Distributed Processing Methods for Extra Large Scale MIMO

Amiri, Abolfazl

DOI (link to publication from Publisher):
10.54337/aau466211655

Publication date:
2021

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Amiri, A. (2021). *Distributed Processing Methods for Extra Large Scale MIMO*. Aalborg Universitetsforlag.

DISTRIBUTED PROCESSING METHODS FOR EXTRA LARGE SCALE MIMO

BY
ABOLFAZL AMIRI

DISSERTATION SUBMITTED 2021

AALBORG UNIVERSITY
DENMARK

Distributed Processing Methods for Extra Large Scale MIMO

Ph.D. Dissertation

Abolfazl Amiri

Dissertation submitted month 10, 2021

Dissertation submitted:	October 2021
PhD supervisor:	Prof. Elisabeth de Carvalho Aalborg University
Assistant PhD supervisor:	Assoc. Prof. Carles Navarro Manchón Prof. Petar Popovski Aalborg University
PhD committee:	Associate Professor Troels Bundgaard Sørensen (chairman) Aalborg University Professor Ana Garcia Armada Universidad Carlos III de Madrid Professor Peter Höher University of Kiel
PhD Series:	Technical Faculty of IT and Design, Aalborg University
Institut:	Department of Electronic Systems

ISSN (online): 2446-1628
ISBN (online): 978-87-7573-986-8

Published by:
Aalborg University Press
Kroghstræde 3
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk

Printed in Denmark by Rosendahls, 2021

Curriculum Vitae

Abolfazl Amiri

He received the B.S. degree in Electrical and Communications Engineering (with honours) from the University of Tabriz, Tabriz, Iran in 2013, and the M.S. degree in Electrical and Communication Systems Engineering from the University of Tehran, Tehran, Iran in 2016. He was with Huawei technologies Tehran office from 2016 to 2017 as an RF engineer and cellular network optimizer. He was a research assistant on multi-antenna signal processing at Aalborg University from 2017 to 2018. He is currently a PhD fellow in the Electronic systems department of Aalborg University, Aalborg, Denmark. His research interests include applications of signal processing and machine learning in wireless communications and multi-antenna systems.

Abstract

Massive MIMO (multiple-input multiple-output) systems are key candidates for the fifth generation (5G) of cellular networks. Having a lot of antenna elements at the base station (BS) is an important enabler to provide a very high spatial resolution. Therefore, systems beyond 5G rely on increasing the number of elements at the BS to support future applications. At very large dimensions, e.g. aperture sizes bigger than 100 wavelengths, a new type of array called extra-large scale MIMO (XL-MIMO) emerges that offers enhanced spectral and energy efficiency. However, practical implementation of such arrays requires overcoming several challenges such as computational complexity, hardware limitations and non-stationary propagation patterns.

This thesis presents several techniques to handle major existing concerns in the XL-MIMO arrays, namely: computational complexity of receiver algorithms, scalability and interconnection overheads. In order to address the complexity issue, different low complexity methods are proposed. One of the main differences between these methods and conventional linear receivers in massive MIMO systems is, that they exploit the information about user energy patterns over the array to operate more effectively. Another approach is to distribute the receiver processing tasks between several nodes and create a hierarchy between processing nodes. The thesis studies different architectures and mostly focuses on a distributed way that uses sub-arrays to obtain local estimates at local nodes. Then, a central node collects all the local data to perform a global decision. Furthermore, the thesis suggests several antenna selection methods to limit the area of the array being processed and control the amount of computations. These methods directly use the received energy patterns at the BS to find the best active antenna sets and turn off the rest of the array to save energy. Moreover, to address the hardware considerations such as scalability and inter-connection overheads, a fully decentralized method is proposed that works without a central node.

In summary, the main outcome of the thesis is the proposal of signal processing enablers for the XL-MIMO systems. The proposed methods address the aforementioned challenges while providing acceptable performance.

Resumé

Massive MIMO (multiple-input multiple-output) systemer er vigtige kandidater til den femte generation (5G) af mobilnetværk. At have mange antenneelementer ved basestationen (BS) er en vigtig muliggør for at give en meget høj rumlig opløsning. Derfor beror systemer over 5G på at øge antallet af antenneelementer på BS for at understøtte fremtidige applikationer. Ved meget store dimensioner, f.eks. ved aperturer større end 100 bølgelængder, opstår en ny type antennearray kaldet ekstra stor MIMO (XL-MIMO), der giver forbedret spektral- og energieffektivitet. Praktisk implementering af sådanne arrays kræver imidlertid at flere udfordringer overvindes, såsom beregningskompleksitet, hardware -begrænsninger og ikke - stationære udbredelsesmønstre.

Denne afhandling præsenterer forskellige teknikker til håndtering af betydelige udfordringer ved XL-MIMO-arrays, navngivelig: beregningsmæssig kompleksitet af modtageralgoritmer, skalerbarhed og sammenkoblings omkostninger. For at løse kompleksitetsproblemet foreslås forskellige lav kompleksitets metoder. En af de vigtigste forskelle mellem disse metoder og konventionelle lineære modtagere i massiv MIMO-systemer er, at de udnytter informationen om brugerens energimønstre over arrayet for at operere mere effektivt. En anden tilgang er at distribuere modtagerens behandlings opgaver mellem flere noder og danne et hierarki mellem behandlingsnoder. Afhandlingen præsenterer forskellige arkitekturer og fokuserer hovedsageligt på en distribueret metode, der bruger underarrays til at opnå lokale estimater på lokale noder. Derefter indsamler en central node alle de lokale data for at udføre en global beslutning. Desuden foreslår afhandlingen adskillige antenneudvælgelsesmetoder for at begrænse arealet af det array, der behandles, og kontrollere antallet af beregninger. Disse metoder bruger direkte de modtagne energimønstre på BS til at finde de bedste sæt af aktive antenner og slukke for resten af arrayet for at spare energi. Desuden foreslås en fuldstændig decentral metode, der fungerer uden en central node, for at imødegå hardwareovervejelser såsom skalerbarhed og sammenkoblings omkostninger.

Sammenfattet er hovedresultatet af afhandlingen forslaget om signalbehandlings aktiverere til XL - MIMO - systemerne. De foreslåede metoder adresserer de førnævnte udfordringer og giver samtidig acceptabel ydeevne.

Contents

Acknowledgements

First and foremost, I would like to thank my supervisors Prof. Elisabeth de Carvalho and Carles Navarro Manch'on who supported me during my Ph.D. studies. Our weekly meetings were the best opportunities to discuss every detail in the research topics and were very helpful. I would also like to express my gratitude to Prof. Petar Popovski who helped me in different stages of my studies and his critical comments improved a lot of my papers.

During my Ph.D. studies, I was fortunate to visit Prof. Taufik Abrao at State University of Londrina (UEL) in Brazil. This visit turned out to affect a considerable part of my Ph.D. studies much more than anticipated. I would like to thank Taufik for his hospitality during my visit, introducing me to his research group, and for providing so many valuable comments on the papers and problems we were trying to address. I would also like to thank Marinello, Joao and Victor for their contribution in many discussions that led to several good publications.

I would like to thank Ana García Armada, Peter Höher, and Troels B. Sørensen for accepting to be committee members at my Ph.D. defense. Thanks to all members of the Connectivity group at Aalborg University and to Dorthe Sparre for her help in administrative tasks. I would also like to thank Sajad Rezaie for joining supervision and study meetings and providing constructive comments to my work. Finally, I would like to express my sincere gratitude to my family and friends for their support.

Abolfazl Amiri
Aalborg University, October 29, 2021

Part I

Introductory Chapters

Chapter 1

Introduction and Motivation

In today's society, we are surrounded by various kinds of connectivity devices and our jobs, social lives and well-being depend on them. Use cases for each of these devices are countless and still keep growing. Unlike 20 years ago, when mobile phones could only make calls or send text messages, nowadays they can be used for many applications such as video conferencing, web surfing and navigation. Moreover, there are many types of user equipment other than mobile phones such as internet of things (IoT) devices or vehicular communication that need wireless connectivity. This rapid growth in the types of services for the mobile (or cellular) networks is a result of the development in wireless technologies. With every new generation of the mobile networks, several improvements and new services are being introduced. Currently, early implementations of the fifth-generation (5G) with limited services are being deployed [1].

There are many key changes in the 5G compared to the older generations that target three goals: very high data rate for interactive applications, ultra-reliable low-latency communications for mission-critical applications and massive machine type connectivity for the internet of things (IoT) use cases. Each of these goals has a different requirement in terms of data rate, latency, reliability and number of connected devices [2]. In order to ensure a certain quality of service (QoS) for any application, a system designer should carefully describe the specifications of it in terms of these 4 requirements. For instance, an application that uses video calling needs a very high data rate, low latency, moderate reliability and a few connected devices to perform as expected. To satisfy these demands, different technologies have been proposed [3]. One of the main technologies for the 5G deployment is massive multiple-input multiple-output (M-MIMO) arrays that use tens of antennas at the base station (BS). Having a lot of antenna elements allows a very high spatial resolution providing a high spectral efficiency (SE) for multiple users [4]. This feature can be used for all three key services in the 5G [8, 9], e.g. high data-rate connectivity. Therefore, the M-MIMO systems attracted a lot of interest in the community.

Multiple antenna arrays were in use way before the idea of the M-MIMO systems. For instance, many commercial systems such as TV broadcasting, radars, localization and navigation systems at airports were using arrays with several antennas. In the late 90s, the idea of space-time codes to improve the bit error rate of a 2×1 single-input multi-output (MISO) system was proposed in [5]. Since then, researchers focused on adding more antennas at both the transmitter and receiver devices to increase achievable data rates. The concept of the M-MIMO communications was first proposed in [6] introducing phenomenal changes in the MIMO systems by assuming a BS with an infinite number of antennas. Even though the idea seemed very theoretical and far from any practical system in the beginning, it got a lot of attention later due to very interesting features of such array systems. Most of the works [4, 16, 17] consider a canonical M-MIMO system, where a multi-carrier and multi-cell network operates in a synchronous time division duplex (TDD) protocol. This allows much more efficient channel estimation methods that only scale with the number of users and not with the number of the BS antennas. Moreover, it is assumed that the number of antennas at the BS is large enough to achieve channel hardening[1] and significantly larger than the number of active single antenna user devices [16]. Furthermore, each of the BSs uses linear processing techniques for signal recovery/transmission independently. Some of the research topics with these assumptions are reaching a finish line. For example, energy and spectral efficiency optimization [18, 19], power optimization [21] and pilot contamination control [20]. Thus, it is natural for any reader to ask *"What is going to happen next? "*. We will get back to this question shortly. But first, we review some of the physical implementation examples of M-MIMO systems to get a more realistic perspective of them.

There are several commercial implementations of M-MIMO arrays that are being used on different network deployments. For instance, Ericsson AIR 6468 [22] which has 64 antenna elements connected to 64 fully digital transceiver chains in both uplink and downlink, and it was designed for 4G LTE and 5G applications in 2018 [13]. Similarly, Huawei has announced that most of their current commercial products have either 32 or 64 antennas [23]. Therefore, one can completely disregard the old claim of high complexity and non-practicality of the M-MIMO arrays, since they are already being used in commercial networks. Now, we can argue that the number of antenna elements can go beyond the above mentioned values in near future to extend the quality of existing services or to enable new services.

[1] Channel hardening makes a fading channel behave as deterministic.

Here, we get back to the question about the possible future trends. Recently, the idea of scaling up the array size in the M-MIMO up to hundreds of elements has become a hot research topic [30–32]. We call this generation of antenna systems, **extra-large scale MIMO (XL-MIMO)** arrays. More precisely, in this thesis, we define antenna arrays with more than 200 elements, which have a physical size of tens of meters[2], as XL-MIMO arrays. The main benefits of the XL-MIMO arrays are threefold:

1. They offer the potential of achieving high SE due to the ability to multiplex a large number of users over the same time-frequency resources. This will support beyond 5G services such as virtual reality gaming.

2. They can provide very high EE that enables extreme obtainable throughput while assuring practical power consumption for the base station.

3. They can easily be deployed on the existing infrastructure such as walls and ceilings in malls, airports and large venues to help with the coverage issues in crowded places [70] [7].

In addition to these merits, all the advantages from M-MIMO systems such as cell edge coverage are also valid for the XL-MIMO arrays. In the following, we elaborate more on each of the mentioned unique benefits. As discussed above, with more antennas at the BS, the ability to fully exploit the spatial degrees of freedom (DoF) enhances. As a result, we can achieve higher data rates or higher reliability in a lower transmission latency. This makes it possible to replace wires with wireless devices in some of the systems such as entertainment tools, factories and surgery tools [11]. The second benefit is about the ability of such arrays to reduce the required power consumption to achieve a certain area throughput. This can indeed save a substantial amount of power and become an enabler for greener network solutions [14]. Finally, the last advantage comes into place when we want to install an XL-MIMO array, where we can use the existing infrastructure, i.e. walls and ceilings [13]. This indeed reduces the costs of the system deployment. For instance, XL-MIMO arrays can be installed all around the walls of a stadium to better cover very high number of simultaneously active users. It is worth mentioning that each of the antenna elements is made of very cheap materials, but together they can provide many different services [14].

