$.

C.2 Proofs from Section 5

We report below the proofs from Section 5.

C.2.1 Proof of Theorem 5.1

The processed signal in (5.3) can be treated as the discrete memoryless interference channel in Lemma 3.5 on p. 248 with a random channel response $h = \mathbf{v}_k^{\text{H}} \mathbf{D}_k \widehat{\mathbf{h}}_k$, the input $x = s_k$, the output $y = \mathbf{v}_k^{\text{H}} \mathbf{D}_k \mathbf{y}^{\text{ul}}$, and the realization $u = \{\mathbf{D}_k \widehat{\mathbf{h}}_i : i = 1, \ldots, K\}$ that affects the conditional variance of the interference. The effective noise $\mathbf{v}_k^{\text{H}} \mathbf{D}_k \mathbf{n}$ may not be Gaussian and all the interference and noise are included in v with $n = 0$ in Lemma 3.5 on p. 248. The input power is $p = \mathbb{E}\{|s_k|^2\} = p_k$. The term v is given by

$$v = \sum_{\substack{i=1 \\ i \neq k}}^{K} \mathbf{v}_k^{\text{H}} \mathbf{D}_k \widehat{\mathbf{h}}_i s_i + \sum_{i=1}^{K} \mathbf{v}_k^{\text{H}} \mathbf{D}_k \widetilde{\mathbf{h}}_i s_i + \mathbf{v}_k^{\text{H}} \mathbf{D}_k \mathbf{n}. \tag{C.3}$$

To prove the theorem, we need to show that the requirements in Lemma 3.5 on p. 248 are satisfied and then compute the conditional variance $p_v(h,u) = \mathbb{E}\{|v|^2 \,|h,u\}$. First, we notice that the realizations of h and u are known at the CPU and v has conditional zero-mean given (h,u), i.e., $\mathbb{E}\{v|h,u\} = 0$ since the symbols $\{s_i : i = 1, \ldots, K\}$ and the noise vector $\mathbf{n}$ are independent of the channel estimates and zero-mean estimation errors. The conditional variance given (h,u) is

$$p_v(h,u) = \mathbb{E}\left\{|v|^2 \,|h,u\right\} = \mathbb{E}\left\{|v|^2 \,|\left\{\mathbf{D}_k \widehat{\mathbf{h}}_i\right\}\right\}$$

$$
\begin{aligned}
&= \sum_{\substack{i=1 \\ i \neq k}}^{K} p_i \left| \mathbf{v}_k^{\mathrm{H}} \mathbf{D}_k \widehat{\mathbf{h}}_i \right|^2 + \sum_{i=1}^{K} p_i \mathbf{v}_k^{\mathrm{H}} \mathbf{D}_k \mathbb{E}\left\{ \widetilde{\mathbf{h}}_i \widetilde{\mathbf{h}}_i^{\mathrm{H}} \right\} \mathbf{D}_k \mathbf{v}_k \\
&\quad + \mathbf{v}_k^{\mathrm{H}} \mathbf{D}_k \mathbb{E}\{\mathbf{n}\mathbf{n}^{\mathrm{H}}\} \mathbf{D}_k \mathbf{v}_k \\
&= \sum_{\substack{i=1 \\ i \neq k}}^{K} p_i \left| \mathbf{v}_k^{\mathrm{H}} \mathbf{D}_k \widehat{\mathbf{h}}_i \right|^2 + \sum_{i=1}^{K} p_i \mathbf{v}_k^{\mathrm{H}} \mathbf{D}_k \mathbf{C}_i \mathbf{D}_k \mathbf{v}_k + \sigma_{\mathrm{ul}}^2 \mathbf{v}_k^{\mathrm{H}} \mathbf{D}_k \mathbf{D}_k \mathbf{v}_k
\end{aligned}
\tag{C.4}
$$

which is equal to the denominator of (5.5). In the derivation, we have used the fact that the individual terms of υ are mutually uncorrelated and the combining vector $\mathbf{v}_k$ depends only on the channel estimates, which are independent of the estimation errors. As a final requirement for using Lemma 3.5, we note that the input signal $x = s_k$ is conditionally uncorrelated with υ given (h, u), i.e., $\mathbb{E}\{s_k^\star \upsilon | \{\mathbf{D}_k \widehat{\mathbf{h}}_i\}\} = 0$ due to the independence of the different symbols and the zero-mean channel estimation errors.