One way to implement an XL-MIMO system is to use a group of subarrays that are composed of a number of antenna elements. Fig. 1.1 shows an example of such XL-MIMO system deployment. They can have equal or unequal sizes. The first case is a good option if the environment is homogeneous in terms of user distribution and the second case can be used in

[2] We assume sub-6 GHz frequencies and half a wavelength antenna distancing in a linear array.

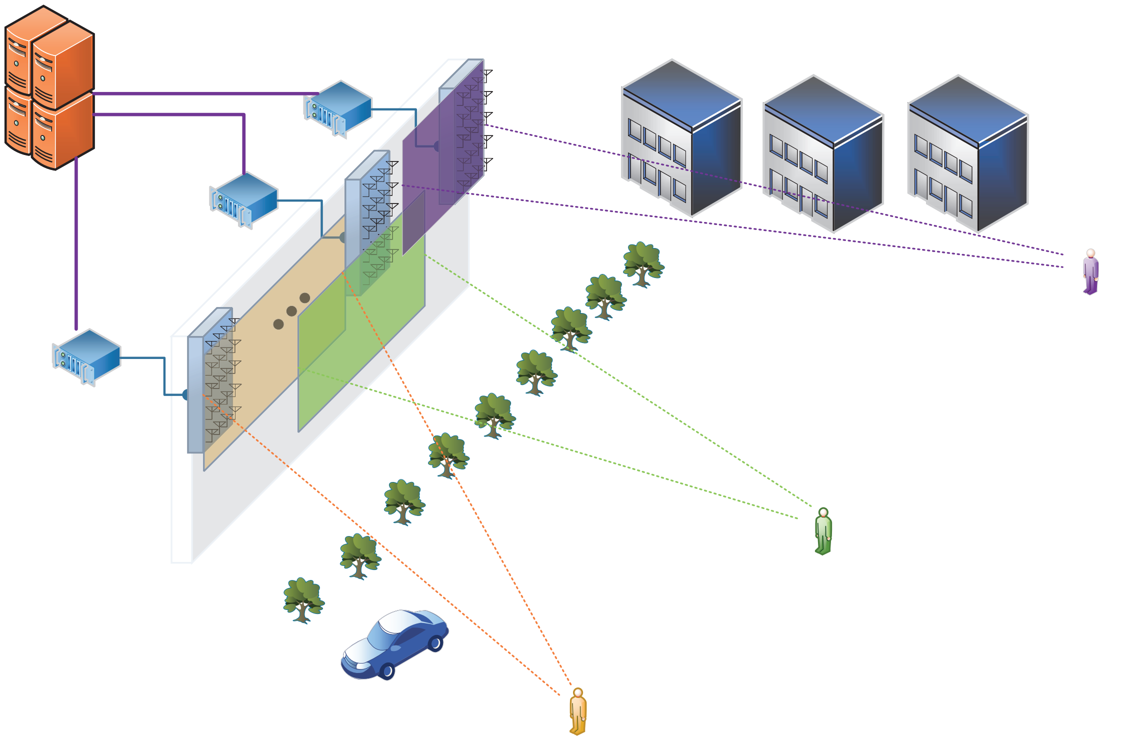


Fig. 1.1: An example of the distributed architecture of the XL-MIMO system with presence of the energy spatial non-stationarities.

non-homogeneous scenarios. The bright side about using sub-arrays is that a part of the processing tasks can be done locally at each sub-array unit and then forwarded to a central unit for a final symbol detection procedure (see Fig. 1.1). This kind of architectures can help in many practical issues such as scaling up the number of antennas, high power central processing units and high communication overhead in the system. On the other hand, the performance degradation due to distributed processing is inevitable and it goes until a point that the network performance requirements cannot be met.

One of the main differences between the XL-MIMO arrays and other MIMO systems is the physical size of the array. Due to its large size, different antenna elements face dissimilar environments [36]. Thus, each part of the array is only visible to subsets of clusters of scatterers in the propagation environment resulting in an uneven received energy distribution along the array. This variation on the received user energies is called spatial non-stationarities. This is different from the conventional M-MIMO case where the energy distributions are assumed to be stationary and independent of the antenna indices. As shown in Fig. 1.1, different parts of the array is visible to each of the users in the area. This variability is degrading the performance of the system and we investigate it in detail in Chapter 3.

1 Motivation

As discussed above, the idea of adding more antennas at the BS has brought a lot of benefits for wireless communications. However, we believe that it does not stop with current M-MIMO array technologies and the antenna count can be pushed even further. In this direction, XL-MIMO arrays have the potential to be a proper candidate for the next generation of the MIMO communications. Due to the benefits of large antenna arrays, we conjecture that the trend of increasing the number of antenna elements in base stations will continue to be the main driver of future systems. Hence, we set out to devise the necessary signal processing algorithms to make such an increase possible and feasible. In short, we try to tackle the problem of high computational complexity, scalability and channel non-stationarity with our methods. A detailed description of the challenges and research questions is presented in the next chapter.

2 Structure of the thesis

The rest of the thesis is organized as the following; first, we start with stating the problem and research questions that this thesis is trying to answer. Then, we continue with categorizing the contributions of the thesis into several groups. Next, for each of the groups, we review the related literature, highlight the research gaps, demonstrate our proposed methods to tackle the existing challenges and present a summary of our papers. We conclude the thesis by discussing the advantages of the proposed receivers and directions for future research.

Chapter 2

Problem statement

In this chapter, we focus on identifying the challenges of XL-MIMO systems and then try to address each of the challenges with a research question. Next, we introduce our methodology together with comparison measures that have been used to evaluate our methods.

1 Challenges

Before listing the existing challenges, we start with a clear definition of an XL-MIMO array and what we mean by it throughout this thesis. An XL-MIMO array is a MIMO array with tens of hundreds of elements where its physical size is in the range of tens of meters for the sub-6 GHz frequency bands. Unlike most of the works in the area of M-MIMO arrays, in this thesis we are not aiming at making large-system or asymptotic analysis of M-MIMO systems; but instead, we would like to focus on the more practical signal processing aspects operating with these very large arrays. For instance, we do not employ random matrix theory, large-system analysis, and asymptotic results on simple theoretical channel models to analyze the problem, and instead, put more emphasis on the processing aspects for arrays with a very large (but finite) number of elements and rely on more realistic and sophisticated models.

Similar to any other technology, there comes a lot of difficulties in the design and development phases. Naturally, a group of the challenges are inherited from the M-MIMO systems. For instance issues regarding channel state information (CSI) acquisition, pilot contamination, hardware non-linearities, etc. However, the focus of this thesis is to study the challenges that arise when we go extra-large scale and for this reason, we avoid addressing typical M-MIMO problems for the rest of the thesis. The main concerns regarding the XL-MIMO systems can be grouped into three major challenges:

1. **Computational complexity:** Larger arrays support more users and require more computational capacity at the processing unit. Due to the large dimensions of the channel matrices, some of the conventional linear signal processing techniques that deal with matrix inversion become unfeasible.

2. **Non-stationarities on the energy patterns:** An important modification on the channel assumptions for this type of array is that the average channel gain varies along the array. We call this uneven distribution of the receiving power over the array channel non-stationarities. This phenomenon is limiting the number of effective antenna elements in the signal transmission and therefore results in a reduction in the array gain and the SE.

3. **Scalability:** Managing more and more antenna elements at the BS causes many deployment issues including the scalability and signalling overhead. For instance, many commercial products are made with a certain number of antenna elements and putting them together to make an extra-large array is not straightforward. The number of interconnections, signalling and total back-haul bandwidth of the overall system should be allocated carefully to ensure the desired performance.

Next, we address each of the mentioned challenges in one research question that this thesis is trying to answer.

2 Research questions (RQs)

RQ1: Can receiver processing methods with feasible complexity lead to a performance close to those of classical receivers, such as zero-forcing (ZF) or minimum mean square error (MMSE), which are infeasible in the XL-MIMO regime? Which techniques are suitable for this task?

RQ2: What is the effect of channel non-stationarities in the performance of XL-MIMO arrays? Can these be accounted for in order to obtain better performing receiver algorithms?

RQ3: Is there a way to make a scalable architecture for an XL-MIMO array? In other words, can we adapt our signal processing methods to work on multiple nodes with minimum signalling overhead between them?

In the following, we investigate each of the questions and provide a short description of our approach to tackle them.

As mentioned before, with an increase in the number of elements in the antenna array and the number of serving users, conventional linear processing methods become computationally heavy. We aim to use distributed processing units to divide the signal processing tasks. Moreover, we try to investigate non-linear and low-complexity techniques to keep the operations as simple as possible. Therefore, the main problem is converted into finding the best way to supervise the distributed units. In other words, we try to design a multi-node system where each node is connected to a sub-set of antenna elements and has limited processing capability. Each of these nodes produces local estimates and the goal is to mix these estimates in a way that the performance of the whole system is close to the one with all nodes working in a central manner. Our main candidate methods are message-passing, randomized and genetic algorithms that have low-complexity and are multi-node compatible. We discuss each of these methods in detail in Chapter 3.

In order to explain the second question, it is worth noting that the energy non-stationarities along the array cause uneven importance for the signals observed at the antenna elements. This property gets even more relevant when we employ distributed processing nodes. In that case, some nodes will have better channel conditions and receive lower interference signals which make their local estimates more *reliable*. This indeed adds to the difficulty of the problem since a uniform mix of the estimates will not result in the best performance. Therefore, there is a need for a smarter decision-making system that mixes the local estimates based on their reliability and signal quality.

Last but not least, the issue with the scalability of the XL-MIMO arrays arises when we try to add more antenna elements to an existing MIMO system to expand its quality of service. For example, if a centralized processing technique is employed, then adding new elements will require connecting each of them to the central unit. In some cases, this might not be possible if the processor is far from the array or its input ports are all in use. Thus, using a distributed architecture would make sense because of lowering the need for wiring and computational capacity. However, this new design should consider the performance loss, delay and overhead caused by decentralizing the array processing unit. Moreover, more relevant to the studies of this thesis, it requires signal processing methods that can deal with these issues and whose complexity scales well in such situations.

3 Methodology

Throughout this thesis, we have used computer simulations implemented by MATLAB. The simulations use the Monte-Carlo method for repeated random experiments since closed-form performance measures are not available for the proposed techniques. Our performance assessment is based on synthetic channel realizations generated using stochastic channel models. The main sources of randomness in the simulations are the additive Gaussian noise, the fading channel gain matrix and the transmitted data symbols for each of the users.

In order to compare the performance of the proposed methods, we use several measures. Some of these metrics include symbol and bit error rates (coded and uncoded), sum-rate, energy efficiency, spectral efficiency and post-processing signal to noise and interference ratio. Moreover, the complexity of the algorithms is compared in terms of the number of complex multiplications and the size of signalling overhead between the distributed units.

With the aim of finding out how good our methods are, we have implemented several well-known benchmark techniques. In general, these techniques can be classified into two groups of linear and non-linear methods. Each of the methods within these groups has two implementation types: 1. centralized, where all of the processing tasks are done in a central node, 2. distributed mode, where several nodes are operating in co-operation. It is worth mentioning that the main idea is to propose a distributed method that works very close to a centralized processing technique.

3.1 Scope of the study and main assumptions

In this subsection, we mention the main assumptions made throughout this thesis. We study single-cell scenarios with orthogonal pilots for all the active users. We consider synchronous TDD cases where the reciprocity between the uplink and downlink channel holds. Channel state information (CSI) is assumed to be known at the BS. We leave channel coding out of the scope of our research and the objective of our proposed receivers is to recover the modulated user symbols. For the most part of the papers (except papers G and H), we assume the uplink data transmission, where the BS detects user symbols. Similarly, in all of the papers except paper H, we assume a digital beamforming system where each antenna element is connected to an RF chain unit. Moreover, we assume that all the active users are in connected mode and have data to transmit. The performances are evaluated at a link-layer level (physical layer). We have investigated various types of channel models for simulating the wireless environment including uncorrelated and correlated Rayleigh fading models.

Chapter 3

Background and Contributions

In this chapter, first, we categorize the areas of contribution of the thesis into 4 groups and describe the research questions that each group tries to answer. Next, for each of the groups, we study the state of the art (SoA) and demonstrate the research gaps in the SoA and how the studies in each group aim to go beyond and solve the challenges. We finalize the discussion of each group with a short summary of the papers published in each group. We conclude the chapter by summing up the contributions of the thesis.

1 Grouping of the contribution areas

In order to discuss the logic behind classifying the types of the contributions in the thesis, we begin with discussing briefly the aim of the thesis. As mentioned in Chapter 2, our focus in this thesis is to answer 3 research questions. These questions are the concerns about: complexity performance trade-off, dealing with non-stationarities and scalable receiver design for the XL-MIMO systems. Thus, the challenge is to find signal processing techniques that are low complexity, scalable and less sensitive to the channel non-stationarities. In short, the main outcomes of the thesis are various distributed low complexity methods that are working better than other conventional techniques in the XL-MIMO channel models.

We arrange the areas of the contributions in 4 groups, depending on the type of problems we solved and the methods we exploit in each of them to answer the RQs. Fig. 3.1 shows a schematic of these areas and consequently, the papers within their corresponding categories. The first group discusses the main advantages and concerns of having an extra-large array and demon-

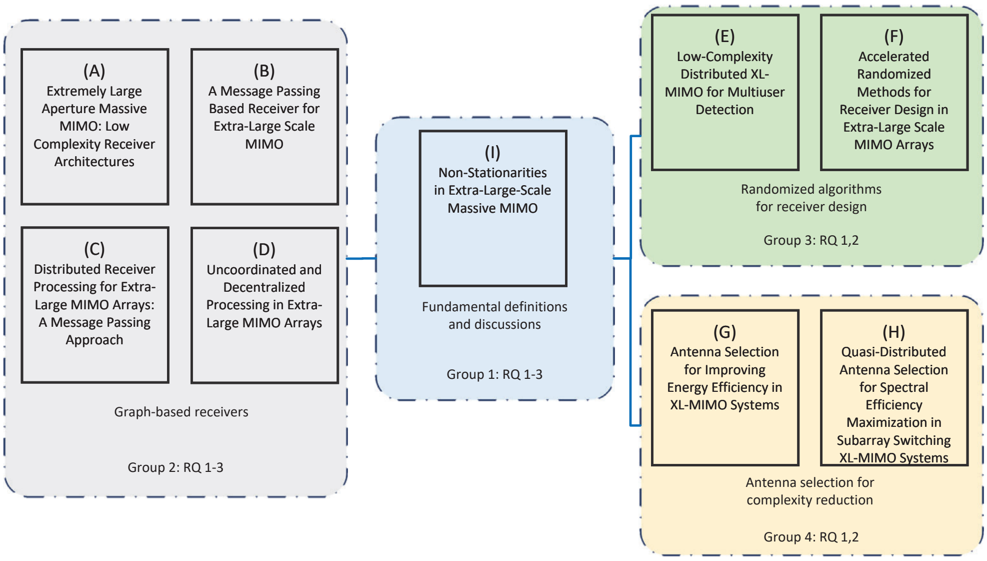

(A)
Extremely Large Aperture Massive MIMO: Low Complexity Receiver Architectures
(B)
A Message Passing Based Receiver for Extra-Large Scale MIMO
(C)
Distributed Receiver Processing for Extra-Large MIMO Arrays: A Message Passing Approach
(D)
Uncoordinated and Decentralized Processing in Extra-Large MIMO Arrays
Graph-based receivers
Group 2: RQ 1-3
(I)
Non-Stationarities in Extra-Large-Scale Massive MIMO
Fundamental definitions and discussions
Group 1: RQ 1-3
(E)
Low-Complexity Distributed XL-MIMO for Multiuser Detection
(F)
Accelerated Randomized Methods for Receiver Design in Extra-Large Scale MIMO Arrays
Randomized algorithms for receiver design
Group 3: RQ 1,2
(G)
Antenna Selection for Improving Energy Efficiency in XL-MIMO Systems
(H)
Quasi-Distributed Antenna Selection for Spectral Efficiency Maximization in Subarray Switching XL-MIMO Systems
Antenna selection for complexity reduction
Group 4: RQ 1,2

Fig. 3.1: Graphical representation of the contribution areas and the papers in each of the groups.

strating all three RQs in detail and proposing possible directions to deal with them. The second group focuses on all the RQs and uses graph-based techniques to lower the receiver complexity. The third group mostly targets RQ1 and RQ2 and exploits randomized methods to propose distributed receiver processing schemes. Finally, the last group pays attention to the problem of antenna selection using genetic algorithms to lower the computational costs and is concerned about RQ1 and RQ2.

In the following, we study each of the groups in detail by reviewing the related literature, pointing out the gaps in the state-of-art and then titles of the papers and their approach to handle the challenges.

2 Group 1: Fundamental definitions and modeling

As discussed above, this group introduces the benefits and challenges of the XL-MIMO arrays. It formulates the RQs and gives preliminary ideas and results on the possible ways to tackle the obstacles of such systems. In the following, we study several relative background topics for this group.

2.1 From MIMO to Massive MIMO

Scaling up the size of the conventional MIMO arrays was first mentioned in [6] and then more practical approaches were introduced in [15] and [16]. Authors in [4] list five main improvements of the M-MIMO compared to the MIMO arrays that are: 1. A 10-fold increase in the capacity and 100 times better energy efficiency at the same time. 2. the possibility of using inexpensive and low-power components 3. significant reduction in the latency on the air interface 4. simplified multiple access layer and 5. increased robustness to both man-made interference and intentional jamming. All of these points are in line with the main targets of 5G and can be exploited for different use cases [3].

One of the main features of the M-MIMO systems is their ability to suppress the interference caused by serving multiple users at the same time. This is a result of favorable propagation that is discussed next. The quantity of the interference between two users can be measured with an inner product between their channel vectors. With a large number of antennas at the BS, the channel vector of each of the users has many elements. Using random matrix theory and basic assumptions, it can be seen that while increasing the vector dimensions, the inter-user interference asymptotically approaches to zero [17]. However, this is only valid when the user channels have independent identically distributed (i.i.d.) complex Gaussian entries which do not happen often in practice [24]. On the other hand, adding more antennas to a BS array with a fixed aperture saturates the asymptotic orthogonality [25]. Therefore, an effective way to achieve favorable propagation is to increase the array aperture together with the number of antennas. In the next section, we discuss arrays with a larger aperture and their practical challenges.

2.2 MIMO systems with extra-large arrays

Following a similar approach of scaling up the spatial size of the BS array, we arrive at the XL-MIMO arrays. The XL-MIMO systems use antenna arrays with extremely large apertures to gain their high spatial resolution [26]. This indeed boosts the ability of the array to direct the beam to each of the desired users while in the meantime mitigating the inter-user interference. Moreover, employing a large number of antennas implies that cheaper antennas can be used in the array and their flaws can be compensated when they are operating together [27].

Here, we briefly talk about two other variants of the MIMO systems with a large number of antenna elements; large intelligent surface (LIS) and cell-free M-MIMO. An LIS can be seen as a planar array consisting of a large number of elements (often modelled as a continuous electromagnetically active area). The main idea of the LISs is their ability to carefully control electromagnetic fields in their environment that makes it possible to tightly focus energy in three-dimension space bringing entirely new capabilities [33]. The LISs have three different subgroups: active arrays that can detect the transmitting signals, passive ones that only act as a relay and hybrid LISs that use a part of the array for sensing and the rest of it for data transmission [28] [29]. The most attractive subgroup is the passive ones since they offer fundamental improvements to the wireless environment while consuming a very small amount of power. These arrays are also called intelligent reflective surfaces (IRS) that can be used for sensing and channel enhancement purposes [34]. The main issues with these arrays are the size of the sensed data together with the size of signalling overhead. For instance, for the channel enhancement applications, where the IRS is helping the M-MIMO BSs to achieve better channel conditions, the BS has to send all of its obtained channel gains to the IRS. The amount of this signalling overhead can be prohibitive in a multi-cell MIMO system. On the other hand, the cell-free M-MIMO can be seen as an extension to distributed MIMO systems, where a network of M-MIMO BSs are located in a large area and are connected to a cloud processor via high bandwidth back-hauls. The key point in these systems is that each individual BS forwards the received signals to the central processing unit and a global decision is made there. In other words, in the cell-free M-MIMO, the main idea is to use all of the BSs in a cooperative way to increase network performance metrics such as the SE and user fairness [35]. However, the main differences with the XL-MIMO array are in the array configuration and back-haul limitations. While in the XL-MIMO case the antennas are co-located and inter-connection delay is negligible, The cell-free systems try to deal with transferring pre-processed data of each of the BSs.

2.3 Spatial non-stationarities of XL-MIMO arrays

We start with one of the fundamental differences of the XL-MIMO arrays compared to other multi-antenna systems. With this type of array, the antenna elements occupy hundreds of wavelengths and therefore different parts of the array see different propagation environments. To be specific, the received signal at each part of the array comes from different scatterers causing uneven energy distribution along the array [26] [37]. Furthermore, spherical propagation of the electromagnetic waves conveys that the channel propagation features are dependent on the relative position of the users to the BS array and the size and magnetic properties of the antenna elements. Therefore, the antenna elements will have an unequal impact on the signal reception. As a consequence, the channel can not be considered wide sense stationary (WSS) in the spatial domain.

The concept of the Visibility regions (VRs) was first introduced in COST 2100 channel model [38]. The standard defines a VR as a terminal geographical area; meaning that when the terminal is located in this area, it can see a given set of clusters and this set of clusters associate with the VR. On the other hand, when the terminal moves out of that VR, it sees a different set of clusters and a different VR. In paper I, we extend the concept of VR to indicate a part of the BS array from which a given set of clusters is visible. Thus, we discriminate between VRs in the terminal domain VR-T and in the array

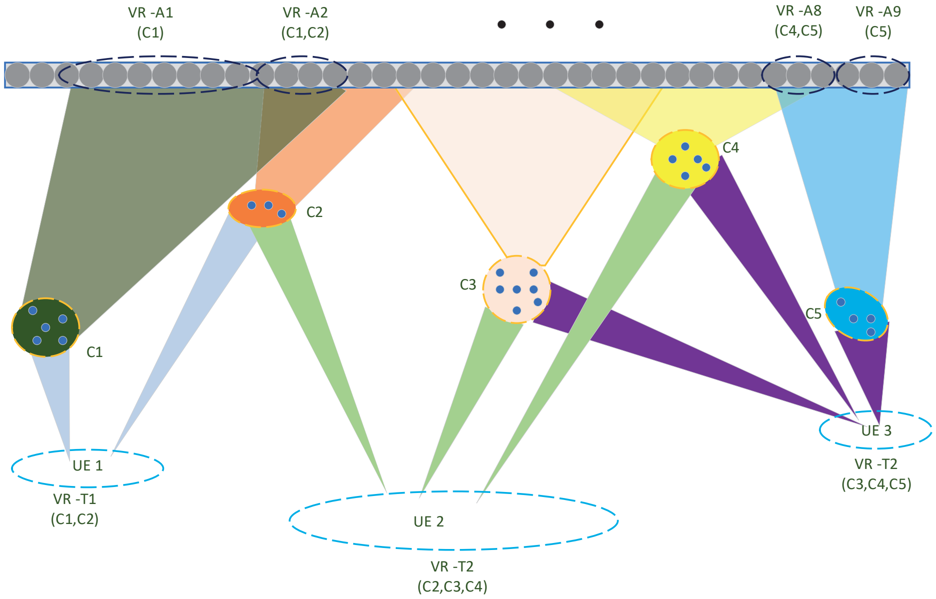


Fig. 3.2: An XL-MIMO array and VRs at both terminal and array domains.

domain VR-A. Fig. 3.2 illustrates an example of an XL-MIMO array where each user equipment (UE) is seeing a subset of the existing clusters (VR-T) and each portion of the array is visible to a group of the clusters (VR-A) [7]. In order to have a unified and compact definition of the VR, we will use the following description henceforth in the thesis: VRs are the portions of the array where most of each user's channel energy is concentrated.

One of the major drawbacks of having spatial non-stationarities is the performance loss experienced by the conventional linear receivers. For instance, authors in [39] show a significant performance degradation due to the non-stationarities compared to the WSS channels for most of the scenarios. The only case where a non-stationary channel helps boosting the linear receiver is when the VRs of the user are non-overlapping and thus, the inter-user interference is zero. Moreover, in Paper E we show the effect of the VR size in the performance of regularized zero-forcing (RZF) when the system load, i.e. number of antennas per user, changes. The results confirm a major increase in the bit error rate (BER) of the RZF receiver. One of the main conclusions from these results is that the linear receivers are not optimal for the XL-MIMO arrays for two reasons: first, their lack of ability to deal with non-stationary energy distributions and correlated channels and second, their impractical computational complexity when the number of antennas and users is very high.

2.4 Paper summaries

This group only contains one paper:
Paper I: (*published*) Elisabeth de Carvalho, Anum Ali, Abolfazl Amiri, Marko Angjelichinoski, Robert W Heath, “Non-Stationarities in Extra-Large-Scale Massive MIMO" , IEEE Wireless Communications Magazine 27, no. 4, pp. 74-80, IEEE, 2020.