As a last step, we note that only the fraction τ_u/τ_c of the samples is used for uplink data transmission, which results in the lower bound on the capacity that is stated in the theorem and measured in bit/s/Hz.

C.2.2 Proof of Theorem 5.2

By adding and subtracting the term $\mathbb{E}\{\mathbf{v}_k^{\mathrm{H}} \mathbf{D}_k \mathbf{h}_k\} s_k$, the received signal in (5.3) can alternatively be expressed as

$$
\widehat{s}_k = \mathbb{E}\{\mathbf{v}_k^{\mathrm{H}} \mathbf{D}_k \mathbf{h}_k\} s_k + \upsilon_k \tag{C.5}
$$

where

$$
\upsilon_k = \left(\mathbf{v}_k^{\mathrm{H}} \mathbf{D}_k \mathbf{h}_k - \mathbb{E}\{\mathbf{v}_k^{\mathrm{H}} \mathbf{D}_k \mathbf{h}_k\} \right) s_k + \sum_{\substack{i=1 \\ i \neq k}}^{K} \mathbf{v}_k^{\mathrm{H}} \mathbf{D}_k \mathbf{h}_i s_i + \mathbf{v}_k^{\mathrm{H}} \mathbf{D}_k \mathbf{n}.
$$

This is a deterministic channel (since $\mathbb{E}\{\mathbf{v}_k^{\mathrm{H}} \mathbf{D}_k \mathbf{h}_k\}$ is constant) with the additive interference plus noise term υ_k, which has zero mean since $\{s_i : i = 1, \ldots, K\}$ and $\mathbf{n}$ have zero mean. Although υ_k contains the desired signal s_k, it is uncorrelated with it since

$$
\mathbb{E}\left\{ s_k^\star \upsilon_k \right\} = \underbrace{\mathbb{E}\left\{ \left(\mathbf{v}_k^{\mathrm{H}} \mathbf{D}_k \mathbf{h}_k - \mathbb{E}\{\mathbf{v}_k^{\mathrm{H}} \mathbf{D}_k \mathbf{h}_k\} \right) \right\}}_{=0} \mathbb{E}\{|s_k|^2\} = 0. \tag{C.6}
$$

Therefore, we can apply Lemma 3.3 on p. 245 with $h = \mathbb{E}\{\mathbf{v}_k^{\mathrm{H}}\mathbf{D}_k\mathbf{h}_k\}$, $x = s_k$, $p = p_k$, $\upsilon = \upsilon_k$, and $\sigma^2 = 0$. By noting that the signals of different UEs are independent and that the noise contributions at different APs are independent, we have that

$$\mathbb{E}\{|\upsilon_k|^2\} = \sum_{i=1}^{K} p_i \mathbb{E}\left\{|\mathbf{v}_k^{\mathrm{H}}\mathbf{D}_k\mathbf{h}_i|^2\right\} - p_k \left|\mathbb{E}\left\{\mathbf{v}_k^{\mathrm{H}}\mathbf{D}_k\mathbf{h}_k\right\}\right|^2 + \sigma_{\mathrm{ul}}^2 \mathbb{E}\left\{\|\mathbf{D}_k\mathbf{v}_k\|^2\right\}. \tag{C.7}$$

The SE expression presented in the theorem then follows from Lemma 3.3. As a last step, we note that only the fraction τ_u/τ_c of the samples is used for uplink data transmission, which results in the lower bound on the capacity that is stated in the theorem and measured in bit/s/Hz.

C.2.3 Proof of Theorem 5.4

Since the CPU does not have knowledge of the channel estimates, it needs to treat the average channel gain $\mathbf{a}_k^{\mathrm{H}}\mathbb{E}\{\mathbf{g}_{kk}\}$ as the true deterministic channel. Hence, the signal model is

$$\widehat{s}_k = \mathbf{a}_k^{\mathrm{H}}\mathbb{E}\{\mathbf{g}_{kk}\}s_k + \upsilon_k \tag{C.8}$$

which is a deterministic channel with the additive interference-plus-noise term

$$\upsilon_k = \left(\mathbf{a}_k^{\mathrm{H}}\mathbf{g}_{kk} - \mathbf{a}_k^{\mathrm{H}}\mathbb{E}\{\mathbf{g}_{kk}\}\right)s_k + \sum_{\substack{i=1 \\ i\neq k}}^{K} \mathbf{a}_k^{\mathrm{H}}\mathbf{g}_{ki}s_i + n_k'. \tag{C.9}$$