The main idea in this paper is to study the primary differences of the XL-MIMO arrays with several conventional systems and discuss research possibilities and challenges in this topic. The paper can be seen as an introduction to very large array MIMO systems. The ideas and conclusions are backed up with several measurements and simulation results. Moreover, the paper suggests several research directions in the area of such arrays and considers the implementation concerns for them. It proposes that non-stationarity-aware and distributed solutions are the way to use XL-MIMO systems in practice.

3 Group 2: Graph-based receivers

This group focuses on all the RQs. Several graph-based receivers that utilize non-stationarities to cut the computational costs are proposed to address the existing challenges. The main idea is to present the user symbol detection problem with a graph and then use graph-based techniques to solve the problem. We begin this part with a short discussion about the problem of high computational complexity and message-passing based techniques and then we review some of the literature that are using graphs for the detection problems. Next, we introduce the papers in this group with a short summary. The papers in this group provide a wide range of schemes each with different performance-complexity characteristics.

3.1 Curse of dimensionality and receiver types

Having an array with hundreds or thousands of antenna elements requires a powerful central processing unit (CPU) that can manage the complexity of the implemented receive combining method. This complexity is a function of the number of the active users K and antenna elements at the BS M and the user equipment N_u for the linear methods. In general, the received signal in the uplink (UL) of a multi-user MIMO system with single antenna users is modelled as:

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n} \tag{3.1}$$

where, $\mathbf{y} \in \mathbb{C}^{M}$ is the received signal vector, $\mathbf{H} \in \mathbb{C}^{M\times K}$ is the channel matrix, $\mathbf{x} \in \mathbb{C}^{K}$ is the transmitted signal vector and $\mathbf{n} \in \mathbb{C}^{M}$ is the noise vector at

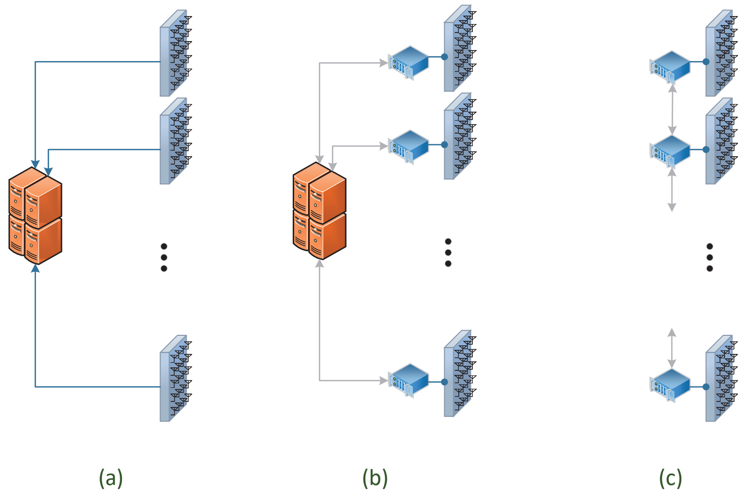


Fig. 3.3: Three different receiver design architectures for the XL-MIMO arrays. (a) fully centralized mode, with a single CPU, (b) hybrid mode with a CPU and several LPU at each sub-array and (c) distributed mode with only LPU connected to their neighbors.

the receiver. The main objective of the receivers in this thesis is to recover the user symbols[1] x_k, $k = 1, \dots, K$. For instance, linear receivers try to find a combiner matrix $\mathbf{F}$ to obtain estimates of the symbols as $\hat{x} = \mathbf{F}\mathbf{y}$. As an example, for the ZF combiner, this matrix is $\mathbf{F} = (\mathbf{H}^H\mathbf{H})^{-1}\mathbf{H}^H$, with $(\cdot)^H$ denoting matrix conjugate transpose operation. The design of the linear combiner matrices can be done in three different ways; centralized, distributed and hybrid modes. Fig. 3.3 illustrates the difference between each of these designs. In the following, we discuss each of the modes.

The original derivation and implementation of the linear receivers are all in the centralized mode. These designs demand a single processing node that collects the signals from all the elements and applies the matrix operations at the same time. The only linear centralized receiver that has a linear complexity growth with respect to M and K is the maximum ratio combiner (MRC). However, this method is unable to cancel the interference between the users and therefore, has a poor performance in most of the multi-user scenarios. On the other hand, other linear receivers such as the ZF and minimum mean square error (MMSE) receivers use matrix inversions to cope with the inter-user interference which adds non-linear terms[2] to the complexity expression [14]. With a lot of antenna elements at the XL-MIMO BS together with many active users, the absolute value of the flops needed for the receiver processor is much bigger than any other M-MIMO system [40]. Therefore, us-

[1] As discussed in the previous chapter, we omit the channel coding in this thesis and thus, the receivers are only recovering the user symbols instead of detecting the transmitted information bits.

[2] Third and second-order relation with the number of users

ing a centralized technique is not desirable for the XL-MIMO case due to the high computational capacity requirement and large processing delay times.

With the aim of dealing with the problem of the centralized methods, two other modes can be employed: 1. distributed methods that only rely on local processing units (LPU) and 2. hybrid methods that use both the CPU and LPUs for signal processing tasks. The distributed receivers for the multi-antenna systems are made of several local nodes that are connected to a subset of the antenna elements called sub-arrays. The extreme case is when each of the antennas has a simple processor and the size of the sub-array is 1 [41]. The main idea in these types of receivers is to obtain several local estimates of the user symbols and then try to refine these estimates by partial existing connectivity between the LPUs. For instance, authors in [41] use a method called daisy-chain to approximate the ZF by cancelling the user interference by processing each of the antennas in series. One major drawback of the daisy-chain techniques is that they weigh the contribution of all the antennas equally, while in the XL-MIMO arrays, the spatial non-stationarities change this equality and degrade the performance. Therefore, there is a need for accounting for the non-stationarities in the design of the receivers.

In order to overcome the issues of the fully distributed receivers, the hybrid methods use a CPU to orchestrate the receiving tasks between the local nodes. The main idea here is to combine the local estimates in the best way possible (which in some cases is not a low-complexity solution) to improve the results. For example, in [42], authors use a hierarchical structure to update the local estimates of the sub-arrays where the CPU calculates the global symbol estimates and then sends it back to the LPUs at each step of their proposed heuristics. In paper C, we use a two-tier receiver that uses both data fusion and successive interference cancellation (SIC) at the CPU and then propagates the updated estimates back to the sub-arrays until all the user symbols are detected. One of the concerns in designing this type of architecture is to handle additional signalling overhead between the nodes, keeping the complexity per node in an acceptable range[3] and limiting the back-haul resources needed to connect all the processors.

3.2 Graph-based and Message-passing based receivers

Using graphs to represent interconnected and complex problems is one of the ways to help visualise the model better and eventually solve it efficiently. For instance, authors in [52] used a graph-based access protocol for M-MIMO systems. Their graph representation models the pilot collision in the access phase of a multi-user scenario. This model helps in realizing the colliding

[3] The main reason to use multiple processing units is to have a much lower load in each of them enabling the use of inexpensive units.

users and makes it easier to implement an effective method to solve the problem. The graph-based methods are widely used in the physical layer problems as well. In [53], a graph-based receiver that iteratively performs soft channel estimation and data detection is presented. The graph shows the correlation between different nodes in the receiver. A novel detection algorithm for MIMO communication systems employing Gaussian trees is presented in [54]. The graph model in the original problem is very loopy. Thus, the authors manage to use tree approximation methods to obtain cycle-free discrete symbol distributions. As discussed before, two of the main objectives of this thesis is to design low-complexity receivers while taking the channel non-stationarities into account. Therefore, inspired by the ability of the graph-based techniques in collision resolution problems, we modelled our multi sub-array multi-user system with a bipartite graph. The idea is to model the non-stationarities with the graph and then use it to iteratively detect the users. See the details of this technique in paper A.

One of the main methods that has been widely used in the context of the M-MIMO systems to lower the complexity is the message-passing methods. In this thesis, we pay special attention to variational Bayesian inference-based and message-passing algorithms [43]. The inference frameworks try to solve approximate probability density/mass functions of the variables that are very hard to solve directly. These frameworks are recognized with two components; beliefs $q(\mathbf{x})$ that are approximating the desired probability function of $p(\mathbf{x})$ and an objective function $F(q)$ that measures the discrepancy of the approximation and the desirable function [44]. Two of the major approaches for variational Bayesian inference are the mean-field (MF) approximation and belief propagation (BP).

We start with describing the MF method. Assume we have a fully Bayesian model where prior distributions of all parameters are given. Moreover, the model can have both latent variables and parameters, and we use $\mathbf{z}$ to show the set of all latent variables and parameters. In the same manner, we denote the set of all observed variables by $\mathbf{x}$. We assume a probabilistic model that specifies the joint distribution $p(\mathbf{x}, \mathbf{z})$, and our goal is to find an approximation $q(\mathbf{z})$ for the posterior distribution $p(\mathbf{z}|\mathbf{x})$ as well as for the model evidence $p(\mathbf{x})$ [45]. Moreover, minimizing the objective function is equivalent to minimizing the Kullback-Leibler divergence between q and p which is defined as $D(q||p) = \sum_{\mathbf{z}} q(\mathbf{z}) \ln(\frac{q(\mathbf{z})}{p(\mathbf{z}|\mathbf{x})})$ [46]. Allowing any possible choice for $q(\mathbf{z})$, then the best approximation happens when the KL divergence vanishes, which occurs when $q(\mathbf{z})$ equals the posterior distribution $p(\mathbf{z}|\mathbf{x})$. However, we restrict the class of functions that we allow for $q(\cdot)$ since working with the true posterior distribution is intractable. One of the famous approaches which is called Naive MF assumes a fully factorized model such as $q(\mathbf{z}) = \prod_{i=1}^{M} q_i(z_i)$, where we partition the elements of $\mathbf{z}$ into disjoint groups of denote by z_i and $i = 1, \cdots, M$. Next, we discuss the BP approach to the approximation problem. In the BP methods, we try to calculate the marginals $q_i(z_i) = \sum_{\mathbf{z} \setminus z_i} p(z)$[4] instead of trying to approximate the full probability function. In BP, we iteratively try to approach the stationary points of the Bethe free energy (unless the underlying factor graph has a tree structure, there is no guarantee that BP will lead to the stationary points) [47].

Both the BP and the MF concepts can be expressed via message-passing algorithms in factor graphs. Factor graph [48] is a tool for graphical representation of a probabilistic model. The message-passing interpretation of the BP and the MF principles are known as the sum-product (SP) algorithm and the variational message-passing (VMP) algorithm, respectively [55].

The SP algorithm can find the exact marginal distribution of the factor graphs without loops. Otherwise, the outcome is an approximation of the desired marginal. The overall strategy is simple message passing; first, we need to form a rooted tree at z_i to compute $q_i(z_i)$. Then, we take the product of descendants at every variable node. Also, for the factor nodes, we take the product of the factor with descendants followed extrinsic sum over the parent of the factor. The SP algorithm has been used in 5G new radio standard for LDPC codes [49].

Similarly, for the MF approximation, the VMP algorithm is minimizing the variational free energy. However, the computation of the exact marginals is not guaranteed [47]. The convergence of the algorithm is assured at each step of the algorithm by confirming that the variational free energy of the computed beliefs is non-increasing. The format of the messages in the VMP algorithm is making it suitable for conjugate-exponential probabilistic models in the wireless communication applications allowing the acquisition of closed-form expressions in many cases. We would like to refer the readers to the papers in this group for more details on the message derivations of both MF and BP frameworks.

There are several works that are using message-passing methods to reduce the receiver complexity in the M-MIMO systems. For instance, [51] proposes a low complexity detector for the M-MIMO systems that uses expectation propagation (EP). The solution suits well with high-dimensional systems with high-order modulations. However, the authors try to come up with reduced-complexity variants of EP to avoid large matrix inversions within the algorithm. A novel approximate probability updating scheme is proposed in [56] that tries to lower the complexity of the message-passing receiver when the number of users or the order of modulation increases. Message-passing methods are also used to compensate some of the hardware non-idealities. As an example, in [58] the authors introduce a low complexity method for an M-MIMO array with low-resolution analog digital convert-

[4] The expression $\mathbf{z} \setminus z_i$ is the set exclusion operation denoting all components of $\mathbf{z}$ except z_i.

ers (ADC) at each antenna. They combine generalized approximate message passing detection with channel decoder to ensure that the information filtered by the ADCs can be recovered as accurate as possible. However, as discussed in the previous chapters, the challenges in the XL-MIMO arrays are different from the M-MIMO ones. Therefore, direct use of the aforementioned methods in the context of the XL-MIMO arrays will not result in the same performance characteristics. Thus, there is a need to develop methods that are compatible with distributed nodes that can talk to each other and update their local information. Furthermore, the effect of non-stationary channels on the performance of such receivers needs to be studied. Motivated by these shortcomings in the literature, we aim to design several message-passing based techniques that have acceptable performance in the XL-MIMO systems, which are presented in papers B,C and D.

3.3 Paper summaries

There are four papers in this category:
Paper A: (*published*) Abolfazl Amiri, Marko Angjelichinoski, Elisabeth de Carvalho, Robert W Heath, “Extremely Large Aperture Massive MIMO: Low Complexity Receiver Architectures" , 2018 IEEE Globecom Workshops (GC Wkshps), pp. 1-6. IEEE, 2018. **Paper B:** (*published*) Abolfazl Amiri, Carles Navarro Manchón, Elisabeth de Carvalho, “A Message Passing Based Receiver for Extra-Large Scale MIMO", 2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), pp. 564-568. IEEE, 2019.
Paper C: (*accepted*) Abolfazl Amiri, Sajad Rezaie, Carles Navarro Manchón, Elisabeth de Carvalho, “Distributed Receiver Processing for Extra-Large MIMO Arrays: A Message Passing Approach", Accepted for publication in IEEE Transactions on Wireless Communications.
Paper D: (*accepted*) Abolfazl Amiri, Carles Navarro Manchón, Elisabeth de Carvalho, “Uncoordinated and Decentralized Processing in Extra-Large MIMO Arrays", Accepted for publication in IEEE Wireless Communications Letters.

Chronologically, Paper A is the first work on the XL-MIMO arrays that focuses on the distributed receiver designs. The motivation of the paper is to propose a better receiving technique that is aware of users’ channel energy non-stationarities. In order to deal with high computational complexity, the paper assumes that the array is composed of many smaller sub-arrays that have a local processing unit. These sub-arrays are in charge of estimating the user symbols locally and then, they forward their information to the central node for data fusion and a global decision. The paper presents a graphical method that uses the energy distributions to construct a bipartite graph that is showing each user’s dominant sub-arrays, the sub-arrays that have

95% of that user's energy. Then, inspired by a collision resolution method from access protocols, the proposed algorithm finds the best schedule of user detection in a successive interference cancellation (SIC) scheme.

In the following, we continue with an overview of the rest of the papers and then mention the main contribution of each of them separately. The main idea of these three papers is to replace conventional linear receivers such as ZF with the VMP receivers. Since the channels are Gaussian, close-form expressions for the messages can be obtained. These VMP receivers can either process the signal from all the array elements or only the signal from a sub-array. The best performance is of course for the first case, but in order to have a more scalable solution, the second option can be used. Then, mixing the local decisions of the sub-arrays and using interference cancellation methods can boost the detection performance to an acceptable level. Another advantage of having a sub-array based architecture is that the contribution of those with lower received energy (due to the non-stationarities) can be adjusted in the final decision. In other words, a non-stationarity-aware receiver can be implemented to reduce the complexity of the central receiver.

Paper B, suggests a centralized VMP receiver design for the XL-MIMO system in crowded scenarios that has a linear complexity behaviour. Next, an extended work to a hybrid architecture is presented in paper C, where local VMP receivers (LPUs) are used to get the local estimates. Moreover, different options are available for the CPU and LPU units to distribute the processing tasks among them. The CPU can fuse the local estimates and apply a SIC procedure on top to enhance the performance. Finally, paper D, uses a combined VMP-BP method in a fully decentralized XL-MIMO system (without a central unit) to deal with the scalability issues of the BS array. The VMP is used for the LPUs while the data exchange process between the LPUs is done using the BP.

4 Group 3: Randomized algorithms for receiver design

In this section, we study another way of lowering the complexity of the XL-MIMO receivers with the use of randomized techniques. Particularly, we are investigating the use of Kaczmarz algorithms in this group and mostly addressing RQ1 and RQ2 here.

4.1 Kaczmarz algorithms

As mentioned before, the important reason for not choosing the centralized linear receivers for the XL-MIMO arrays is their high computational complexity. For instance, the regularized zero-forcing (RZF) technique uses a matrix inversion that is very costly, especially when the number of users and antennas is high. This symbol estimation of the RZF receiver is calculated by $\hat{\mathbf{x}}^{\text{RZF}} = \left(\mathbf{H}^H\mathbf{H} + \xi\mathbf{I}_K\right)^{-1}\mathbf{H}^H\mathbf{y} = \mathbf{F}^{\text{RZF}}\mathbf{y}$, where $\mathbf{F}^{\text{RZF}}$ is the linear combiner matrix, ξ is an arbitrary regularization coefficient[5] and $\mathbf{I}_K$ is an identity matrix of size K. One can reformulate this receiver as a solution of the following optimization problem [50]:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} ||\mathbf{H}\mathbf{w} - \mathbf{y}||_2^2 + \xi||\mathbf{w}||_2^2 \tag{3.2}$$

with $\mathbf{w}^*$ corresponding to $\hat{\mathbf{x}}^{\text{RZF}}$. The cost function can be written in a more compact way as $||\mathbf{B}\mathbf{w} - \mathbf{y}_0||_2^2$, where $\mathbf{B} = [\mathbf{H}; \xi\mathbf{I}_K]$ and $\mathbf{y}_0 = [\mathbf{y}_0; \mathbf{0}_{K\times 1}]$[6]. The solution to $\mathbf{B}\mathbf{w} - \mathbf{y}_0 = 0$, which is a system of linear equations (SLE), can be obtained with various techniques. One way to keep the complexity controlled is to select a subset of these equations instead of solving the full SLE. This indeed reduces the accuracy of the solutions, but with a good selection of the equations and the algorithm the performance gap can approach zero. One of these methods is Kaczmarz algorithms (KA) that were first introduced by a Polish mathematician Stefan Kaczmarz and have been widely used in signal processing applications [59].

The authors in [60] used another variant called randomized Kaczmarz (rKA) that has a much better convergence and complexity behaviour than the original version. In summary, the rKA can be seen as a particular case of stochastic gradient descent. The algorithm chooses the equations in a random way based on a predefined probability distribution. Then, it computes a residual that is an orthogonal projection of the last iterative solution onto the solution hyper-plane. Then, this residual is normalized by the energy of the chosen equation and the algorithm continues until a stopping criterion is met. This algorithm has been used in detection problems of MIMO systems [50, 61, 62]. In general, these papers try to describe the data detection problem in the MIMO and M-MIMO systems with an SLE and use variants of KA to approximate an exact linear detection technique (such as RZF) to develop a low complexity solution. However, we believe their extension for an XL-MIMO array needs more investigations to adapt the KA methods for such arrays.

The three key reasons that make these algorithms interesting for the XL-MIMO are : 1. simple calculations that are done in vector and scalar multiplications instead of matrix operations. 2. A graceful degradation that gives a fine balance between the complexity and accuracy and can be managed very easily. 3. The randomization function can exploit the non-stationarities of

[5] One suitable option is to choose the inverse of the pre-processing user transmit signal to noise ratio (SNR) for ξ which will resemble the linear minimum mean square error (LMMSE) receiver.

[6] The expression $\mathbf{0}_{K\times 1}$ is denoting a vector of zeros of size K.

the channel in a straightforward way. In other words, the importance of the users can be seen as the size of their VR and the stronger users are selected for detection first. Moreover, the algorithm does an implicit interference cancellation after each projection resulting in improved performance compared to other iterative methods. In the following, we introduce two papers where we evaluate the use of randomized methods for the XL-MIMO systems.

4.2 Paper summaries

This group has two papers:
Paper E: (*published*) Victor Croisfelt Rodrigues, Abolfazl Amiri, Taufik Abrão, Elisabeth de Carvalho, Petar Popovski, “Low-Complexity Distributed XL-MIMO for Multiuser Detection", 2020 IEEE International Conference on Communications Workshops (ICC Workshops), pp. 1-6. IEEE, 2020.
Paper F: (*published*) Victor Croisfelt Rodrigues, Abolfazl Amiri, Taufik Abrão, Elisabeth de Carvalho, Petar Popovski, “Accelerated Randomized Methods for Receiver Design in Extra-Large Scale MIMO Arrays", , IEEE Transactions on Vehicular Technology vol. 70, no. 7, pp. 6788-6799, IEEE, July 2021.

Both of the papers propose low-complexity and distributed approximations of the ZF and RZF receivers. The main idea is to use the spatial non-stationary information to simplify the aforementioned linear receivers. Moreover, the proposed receivers are implemented in a distributed way which is practical for the XL-MIMO systems with sub-arrays. These receivers use the rKA to mimic the RZF and utilize VRs to fine-tune the randomness design of the rKA. A novel equation selection probability function is introduced that is using the VRs and normalized user energies to prioritize the stronger users first. This algorithm works at each of the sub-arrays to approximate the combining matrix of that sub-array. Then, all these matrices are concatenated together at the CPU to find the global receiver. One of the key features of the rKA is the graceful degradation manner of it allows the designer to handle the complexity- performance trade-off only by changing the number of algorithm iterations.

In paper E, the use of rKA for the context of XL-MIMO arrays is introduced, where the randomization function uses VR-related information to obtain better results than uniform random selection. Aiming to extend the contributions of paper E, paper F uses several acceleration techniques for further complexity reduction of the receivers. The idea is to use these techniques to overcome the convergence issues and help boost the convergence speed and eventually reduce the computational cost of the updated algorithm. The main takeaway of the paper is the ability of the three proposed methods to adjust the computational cost of the XL-MIMO receiver and offer several quality of service options.

5 Group 4: Antenna selection for complexity reduction

The main approach in this group is quite different from the rest of the groups. While most of the groups try to reduce the receiver complexity in the uplink, this group is analyzing the downlink transmission and beamforming techniques. Most of the focus of this group is on RQ1 and RQ2. The main idea here is to find a good antenna selection (AS) scheme that can cut the complexity of the transmitter at the BS. We start with an introduction to the necessity of AS solutions and then examine two widely used approaches. We conclude the section by summarizing the papers of this group.

Before discussing any details, we would like to clarify our definition of the AS in this thesis. Our AS methods provide a set of candidate antenna set $\mathcal{A}$ that is a subset of all the antenna elements in the XL-MIMO array. This selection is done based on an optimization solution and is maximizing a utility function that we will introduce in the following. Earlier in this chapter, we mentioned that the complexity is a function of the number of users K and the number of antennas M. So far, we reviewed several methods to propose receiver design functions with complexities that grow slower with respect to these two parameters. We also introduced the distributed and hybrid methods that are composed of smaller antenna clusters called sub-arrays. Now, we consider the downlink transmission in an XL-MIMO system. The main idea here is to first restrict M for each of the LPUs to deal with a lower number of antennas. Moreover, due to the user energy non-stationarities, the number of active users will decrease at each sub-array[7]. In other words, each sub-array will have a certain number of dominant users (smaller than the total number of active users in the system) that have an almost spatially stationary channel over the elements of that sub-array. The problem of AS has been studied vastly for the M-MIMO systems mostly because of the hardware limitations of the fully digital implementations [63]. There are several techniques that are employed to select the best antennas based on some predefined cost function such as power consumption or RF chain connections [64]. For instance, [65] selects an optimal channel sub-matrix with the largest minimum singular value. [66] and [67] use machine-learning based techniques to maximize the energy efficiency of the array. In the following, we study several methods for the AS in the XL-MIMO arrays.

[7] We assume that the impact of the users that have less than a certain percentage of their energy on a specific sub-array can be ignored on that sub-array without a major performance loss.

5.1 Simple AS techniques

One of the major concerns about the AS problem in the context of the XL-MIMO systems is to control the complexity of the AS solution. In other words, the computations needed for the AS problem plus the transmitter block should be less than the one needed for a centralized method; otherwise, the objective of complexity reduction will be violated. The simplest AS method is a fixed division method where the array is divided equally into B sub-arrays and at each instance, a subset of these sub-arrays will transmit. This method has almost no complexity burden but it is not optimal in the sense of optimizing the EE and SE, especially for the XL-MIMO systems where the distribution of the users' energies is not uniform. Another simple technique is to use the channel matrix or second-order statistics of the channel to find the parts of the array, let us call them effective antennas, that have the most of the user energies. There are two important issues with this method: 1. the effective antennas for each of the users are different and we might end up selecting the whole array by choosing each users' part. 2. The effect of interference is neglected in this solution, meaning that an effective area for a user can cause and unaccounted interference to other users, resulting in a poor symbol detection outcome. Therefore, we seek to find smarter methods that consider the shortcomings of these two simple schemes.

5.2 Genetic algorithm for the AS problem

The genetic algorithm (GA) is one of the famous bio-inspired methods that has been widely used for optimization problems [68]. The main idea consists of a simulation of natural selection. In short, the GA is a local search technique that tries to find an approximate solution for optimization and search problems.

The implementation of the GA contains the following phases:

1. *Elitism:* The elitism intends to keep the best individuals of the current generation without any change. It ensures that the score of the best individual over the generations is non-decreasing.

2. *Tournament selection:* During the tournament selection, the score of randomly chosen individuals are compared in pairs.