The term υ_k has zero mean and is uncorrelated with the signal term in (C.8) since

$$\underbrace{\mathbb{E}\left\{\mathbf{a}_k^{\mathrm{H}}\mathbf{g}_{kk} - \mathbf{a}_k^{\mathrm{H}}\mathbb{E}\left\{\mathbf{g}_{kk}\right\}\right\}}_{=0}\mathbb{E}\left\{|s_k|^2\right\} = 0. \tag{C.10}$$

Therefore, we can apply Lemma 3.3 on p. 245 with $h = \mathbf{a}_k^{\mathrm{H}}\mathbb{E}\{\mathbf{g}_{kk}\}$, $x = s_k$, $p = p_k$, $\upsilon = \upsilon_k$, and $\sigma^2 = 0$. By noting that the signals of different UEs are independent and that the received noise at different APs are independent, we have that

$$\mathbb{E}\{|\upsilon_k|^2\} = \sum_{i=1}^{K} p_i \mathbb{E}\{|\mathbf{a}_k^{\mathrm{H}}\mathbf{g}_{ki}|^2\} - p_k \left|\mathbf{a}_k^{\mathrm{H}}\mathbb{E}\{\mathbf{g}_{kk}\}\right|^2 + \mathbf{a}_k^{\mathrm{H}}\mathbf{F}_k\mathbf{a}_k. \tag{C.11}$$

The SE expression presented in the theorem then follows from Lemma 3.3. As a last step, we note that only the fraction τ_u/τ_c of the samples is used for uplink data transmission, which results in the lower bound on the capacity that is stated in the theorem and measured in bit/s/Hz.

C.2.4 Proof of Corollary 5.6

The proof consists of a direct computation of the three types of expectations that appear in (5.26). We begin with

$$\begin{aligned}\left[\mathbb{E}\left\{\mathbf{g}_{ki}\right\}\right]_l &= \mathbb{E}\left\{\mathbf{v}_{kl}^{\mathrm{H}}\mathbf{D}_{kl}\mathbf{h}_{il}\right\} = \mathrm{tr}\left(\mathbf{D}_{kl}\mathbb{E}\left\{\widehat{\mathbf{h}}_{il}\widehat{\mathbf{h}}_{kl}^{\mathrm{H}}\right\}\right) \\ &= \begin{cases} \sqrt{\eta_k\eta_i}\tau_p \mathrm{tr}\left(\mathbf{D}_{kl}\mathbf{R}_{il}\boldsymbol{\Psi}_{t_kl}^{-1}\mathbf{R}_{kl}\right) & i \in \mathcal{P}_k \\ 0 & i \notin \mathcal{P}_k \end{cases}\end{aligned} \tag{C.12}$$

where the second equality follows from the second identity of Lemma B.2 on p. 445 and the fact that $\widetilde{\mathbf{h}}_{il}$ and $\widehat{\mathbf{h}}_{kl}$ are independent. The third equality follows from (4.19) and gives the expression in (5.33). Similarly,

$$\left[\mathbf{F}_k\right]_{ll} = \sigma_{\mathrm{ul}}^2\mathbb{E}\left\{\|\mathbf{D}_{kl}\mathbf{v}_{kl}\|^2\right\} = \sigma_{\mathrm{ul}}^2\mathrm{tr}\left(\mathbf{D}_{kl}\mathbb{E}\left\{\widehat{\mathbf{h}}_{kl}\widehat{\mathbf{h}}_{kl}^{\mathrm{H}}\right\}\right) = \sigma_{\mathrm{ul}}^2\left[\mathbb{E}\left\{\mathbf{g}_{kk}\right\}\right]_l \tag{C.13}$$

where we used the second identity from Lemma B.2 and then identify the expression of $\left[\mathbb{E}\left\{\mathbf{g}_{kk}\right\}\right]_l$. This gives us the expression in (5.35).