3. *Crossover:* The crossover phase aims to mix chromosomes of the winners of the tournament in order to get new solutions, with a focus on exploring the search space. Note that, the chromosomes are a set of optimization variables for any candidate solution.

4. *Mutation:* The mutation phase intends to add random small changes to the generated offspring by the crossover step. This phase increases

the variability among the individuals, exploring different regions of the feasible set.

The main motivations for using this algorithm for the AS problem in the context of the XL-MIMO array are threefold: 1. the algorithm has low complexity and uses pay-off information instead of derivatives. 2. It can easily be used in paralleled implementations. 3. It support multi-objective optimization problems [69]. Therefore, this algorithm can give us a set of the best antennas that can maximize for example, the energy efficiency of the system. In this group of papers, we are particularly interested in the spectral and energy efficiencies for the XL-MIMO arrays. We try to look deeper into the effect of non-stationarities on the AS problem, which is missing in the literature and exploit it to achieve even lower complexity AS solutions. In the following, we introduce the papers that explore the GA method for the AS problem in the XL-MIMO arrays.

5.3 Paper summaries

This group has two papers:
Paper G: (*published*) José Carlos Marinello, Taufik Abrão, Abolfazl Amiri, Elisabeth De Carvalho, Petar Popovski, “Antenna Selection for Improving Energy Efficiency in XL-MIMO Systems", IEEE Transactions on Vehicular Technology 69, no. 11, pp. 13305-13318, IEEE, 2020.
Paper H: (*published*) João Henrique Inacio de Souza, Abolfazl Amiri, Taufik Abrão, Elisabeth de Carvalho, Petar Popovski, “Quasi-Distributed Antenna Selection for Spectral Efficiency Maximization in Subarray Switching XL-MIMO Systems", in IEEE Transactions on Vehicular Technology, vol. 70, no. 7, pp. 6713-6725, IEEE, July 2021.

These papers, as it is obvious from their titles, propose several AS techniques for the XL-MIMO arrays in the downlink transmission. The utility function is EE in the first one and is SE in the second one. The proposed optimization solutions target two main concerns with the XL-MIMO arrays: 1. the energy consumption due to activation of the RF chains and 2. the use of energy for computational tasks. Papers discuss the cases, where there are much fewer active users in the system than the number of antennas and their signals will only cover a part of the array. Thus, exploiting a good antenna selection strategy can help reduce the energy usage and shed a light towards more green solutions.

Since the problem of AS is a combinatorial optimization problem, papers G and H, use the GA to solve the AS problem subjected to energy efficiency and spectrum efficiency maximization constraints, respectively. The solutions in paper G aim to minimize the energy consumption at the BS considering the power used at the RF chains and also for the signal processing purposes.

Several AS schemes are proposed that work in a centralized manner. On the other hand, in paper H, the focus is to find distributed techniques to find the best active antenna set that guarantee a sum-rate that is very close to that of the centralized receivers. Moreover, the problem of power allocation is considered jointly with the SE maximization to have more realistic constraints for the optimization problem.

6 Summary of the thesis's contributions

In this section, we summarize the main novelties of this thesis. So far, we have discussed the newness in each of the paper groups and here, we try to put them together and give a higher-level perspective of the contributions done in the thesis. These contributions are addressing the RQs introduced in Chapter 2. In the following, we try to further discuss the technical contents in a compact way.

6.1 Low complexity and scalable algorithm designs

The main concern about the practical implementation of the XL-MIMO systems is the required computational capacity for them to operate properly. In this thesis, we target this issue and come up with several techniques to alleviate the computational cost of signal reception at the XL-MIMO BS. We used three approaches in this regard: distributing the processing tasks, using simpler non-linear reception methods and limiting the processing area size by antenna selection techniques. We employed several decentralizing schemes that use local nodes at each sub-array for a part of the signal reception. Then, these nodes either communicate with each other or forward their estimates to a central processing unit for further decisions. One major outcome of these processing methods is their flexibility to control the complexity-performance trade-off. Further, we used several low-complexity heuristics such as message-passing and Kaczmarz tools to approximate the exact symbol detection process. This approximation results in a much lower computational complexity compared to the centralized linear processing methods with the cost of degraded performance which we try to minimize. Last but not least, the antenna selection techniques that mostly use genetic algorithms aid the complexity reduction. We focus on several antenna selection criteria to target complexity reduction, RF chain power consumption and sum-rate maximization. Using the distributed schemes, gave us the ability to scale up our solutions to any size of the array, without the need for a complete modification of the algorithms. Our multi-node processing tools can be used in any other system that try to have scalable designs.

6.2 Spatial non-stationarity aware receivers

One of the main differences of XL-MIMO arrays with ordinary M-MIMO ones is the appearance of visibility regions due to uneven energy distributions over the array. In this thesis, we introduce numerous methods to first recognize these VRs and then incorporate them in the receiver design process. We question the applicability of conventional linear receivers for the XL-MIMO arrays and show that they suffer from a substantial degradation due to non-stationarities of the channel. Our receivers try to exploit the VRs and employ interference cancellation techniques using the natural signal separation made by the VRs. In other words, non-overlapping parts of the VRs , that are less affected by the interference, can be used to recover stronger user signals and then a SIC receiver can recover the signals from the overlapped parts. Furthermore, we show that with smart VR-aware methods the computational costs can be cut considerably while the system performance stays almost the same.

Chapter 4

Concluding remarks

This chapter aims to wrap up the main content of the thesis by reviewing the key benefits of employing XL-MIMO arrays together with the signal processing techniques introduced here. Further, different possible applications and future research directions and extensions are provided.

1 Conclusions

In this section, we begin with a general review of the findings of the thesis. Then, we revisit the research questions that were asked at the beginning of the thesis and then show how our studies answer/ shed some light on them.

This thesis focuses on providing several signal processing methods for the XL-MIMO systems. It argues with a common understanding that the conventional linear receivers, that operate close to optimal for the M-MIMO systems, can also be used for the XL array as well. The thesis disagrees with this idea and tries to first find evidence that usual M-MIMO reception techniques fail greatly in the XL-MIMO scenarios and then, studies various methods to combat the challenges.

In order to answer RQ1, the proposed receiver designs combine different strategies to deal with the high number of antenna elements at the BS and also the channel non-stationarities. They make use of distributed and parallel processing techniques to divide the computational burden into manageable smaller tasks that can be done in remote and cheaper units. Each of these processing units is connected to a sub-array that can have either a fixed or dynamic size. Most of the receivers introduced in the thesis are working with the first case while the antenna selection methods propose new ways to assign the sub-arrays. One remark that should be considered for any type of antenna selection technique is that the overhead complexity of this selection must not make the whole process more complex than a full-array selection unless the designer wishes to limit the power consumption. In short, in response to RQ1, our methods offer performances close to those of classical receivers with feasible computational complexities.

Regarding the second RQ, our studies first analyze the effect of user channel energy non-stationarities in the XL-MIMO systems, and then, try to exploit it to design efficient algorithms. It was believed that the capacity of the M-MIMO system can grow boundlessly if the number of antennas at the BS goes to infinity. However, this thesis is looking into more realistic scenarios where the channel correlations and energy non-stationarities do not allow the above statement. After recognizing the concern of uneven energy patterns, the thesis aims to utilize the information about the visibility regions to design smarter algorithms. The main idea is to use this information as a tool to cut the complexity in the first place and then try to improve the performance by means of interference cancellation techniques.

Last but not least, in response to RQ3, our proposed distributed methods can provide scalable solutions for the XL-MIMO arrays. In other words, they use several multi-node processing techniques that can handle any size of the system. The key point is to divide the central processing problem into many sub-problems (in the local units) and solve them and then, find a way to connect all these sub-solutions to each other. One main challenge is to have control over the added information exchanges between the local units and the central unit. The solutions mentioned in the thesis try to keep the overheads very low. For instance, instead of transmitting the raw received signals (that scale with the number of antennas at each local unit), they send their local estimates which scale with the number of users. This indeed has a much lower size and will not grow with the number of antennas in the XL-MIMO array.

So far, we have discussed a lot about the advantages of the methods in the thesis and here we mention some of the disadvantages of the methods as well. These points can be seen as a motivation to continue the topics studied in this thesis with other techniques. One of the main shortcomings of the designed receivers is their dependency on accurate channel information and imperfect channel estimates can substantially degrade the performances. One way to compensate this effect is to account for more robust techniques that account for such imperfections. For the graph-based techniques, the modulation order can play a crucial role in defining the complexity. While we mostly used orders of 4 and 8, extending the works to higher orders might encounter a prohibitive amount of computations. A feasible approach to tackle this issue would be the use of continuous modulated symbol space instead of discrete modulated points to limit the complexity. Regarding the heuristics used in the third and fourth groups of the contribution areas, the problem and convergence can be critical. Especially, when the selection of the starting points of the algorithms is not done properly.

2 Applications

Here, we try to give a few examples of possible applications of our findings in current or future multi-antenna systems. The first obvious application is for the extra-large size multi-antenna systems. As an example, a refined version of stripe-line antennas is introduced to improve coverage [70]. These arrays are composed of thousands of cheap antenna elements and can be deployed very easily. Our proposed methods can be applied for these types of systems where the number of elements is huge, the aperture size is big and computational power is limited. In general, the methods of this thesis can still be adopted for any M-MIMO system that has a limited processing capability. Our proposed decentralizing methods can inspire different hybrid architectures in the M-MIMO arrays each covering a range of applications. Moreover, early implementations of the LISs that still use many antenna elements can benefit from our techniques for signal reception and also antenna selection schemes for power management considerations.

3 Future work directions

There are several possibilities to extend the work done in the thesis. The first category is to use more realistic channel models and also consider the CSI acquiring phase in the designs. Therefore, new techniques can be developed that are robust to partial and imperfect CSI. Moreover, the channel coding can be directly embedded within the message-passing methods to boost the performance. On the other hand, more focused research should take care of implementation challenges such as hardware imperfections, backhaul link requirements and local unit capabilities. Another relevant problem is the scheduling of the users in the XL-MIMO systems especially using some VR-related side information. The problem gets more interesting if the user devices can perform beamforming to redirect their signal to non-overlapping areas of the array to maximize their rate. Finally, the use of the machine-learning technique is also an option in different parts of the XL-MIMO systems. Such tools can predict the non-stationary pattern changes and adapt the receiver based on that. Further, they can be utilized for other tasks such as user-sub-array assignment, beamforming design and power control problems.

References

[1] "Where is 5g available?" https://itchronicles.com/5g/where-is-5g-available/, accessed: 2021-03-28.

[2] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, "5g wireless network slicing for embb, urllc, and mmtc: A communication-theoretic view," *IEEE Access*, vol. 6, pp. 55 765–55 779, 2018.

[3] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5g," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 74–80, 2014.

[4] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive mimo for next generation wireless systems," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 186–195, 2014.

[5] S. M. Alamouti, "A simple transmit diversity technique for wireless communications," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 8, pp. 1451–1458, 1998.

[6] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Transactions on Wireless Communications*, vol. 9, no. 11, pp. 3590–3600, 2010.

[7] E. De Carvalho, A. Ali, A. Amiri, M. Angjelichinoski, and R. .W Heath," Non-Stationarities in Extra-Large-Scale Massive MIMO," *IEEE Communications Magazine*, vol. 27, no. 4, pp. 74–80, 2020.

[8] A.-S. Bana, G. Xu, E. De Carvalho, and P. Popovski, "Ultra reliable low latency communications in massive multi-antenna systems," in *2018 52nd Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2018, pp. 188–192.

[9] A.-S. Bana, E. De Carvalho, B. Soret, T. Abrao, J. C. Marinello, E. G. Larsson, and P. Popovski, "Massive mimo for internet of things (iot) connectivity," *Physical Communication*, vol. 37, p. 100859, 2019.

[10] P. von Butovitsch, D. Astely, C. Friberg, A. Furuskär, B. Göransson, B. Hogan, J. Karlsson, and E. Larsson, "Advanced antenna systems for 5g networks," https://www.ericsson.com/en/white-papers/advanced-antenna-systems-for-5g-networks,Whitepaper, accessed: 2019-08-07.

[11] M. Javaid and A. Haleem, "Industry 4.0 applications in medical field: A brief review," *Current Medicine Research and Practice*, vol. 9, no. 3, pp.

102–109, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352081719300418

[12] C. Liaskos, A. Tsioliaridou, A. Pitsillides, S. Ioannidis, and I. Akyildiz, "Using any surface to realize a new paradigm for wireless communications," *Communications of the ACM*, vol. 61, no. 11, pp. 30–33, 2018.