It remains to compute the elements of $\mathbb{E}\left\{\mathbf{g}_{ki}\mathbf{g}_{ki}^{\mathrm{H}}\right\}$. We observe that $\mathbb{E}\left\{\left[\mathbf{g}_{ki}\right]_l\left[\mathbf{g}_{ki}^{\star}\right]_r\right\} = \mathbb{E}\left\{\left[\mathbf{g}_{ki}\right]_l\right\}\mathbb{E}\left\{\left[\mathbf{g}_{ki}^{\star}\right]_r\right\}$ for $r \neq l$ due to the independence of the channels of different APs. Hence, it only remains to compute

$$\left[\mathbb{E}\left\{\mathbf{g}_{ki}\mathbf{g}_{ki}^{\mathrm{H}}\right\}\right]_{ll} = \mathbb{E}\left\{\widehat{\mathbf{h}}_{kl}^{\mathrm{H}}\mathbf{D}_{kl}\mathbf{h}_{il}\mathbf{h}_{il}^{\mathrm{H}}\mathbf{D}_{kl}\widehat{\mathbf{h}}_{kl}\right\} = \mathrm{tr}\left(\mathbf{D}_{kl}\mathbb{E}\left\{\mathbf{h}_{il}\mathbf{h}_{il}^{\mathrm{H}}\mathbf{D}_{kl}\widehat{\mathbf{h}}_{kl}\widehat{\mathbf{h}}_{kl}^{\mathrm{H}}\right\}\right) \tag{C.14}$$

where we utilized the second identity from Lemma B.2.

If $i \notin \mathcal{P}_k$, we can utilize the independence of $\widehat{\mathbf{h}}_{kl}$ and $\mathbf{h}_{il}$ to obtain

$$\begin{aligned}\mathrm{tr}\left(\mathbf{D}_{kl}\mathbb{E}\left\{\mathbf{h}_{il}\mathbf{h}_{il}^{\mathrm{H}}\mathbf{D}_{kl}\widehat{\mathbf{h}}_{kl}\widehat{\mathbf{h}}_{kl}^{\mathrm{H}}\right\}\right) &= \mathrm{tr}\left(\mathbf{D}_{kl}\mathbb{E}\left\{\mathbf{h}_{il}\mathbf{h}_{il}^{\mathrm{H}}\right\}\mathbf{D}_{kl}\mathbb{E}\left\{\widehat{\mathbf{h}}_{kl}\widehat{\mathbf{h}}_{kl}^{\mathrm{H}}\right\}\right) \\ &= \eta_k\tau_p\mathrm{tr}\left(\mathbf{D}_{kl}\mathbf{R}_{il}\mathbf{R}_{kl}\boldsymbol{\Psi}_{t_kl}^{-1}\mathbf{R}_{kl}\right)\end{aligned} \tag{C.15}$$

where we (for brevity) omitted one $\mathbf{D}_{kl}$ term that does not affect the result.

If $i \in \mathcal{P}_k$, we notice that

$$\begin{aligned}&\mathrm{tr}\left(\mathbf{D}_{kl}\mathbb{E}\left\{\mathbf{h}_{il}\mathbf{h}_{il}^{\mathrm{H}}\mathbf{D}_{kl}\widehat{\mathbf{h}}_{kl}\widehat{\mathbf{h}}_{kl}^{\mathrm{H}}\right\}\right)\\&=\mathrm{tr}\left(\mathbf{D}_{kl}\mathbb{E}\left\{\widehat{\mathbf{h}}_{il}\widehat{\mathbf{h}}_{il}^{\mathrm{H}}\mathbf{D}_{kl}\widehat{\mathbf{h}}_{kl}\widehat{\mathbf{h}}_{kl}^{\mathrm{H}}\right\}\right)+\mathrm{tr}\left(\mathbf{D}_{kl}\mathbb{E}\left\{\widetilde{\mathbf{h}}_{il}\widetilde{\mathbf{h}}_{il}^{\mathrm{H}}\mathbf{D}_{kl}\widehat{\mathbf{h}}_{kl}\widehat{\mathbf{h}}_{kl}^{\mathrm{H}}\right\}\right)\end{aligned} \tag{C.16}$$

where the equality follows from separating $\mathbf{h}_{il}$ into its estimate and estimation error. The second term becomes $\eta_k\tau_p\mathrm{tr}(\mathbf{D}_{kl}\mathbf{C}_{il}\mathbf{R}_{kl}\boldsymbol{\Psi}_{t_kl}^{-1}\mathbf{R}_{kl})$ by utilizing the independence of estimates and estimation error and omitting one $\mathbf{D}_{kl}$ term. The first term is computed by utilizing the result from (4.18) to rewrite the estimate as $\widehat{\mathbf{h}}_{il} = \sqrt{\frac{\eta_i}{\eta_k}}\mathbf{R}_{il}\mathbf{R}_{kl}^{-1}\widehat{\mathbf{h}}_{kl}$:

$$\begin{aligned}&\mathrm{tr}\left(\mathbf{D}_{kl}\mathbb{E}\left\{\widehat{\mathbf{h}}_{il}\widehat{\mathbf{h}}_{il}^{\mathrm{H}}\mathbf{D}_{kl}\widehat{\mathbf{h}}_{kl}\widehat{\mathbf{h}}_{kl}^{\mathrm{H}}\right\}\right)\\&=\frac{\eta_i}{\eta_k}\mathrm{tr}\left(\mathbf{D}_{kl}\mathbb{E}\left\{\mathbf{R}_{il}(\mathbf{R}_{kl})^{-1}\widehat{\mathbf{h}}_{kl}\widehat{\mathbf{h}}_{kl}^{\mathrm{H}}(\mathbf{R}_{kl})^{-1}\mathbf{R}_{il}\mathbf{D}_{kl}\widehat{\mathbf{h}}_{kl}\widehat{\mathbf{h}}_{kl}^{\mathrm{H}}\right\}\right)\\&=\frac{\eta_i}{\eta_k}\mathbb{E}\{|\widehat{\mathbf{h}}_{kl}^{\mathrm{H}}\mathbf{D}_{kl}\mathbf{R}_{il}(\mathbf{R}_{kl})^{-1}\widehat{\mathbf{h}}_{kl}|^2\}\\&=\eta_k\eta_i\tau_p^2\left|\mathrm{tr}\left(\mathbf{D}_{kl}\mathbf{R}_{il}\boldsymbol{\Psi}_{t_kl}^{-1}\mathbf{R}_{kl}\right)\right|^2+\eta_k\tau_p\mathrm{tr}\left(\mathbf{D}_{kl}(\mathbf{R}_{il}-\mathbf{C}_{il})\mathbf{R}_{kl}\boldsymbol{\Psi}_{t_kl}^{-1}\mathbf{R}_{kl}\right)\end{aligned} \tag{C.17}$$

where the last step follows from Lemma B.5 on p. 446 and some algebra. By adding these two terms together, we finally obtain (5.34). Note that the proof holds even if $\mathbf{R}_{kl}$ is non-invertible because $\mathbf{R}_{kl}^{-1}\widehat{\mathbf{h}}_{kl} = \sqrt{\eta_k\tau_p}\mathbf{R}_{kl}^{-1}\mathbf{R}_{kl}\boldsymbol{\Psi}_{t_kl}^{-1}\mathbf{y}_{t_kl}^{\mathrm{pilot}} = \sqrt{\eta_k\tau_p}\boldsymbol{\Psi}_{t_kl}^{-1}\mathbf{y}_{t_kl}^{\mathrm{pilot}}$, where there is no inversion. We are only using the inversion to shorten the notation.

C.3 Proofs from Section 6

We report below the proofs from Section 6.

C.3.1 Proof of Theorem 6.1

Since the UE k only has knowledge of the average of the effective channel, $\mathbb{E}\{\mathbf{h}_k^{\mathrm{H}}\mathbf{D}_k\mathbf{w}_k\}$, the received signal in (6.7) at UE k can be expressed as

$$y_k^{\mathrm{dl}} = \mathbb{E}\{\mathbf{h}_k^{\mathrm{H}}\mathbf{D}_k\mathbf{w}_k\}\varsigma_k + \upsilon_k + n_k \tag{C.18}$$

which is a deterministic channel with the additive noise n_k and the additive interference term

$$\upsilon_k = \left(\mathbf{h}_k^{\mathrm{H}}\mathbf{D}_k\mathbf{w}_k - \mathbb{E}\left\{\mathbf{h}_k^{\mathrm{H}}\mathbf{D}_k\mathbf{w}_k\right\}\right)\varsigma_k + \sum_{\substack{i=1 \\ i\neq k}}^{K}\mathbf{h}_k^{\mathrm{H}}\mathbf{D}_i\mathbf{w}_i\varsigma_i. \tag{C.19}$$