[13] E. Björnson, L. Sanguinetti, H. Wymeersch, J. Hoydis, and T. L. Marzetta, "Massive mimo is a reality—what is next?: Five promising research directions for antenna arrays," *Digital Signal Processing*, vol. 94, pp. 3–20, 2019, special Issue on Source Localization in Massive MIMO. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1051200419300776

[14] E. Björnson, J. Hoydis, and L. Sanguinetti, "Massive mimo networks: Spectral, energy, and hardware efficiency," *Found. Trends Signal Process.*, vol. 11, no. 3–4, p. 154–655, Nov. 2017. [Online]. Available: https://doi.org/10.1561/2000000093

[15] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, "Scaling up mimo: Opportunities and challenges with very large arrays," *IEEE Signal Processing Magazine*, vol. 30, no. 1, pp. 40–60, 2013.

[16] E. Björnson, J. Hoydis, and L. Sanguinetti, "Massive mimo networks: Spectral, energy, and hardware efficiency," *Foundations and Trends in Signal Processing*, vol. 11, no. 3-4, pp. 154–655, 2017.

[17] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Aspects of favorable propagation in massive mimo," in *2014 22nd European Signal Processing Conference (EUSIPCO)*. IEEE, 2014, pp. 76–80.

[18] S. K Mohammed, "Impact of transceiver power consumption on the energy efficiency of zero-forcing detector in massive MIMO systems," in IEEE Transactions on Communications, vol. 62, no. 11, pp. 3874–3890, 2014

[19] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," in IEEE Transactions on Communications, vol. 61 , no. 4, pp. 1436–1449, 2013

[20] H Yin, L Cottatellucci, D Gesbert, R. R Müller and G He, "Robust pilot decontamination based on joint angle and power domain discrimination," IEEE Transactions on Signal Processing, vol. 64, no. 11, pp. 2990–3003, 2016

[21] K. Guo, Y. Guo, G. Fodor, and G. Ascheid, "Uplink power control with MMSE receiver in multi-cell MU-massive-MIMO systems," in IEEE International Conference on Communication (ICC), 2014, pp. 5184–5190.

[22] E. Björnson, "A look at an LTE-TDD Massive MIMO product," `http://ma-mimo.ellintech.se/2018/08/27/a-look-at-an-lte-tdd-massive-mimo-product/`, accessed: 2021-09-07.

[23] Huawei, "Huawei launches 5G simplified solution,", `https://www.huawei.com/en/press-events/news/2019/2/huawei-5g-simplified-solution`, accessed: 2021-09-07.

[24] S. Biswas, C. Masouros, and T. Ratnarajah, "Performance analysis of large multiuser mimo systems with space-constrained 2-d antenna arrays," *IEEE Transactions on Wireless Communications*, vol. 15, no. 5, pp. 3492–3505, 2016.

[25] C. Masouros and M. Matthaiou, "Space-constrained massive mimo: Hitting the wall of favorable propagation," *IEEE Communications Letters*, vol. 19, no. 5, pp. 771–774, 2015.

[26] À. O. Martínez, E. De Carvalho, and J. Ø. Nielsen, "Towards very large aperture massive mimo: A measurement based study," in *2014 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2014, pp. 281–286.

[27] C. Mollen, J. Choi, E. G. Larsson, and R. W. Heath, "Uplink performance of wideband massive mimo with one-bit adcs," *IEEE Transactions on Wireless Communications*, vol. 16, no. 1, pp. 87–100, 2016.

[28] X. Ying, U. Demirhan, and A. Alkhateeb, "Relay aided intelligent reconfigurable surfaces: Achieving the potential without so many antennas," *arXiv preprint arXiv:2006.06644*, 2020.

[29] S. Hu, F. Rusek, and O. Edfors, "Beyond massive mimo: The potential of positioning with large intelligent surfaces," *IEEE Transactions on Signal Processing*, vol. 66, no. 7, pp. 1761–1774, 2018.

[30] H. Wang, A. Kosasih, C.-K. Wen, S. Jin, and W. Hardjawana, "Expectation propagation detector for extra-large scale massive mimo," in *IEEE Transactions on Wireless Communications*, vol.19, no. 3, IEEE, 2020, pp. 2036-2051.

[31] X. Yang, F. Cao, M. Matthaiou, and S. Jin, "On the uplink transmission of extra-large scale massive mimo systems," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 12, pp. 15 229–15 243, 2020.

[32] O. S. Nishimura, J. C. Marinello, and T. Abrão, "A grant-based random access protocol in extra-large massive mimo system," *IEEE Communications Letters*, vol. 24, no. 11, pp. 2478–2482, 2020.

[33] S. Hu, F. Rusek, and O. Edfors, "The potential of using large antenna arrays on intelligent surfaces," *2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*, IEEE, 2017, pp. 1–6.

[34] Q. Wu and R. Zhang, "Intelligent reflecting surface enhanced wireless network: Joint active and passive beamforming design," in *2018 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2018, pp. 1–6.

[35] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive mimo versus small cells," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1834–1850, 2017.

[36] X. Gao, F. Tufvesson, and O. Edfors, "Massive mimo channels—measurements and models," in *2013 Asilomar conference on signals, systems and computers*. IEEE, 2013, pp. 280–284.

[37] X. Li, S. Zhou, E. Björnson, and J. Wang, "Capacity analysis for spatially non-wide sense stationary uplink massive mimo systems," *IEEE Transactions on wireless communications*, vol. 14, no. 12, pp. 7044–7056, 2015.

[38] L. Liu, C. Oestges, J. Poutanen, K. Haneda, P. Vainikainen, F. Quitin, F. Tufvesson, and P. De Doncker, "The cost 2100 mimo channel model," *IEEE Wireless Communications*, vol. 19, no. 6, pp. 92–99, 2012.

[39] A. Ali, E. De Carvalho, and R. W. Heath, "Linear receivers in non-stationary massive mimo channels with visibility regions," *IEEE Wireless Communications Letters*, vol. 8, no. 3, pp. 885–888, 2019.

[40] V. C. Rodrigues, A. Amiri, T. Abrão, E. de Carvalho, and P. Popovski, "Low-complexity distributed xl-mimo for multiuser detection," in *2020 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2020, pp. 1–6.

[41] M. Sarajlić, F. Rusek, J. R. Sánchez, L. Liu, and O. Edfors, "Fully decentralized approximate zero-forcing precoding for massive mimo systems," *IEEE Wireless Communications Letters*, vol. 8, no. 3, pp. 773–776, 2019.

[42] H. Wang, A. Kosasih, C.-K. Wen, S. Jin, and W. Hardjawana, "Expectation propagation detector for extra-large scale massive mimo," *IEEE Transactions on Wireless Communications*, vol. 19, no. 3, pp. 2036–2051, 2020.

[43] T.-C. Zhang, C.-K. Wen, S. Jin, and T. Jiang, “Mixed-adc massive mimo detectors: Performance analysis and design optimization,” *IEEE Transactions on Wireless Communications*, vol. 15, no. 11, pp. 7738–7752, 2016.

[44] M. J. Wainwright and M. I. Jordan, *Graphical models, exponential families, and variational inference*. Now Publishers Inc, 2008.

[45] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[46] T. M. Cover and J. A. Thomas, “Information theory and statistics,” *Elements of Information Theory*, vol. 1, no. 1, pp. 279–335, 1991.

[47] J. S. Yedidia, W. T. Freeman, and Y. Weiss, “Constructing free-energy approximations and generalized belief propagation algorithms,” *IEEE Transactions on information theory*, vol. 51, no. 7, pp. 2282–2312, 2005.

[48] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, “Factor graphs and the sum-product algorithm,” *IEEE Transactions on information theory*, vol. 47, no. 2, pp. 498–519, 2001.

[49] “3GPP TS 38.212. NR; Multiplexing and Channel Coding.” Technical Specification Group Radio Access Network, Standard, Jul. 2018.

[50] M. N. Boroujerdi, S. Haghighatshoar, and G. Caire, “Low-complexity statistically robust precoder/detector computation for massive mimo systems,” *IEEE Transactions on Wireless Communications*, vol. 17, no. 10, pp. 6516–6530, 2018.

[51] X. Tan, Y.-L. Ueng, Z. Zhang, X. You, and C. Zhang, “A low-complexity massive mimo detection based on approximate expectation propagation,” *IEEE Transactions on Vehicular Technology*, vol. 68, no. 8, pp. 7260–7272, 2019.

[52] H. Han, Y. Li, and X. Guo, “A graph-based random access protocol for crowded massive mimo systems,” *IEEE Transactions on Wireless Communications*, vol. 16, no. 11, pp. 7348–7361, 2017.

[53] C. Knievel, P. A. Hoeher, A. Tyrrell, and G. Auer, “Multi-dimensional graph-based soft iterative receiver for mimo-ofdm,” *IEEE transactions on communications*, vol. 60, no. 6, pp. 1599–1609, 2012.

[54] J. Goldberger and A. Leshem, “Mimo detection for high-order qam based on a gaussian tree approximation,” *IEEE transactions on information theory*, vol. 57, no. 8, pp. 4973–4982, 2011.

[55] J. Winn, C. M. Bishop, and T. Jaakkola, “Variational message passing.” *Journal of Machine Learning Research*, vol. 6, no. 4, 2005.

[56] J. Zeng, J. Lin, and Z. Wang, "Low complexity message passing detection algorithm for large-scale mimo systems," *IEEE Wireless Communications Letters*, vol. 7, no. 5, pp. 708–711, 2018.

[57] M. Gormley, "Machine learning for structured data course slides," `http://www.cs.cmu.edu/~mgormley/courses/10418/slides/lecture9-bp.pdf,MachineLearningDepartment,SchoolofComputerScience,CarnegieMellonUniversity`, accessed: 2021-10-02.

[58] Y. Xiong, N. Wei, and Z. Zhang, "A low-complexity iterative gamp-based detection for massive mimo with low-resolution adcs," in *2017 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2017, pp. 1–6.

[59] Y. Chi and Y. M. Lu, "Kaczmarz method for solving quadratic equations," *IEEE Signal Processing Letters*, vol. 23, no. 9, pp. 1183–1187, 2016.

[60] T. Strohmer and R. Vershynin, "A randomized kaczmarz algorithm with exponential convergence," *Journal of Fourier Analysis and Applications*, vol. 15, no. 2, pp. 262–278, 2009.

[61] M.-L. Sun, C.-Q. Gu, and P.-F. Tang, "On randomized sampling kaczmarz method with application in compressed sensing," *Mathematical Problems in Engineering*, vol. 2020, 2020.

[62] H. Wu, B. Shen, S. Zhao, and P. Gong, "Low-complexity soft-output signal detection based on improved kaczmarz iteration algorithm for uplink massive mimo system," *Sensors*, vol. 20, no. 6, p. 1564, 2020.

[63] S. Park, A. Alkhateeb, and R. W. Heath, "Dynamic subarrays for hybrid precoding in wideband mmwave mimo systems," *IEEE Transactions on Wireless Communications*, vol. 16, no. 5, pp. 2907–2920, 2017.

[64] X. Gao, O. Edfors, F. Tufvesson, and E. G. Larsson, "Multi-switch for antenna selection in massive mimo," in *2015 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2015, pp. 1–6.

[65] Y. Gao and T. Kaiser, "Antenna selection in massive mimo systems: Full-array selection or subarray selection?" in *2016 IEEE sensor array and multichannel signal processing workshop (SAM)*. IEEE, 2016, pp. 1–5.

[66] J. Joung, "Machine learning-based antenna selection in wireless communications," *IEEE Communications Letters*, vol. 20, no. 11, pp. 2241–2244, 2016.

[67] A. M. Elbir, K. V. Mishra, and Y. C. Eldar, "Cognitive radar antenna selection via deep learning," *IET Radar, Sonar & Navigation*, vol. 13, no. 6, pp. 871–880, 2019.

[68] B. Liu, Z. He, and Q. He, “Optimization of orthogonal discrete frequency-coding waveform based on modified genetic algorithm for mimo radar,” in *2007 International Conference on Communications, Circuits and Systems*. IEEE, 2007, pp. 966–970.

[69] K.-S. Tang, K.-F. Man, S. Kwong, and Q. He, “Genetic algorithms and their applications,” *IEEE signal processing magazine*, vol. 13, no. 6, pp. 22–37, 1996.

[70] “Ericsson refines antenna ‘stripe’ technology to improve coverage,” `https://www.fiercewireless.com/wireless/ericsson-refines-antenna-stripe-technology-to-improve-coverage`, accessed: 2021-04-28.

Part II

Papers

Paper A

Extremely large aperture massive MIMO: Low complexity receiver architectures

Abolfazl Amiri, Marko Angjelichinoski, Elisabeth De Carvalho,
Robert W. Heath

The paper has been published in the
IEEE Globecom Workshops (GC Wkshps) Proceedings, pp.1–6, 2018.