The υ_k term has zero mean (since ς_i has zero mean) and although it contains the desired signal ς_k, it is uncorrelated with it since

$$\mathbb{E}\left\{\varsigma_k^{\star}\upsilon_k\right\} = \underbrace{\mathbb{E}\left\{\left(\mathbf{h}_k^{\mathrm{H}}\mathbf{D}_k\mathbf{w}_k - \mathbb{E}\left\{\mathbf{h}_k^{\mathrm{H}}\mathbf{D}_k\mathbf{w}_k\right\}\right)\right\}}_{=0}\mathbb{E}\{|\varsigma_k|^2\} = 0. \tag{C.20}$$

Therefore, we can apply Lemma 3.3 on p. 245 with $h = \mathbb{E}\left\{\mathbf{h}_k^{\mathrm{H}}\mathbf{D}_k\mathbf{w}_k\right\}$, $x = \varsigma_k$, $p = 1$, $\upsilon = \upsilon_k$, and $\sigma^2 = \sigma_{\mathrm{dl}}^2$. By noting that the signals of different UEs are independent, we have that

$$\mathbb{E}\{|\upsilon_k|^2\} = \sum_{i=1}^{K}\mathbb{E}\left\{\left|\mathbf{h}_k^{\mathrm{H}}\mathbf{D}_i\mathbf{w}_i\right|^2\right\} - \left|\mathbb{E}\left\{\mathbf{h}_k^{\mathrm{H}}\mathbf{D}_k\mathbf{w}_k\right\}\right|^2. \tag{C.21}$$

The SE expression presented in the theorem then follows from Lemma 3.3 on p. 245. As a last step, we note that only the fraction τ_d/τ_c of the samples is used for downlink data transmission, which results in the lower bound on the capacity that is stated in the theorem and measured in bit/s/Hz.

C.3.2 Proof of Theorem 6.2

Let $\gamma_k = \mathsf{SINR}_k^{(\mathrm{ul,c-UatF})}$ denote the value of the effective SINR in (5.9) for the uplink powers $\{p_i : i = 1, \ldots, K\}$ and combining vectors $\{\mathbf{D}_i\mathbf{v}_i : i = 1, \ldots, K\}$. We want to show that $\gamma_k = \mathsf{SINR}_k^{(\mathrm{dl,c})}$ is achievable in the downlink when (6.11) is satisfied for all i. Plugging (6.11) into (6.10) yields the following SINR constraints:

$$\gamma_k = \frac{\rho_k\left|\mathbb{E}\left\{\mathbf{h}_k^{\mathrm{H}}\frac{\mathbf{D}_k\mathbf{v}_k}{\sqrt{\mathbb{E}\{\|\mathbf{D}_k\mathbf{v}_k\|^2\}}}\right\}\right|^2}{\sum\limits_{i=1}^{K}\rho_i\mathbb{E}\left\{\left|\mathbf{h}_k^{\mathrm{H}}\frac{\mathbf{D}_i\mathbf{v}_i}{\sqrt{\mathbb{E}\{\|\mathbf{D}_i\mathbf{v}_i\|^2\}}}\right|^2\right\} - \rho_k\left|\mathbb{E}\left\{\mathbf{h}_k^{\mathrm{H}}\frac{\mathbf{D}_k\mathbf{v}_k}{\sqrt{\mathbb{E}\{\|\mathbf{D}_k\mathbf{v}_k\|^2\}}}\right\}\right|^2 + \sigma_{\mathrm{dl}}^2} \tag{C.22}$$

for $k = 1, \ldots, K$. We define the diagonal matrix $\mathbf{\Gamma} \in \mathbb{R}^{K \times K}$ with the kth diagonal element being

$$[\mathbf{\Gamma}]_{kk} = \frac{1}{\gamma_k} \left| \mathbb{E} \left\{ \mathbf{h}_k^{\mathrm{H}} \frac{\mathbf{D}_k \mathbf{v}_k}{\sqrt{\mathbb{E}\{\mathbf{v}_k^{\mathrm{H}} \mathbf{D}_k \mathbf{v}_k\}}} \right\} \right|^2 \tag{C.23}$$

and let $\mathbf{\Sigma} \in \mathbb{R}^{K \times K}$ be the matrix whose (k, i)th element is

$$[\mathbf{\Sigma}]_{ki} = \mathbb{E} \left\{ \left| \mathbf{h}_k^{\mathrm{H}} \frac{\mathbf{D}_i \mathbf{v}_i}{\sqrt{\mathbb{E}\{\mathbf{v}_i^{\mathrm{H}} \mathbf{D}_i \mathbf{v}_i\}}} \right|^2 \right\} - \begin{cases} 0 & i \neq k \\ \gamma_k [\mathbf{\Gamma}]_{kk} & i = k. \end{cases} \tag{C.24}$$

Using these matrices, the SINR constraint in (C.22) can be expressed as

$$[\mathbf{\Gamma}]_{kk} = \frac{\sum_{i=1}^{K} \rho_i [\mathbf{\Sigma}]_{ki} + \sigma_{\mathrm{dl}}^2}{\rho_k}. \tag{C.25}$$

By rearranging this equation, we obtain $\sigma_{\mathrm{dl}}^2 = \rho_k [\mathbf{\Gamma}]_{kk} - \sum_{i=1}^{K} \rho_i [\mathbf{\Sigma}]_{ki}$. The K constraints can be written in matrix form as $\mathbf{1}_K \sigma_{\mathrm{dl}}^2 = (\mathbf{\Gamma} - \mathbf{\Sigma})\, \boldsymbol{\rho}$ with $\boldsymbol{\rho} = [\rho_1 \ \ldots \ \rho_K]^{\mathrm{T}}$ being the downlink transmit power vector. The SINR constraints in (C.22) are thus satisfied if

$$\boldsymbol{\rho} = (\mathbf{\Gamma} - \mathbf{\Sigma})^{-1} \mathbf{1}_K \sigma_{\mathrm{dl}}^2. \tag{C.26}$$

This is a feasible power if $\mathbf{\Gamma} - \mathbf{\Sigma}$ is invertible, which always holds when $\mathbf{p} = [p_1 \ \ldots \ p_K]^{\mathrm{T}}$ is feasible. To show this, we notice that the K uplink SINR conditions can be expressed in a similar form where $\mathbf{\Sigma}$ is replaced by $\mathbf{\Sigma}^{\mathrm{T}}$ such that $\mathbf{p} = (\mathbf{\Gamma} - \mathbf{\Sigma}^{\mathrm{T}})^{-1} \mathbf{1}_K \sigma_{\mathrm{ul}}^2$. Since the eigenvalues of $\mathbf{\Gamma} - \mathbf{\Sigma}$ and $\mathbf{\Gamma} - \mathbf{\Sigma}^{\mathrm{T}}$ are the same and the uplink SINR conditions are satisfied by assumption, we can always select the downlink powers according to (C.26). Substituting $\mathbf{1}_K = \frac{1}{\sigma_{\mathrm{ul}}^2} (\mathbf{\Gamma} - \mathbf{\Sigma}^{\mathrm{T}})\, \mathbf{p}$ into (C.26) yields

$$\boldsymbol{\rho} = \frac{\sigma_{\mathrm{dl}}^2}{\sigma_{\mathrm{ul}}^2} (\mathbf{\Gamma} - \mathbf{\Sigma})^{-1} (\mathbf{\Gamma} - \mathbf{\Sigma}^{\mathrm{T}})\, \mathbf{p}. \tag{C.27}$$

The total transmit power condition now follows from direct computation by noting that $\mathbf{1}_K^{\mathrm{T}} (\mathbf{\Gamma} - \mathbf{\Sigma})^{-1} \mathbf{1}_K = \mathbf{1}_K^{\mathrm{T}} (\mathbf{\Gamma} - \mathbf{\Sigma}^{\mathrm{T}})^{-1} \mathbf{1}_K$.

To complete the proof, we need to show the power allocation coefficients obtained by (C.27) are positive. Please see [**massivemimobook**] for the final technical details.