Abstract

This paper focuses on new communication paradigms arising in massive multiple-input-multiple-output systems where the antenna array at the base station is of extremely large dimension (xMaMIMO). Due to the extreme dimension of the array, xMaMIMO is characterized by spatial non-stationary field properties along the array; this calls for a multi-antenna transceiver design that is adapted to the array dimension but also its non-stationary properties. We address implementation aspects of xMaMIMO, with computational efficiency as our primary objective. To reduce the computational burden of centralized schemes, we distribute the processing into smaller, disjoint sub-arrays. Then, we consider several low-complexity data detection algorithms as candidates for uplink communication in crowded xMaMIMO systems. Drawing inspiration from coded random access, one of the main contributions of the paper is the design of low complexity scheme that exploits the non-stationary nature of xMaMIMO systems and where the data processing is decentralized. We evaluate the bit-error-rate performance of the transceivers in crowded xMaMIMO scenarios. The results confirm their practical potential.

***Keywords*—** Very large arrays, Massive MIMO, non- stationary, coded random access, 5G

1 Introduction

Massive multiple-input-multiple-output (MIMO) is a key technology in cellular communication systems for increasing area spectral efficiency [1], [2]. The highest gains with massive MIMO are achieved when the antenna array dimension is very large [3, 4]. This has motivated the introduction of new types of deployment where arrays with extremely large dimension are deployed as part of a large infrastructure, for example along the walls of buildings in a mega-city, in airports, large shopping malls or along the structure of a stadium [4]. Similarly, large intelligent surfaces have emerged involving large electromagnetic surfaces [5]. Such a massive MIMO system with antenna arrays of extremely large dimension is denoted as xMaMIMO.

With increased antenna array dimensions, spatial non-wide sense stationary properties appear across the array due to electromagnetic propagation attributes as well as the distance between the users and the array that becomes smaller than the Rayleigh distance (see Fig. 1). In such xMaMIMO systems, different channel models and receiver algorithms are needed that account for this non-stationarity.

In this paper, we consider non-stationary properties through the concept of visibility region. A visibility region is associated to one given user and is defined as the portion of the array that one given user sees, i.e. that is able to receive signals from the user. This behaviour of the channel introduces an inherent *sparsity* to the system model, meaning that the transmitted signal of one user only exists on a small part of the antenna array. Thus, in contrast to ordinary massive MIMO models, user detection can be done by only processing the visibility region of each user. Using this important property of the system, we cut the computation costs of central processing, i.e. processing all antenna elements together, and propose local approaches. Note that the vast majority of the existing works on massive MIMO are based on conventional standard models with stationary characteristics of the channel [7]. In [8], an information theory study on non-wide sense stationary characteristics of massive MIMO channels is available where different parts of the array see different propagation paths. The problem of user assignment in large intelligent surfaces is studied in [9] in an interference-free environment.

To exploit cluster visibility regions, we propose new algorithms for uplink data detection. One of the challenges in xMaMIMO is its practical implementation, especially the enormous computational load that is required. To reduce the computation load of the system, we divide the array into smaller, disjoint units, referred to as subarrays and we distribute the computations among them. Then, we propose two types of uplink receivers. The first receiver is based on distributed linear data fusion (DLDF), where the users are first softly detected per subarray and then linearly fused in a centralized manner to produce the final soft information used to reconstruct the symbols.

Next, relying on the non-stationary nature of the xMaMIMO system and drawing inspiration from coded random access, we propose a decentralized receiver of very low complexity where processing is executed locally per subarray with the fusion centre acting only as a forwarding node, relaying messages among the subarrays. One important factor here is the order of local processes and our proposed method copes delicately with it. The simulation results confirm the practical potential of the proposed receivers for xMaMIMO systems especially in crowded applications.

2 System Model

We consider an xMaMIMO system. As discussed earlier, such infrastructure can be deployed along walls of buildings in urban sprawls, airports, shopping malls, even stadiums and they are envisioned to provide services to massive crowds.

A possible way to deal with the enormous computational load of the xMaMIMO system is to distribute the computation within separate processing units, referred to as *subarrays*. Depending on the specific implementation of the system and the actual physical constraints, a subarray can be defined in various ways. For instance, a subarray can correspond to a separate physical component. To see this, consider a large stadium. To provide high quality connectivity, an xMaMIMO system can be deployed along its walls. Depending on the actual deployment burden and cost, the operator might choose to mount individual arrays and connect them into a central processing unit using a cloud radio access network architecture. In such case, the number and the sizes of the subarrays is fixed. Alternatively, the operator might install a single array and provide logical interconnections between different portions of it. Here, the subarrays can be defined flexibly, adapting their size, number and position to the evolving data traffic conditions. We note that our framework is applicable to both cases as well as any combination in-between.

Let M and K denote the number of antennas and simultaneously active users, respectively. We assume narrow-band transmissions; $\mathbf{x} \in \mathbb{C}^K$ denotes the vector of complex input symbols, $\mathbf{H} \in \mathbb{C}^{M \times K}$ is the complex channel matrix and $\mathbf{n} \sim \mathcal{CN}(0, \sigma_n^2 \mathbf{I}_M)$ is the AWGN ($\mathbf{I}_M$ denotes the identity matrix). We model the received baseband signal $\mathbf{y} \in \mathbb{C}^M$ across the whole array as follows:

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}. \tag{A.1}$$

Let $\mathbf{h}_k$ denote the k-th column of $\mathbf{H}$, corresponding to user k; in this work, we adopt the following channel model [6]:

$$\mathbf{h}_k = \sqrt{\mathbf{w}_k} \odot \bar{\mathbf{h}}_k, \tag{A.2}$$

with $\odot$ denoting the element wise (Hadamard) products between two equal-size vectors. $\mathbf{w}_k$ captures the effect of large scale fading which in turn is a function of the distance of the user from the array, denoted with $\mathbf{d}_k$, and the propagation properties of the environment; here, we employ the following simplified propagation model [11]:

$$\mathbf{w}_k = \beta \mathbf{d}_k^{\gamma}, \tag{A.3}$$

where β is a attenuation coefficient [11] and γ is the pathloss exponent. $\bar{\mathbf{h}}_k \sim \mathcal{CN}(0, \mathbf{I})$ accounts for fast fading.

We split the xMaMIMO system into B subarrays, each with $M^{(b)} \geq K, b = 1, \ldots, B$ antennas such that $\sum_{b=1}^{B} M^{(b)} = M$; the received signal per subarrays is denoted by $\mathbf{y}^{(b)} \in \mathbb{C}^{M_b}$ and can be written as:

$$\mathbf{y}^{(b)} = \mathbf{H}^{(b)}\mathbf{x} + \mathbf{n}^{(b)}, \tag{A.4}$$

for any $b = 1, \ldots, B$. Without loss of generality, in the rest of the paper we will assume that all active users transmit with equal power ($E|x|^2 = 1$).

3 Multiuser Detection Algorithms

In this section, we develop algorithms for multiuser symbol detection in xMaMIMO systems. Throughout, we assume perfect Channel State Information (CSI) at the receiver.

We distinguish between two different regimes of operation of the system: (i) stationary regime, where we assume that the users' energy spread across the whole array (in other words, each user "sees" the whole array), and (ii) non-stationary regime, where we assume that the energy of each user is predominately concentrated on a limited number of antennas (see Fig. A.1), which is usually significantly smaller than M (i.e., each user "sees" only limited portion of the array). Obviously, the inherent, natural regime of operation of the system would be the non-stationary one in general, since the uplink power of each user will be unevenly distributed along the antenna array. Here, the distinction between the two regimes is done according to the knowledge of the receiver, i.e., when we say stationary regime, we mainly refer to the aspect of receiver agnosticism towards the non-stationary nature of the system.

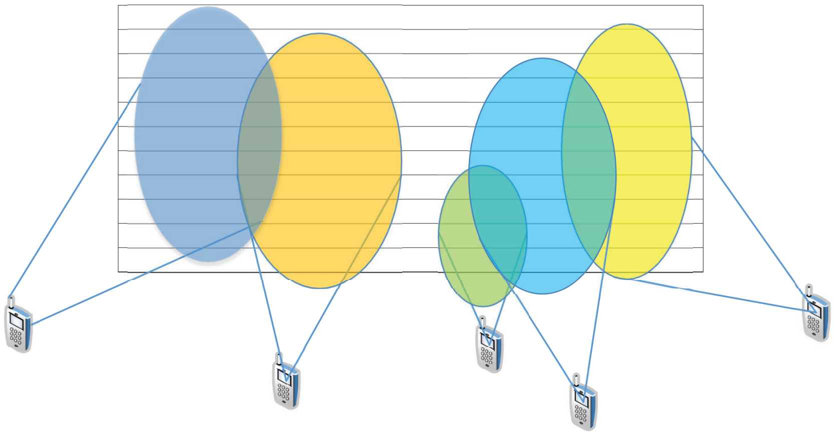


Fig. A.1: An example of extremely large M-MIMO array with spatial non-stationary regions along the array. Each user has a specific visibility region according to the channel conditions.

3.1 Stationary regime

Centralized Zero-Forcing Receiver

Given the main underlying assumption, i.e., each user "reaches" every antenna of the array, a straightforward way to perform multiuser symbol detection is to process the complete received signal $\mathbf{y}$. This can be done simply via Zero-Forcing (ZF); specifically, the ZF receiver for user k, denoted by $\mathbf{F}_{ZF,k}$, can be written as follows:

$$\mathbf{F}_{ZF,k}[\mathbf{H}] = \frac{\mathbf{h}_k^H \mathbf{P}_{\bar{\mathbf{H}}_\mathbf{k}}^{\perp}}{\mathbf{h}_k^H \mathbf{P}_{\bar{\mathbf{H}}_\mathbf{k}}^{\perp} \mathbf{h}_k}, \tag{A.5}$$

with $\mathbf{P}_{\bar{\mathbf{H}}_\mathbf{k}}^{\perp} = \mathbf{I} - \bar{\mathbf{H}}_\mathbf{k}(\bar{\mathbf{H}}_\mathbf{k}^H \bar{\mathbf{H}}_\mathbf{k})^{-1} \bar{\mathbf{H}}_\mathbf{k}^H$ [12]; $\bar{\mathbf{H}}_\mathbf{k}$ is obtained from $\mathbf{H}$ by removing its k^{th} column $\mathbf{h}_k$. The post-processing SNR of the ZF receiver obtains the following form:

$$\text{SNR}_{ZF,k} = \rho\, \mathbf{h}_k^H \mathbf{P}_{\bar{\mathbf{H}}_\mathbf{k}}^{\perp} \mathbf{h}_k, \tag{A.6}$$

with $\rho = 1/\sigma_n^2$. Given the extreme dimension of the aperture and potentially the extremely crowded setup, one should immediately note the computational burden of the centralized scheme. To reduce the computational complexity, we propose two schemes based on subarray processing. In both cases, the underlying idea is simple; instead of processing $\mathbf{y}$ fully, first process $\mathbf{y}^{(b)}, b = 1, \ldots, B$ and then perform linear soft fusion in a centralized manner.

Distributed Linear Data Fusion Receiver

We introduce a simple, distributed linear data fusion (DLDF) method that combines softly the detected signals from each individual subarray. Furthermore, soft information of each user is obtained by

$$\hat{x}_k^{(b)} = \mathbf{F}_{ZF,k}[\mathbf{H}^{(b)}]\mathbf{y}^{(b)} \tag{A.7}$$

```
Algorithm 1 Distributed Linear Data Fusion receiver.
Result: Estimates of $x_k, k = 1, \dots, K$
Initialize: $\mathbf{H}$, $K$, $B$, $M^{(b)}$, $\mathcal{K} = \{1, \dots, K\}$
Stage I: Distributed Linear Data Fusion (DLDF)
 1. compute $\hat{x}_k^{(b)}, k \in \mathcal{K}, b = 1, \dots, B$ via (A.7)
 2. compute $\alpha_k^{(b)}, k \in \mathcal{K}, b = 1, \dots, B$ via (A.12)
 3. compute $\hat{x}_k, k \in \mathcal{K}$ via (A.8)
 4. perform hard decision over $\hat{x}_k, k \in \mathcal{K}$ and terminate
```

For each user k we define the combined DLDF symbol $\hat{x}_k$ as follows:

$$\hat{x}_k = \sum_{b=1}^{B} \alpha_k^{(b)} \hat{x}_k^{(b)}, \tag{A.8}$$

where $\alpha_k^{(b)}$ is the weight for user k using from subarray b; note that $\sum_{b=1}^{B} \alpha_k^{(b)} = 1$. It is worth mentioning that (A.8) is done in the central unit after receiving all the soft information from the subarrays. Also, mean squared error (MSE) of each user on subarrays is defined as:

$$\mathrm{MSE}_k^{(b)} = E|x_k^{(b)} - \hat{x}_k^{(b)}|^2 \tag{A.9}$$

where E denotes the expectation operation. Here, it is taken with respect to the noise.

As the noise is assumed independent across subarrays, the overall MSE when data fusion is performed is:

$$\mathrm{MSE}_k = \sum_{b=1}^{B} \alpha_k^{(b)2} \mathrm{MSE}_k^{(b)}, \tag{A.10}$$

The objective is to minimize MSE_k with the constraint $\sum_{b=1}^{B} \alpha_k^{(b)} = 1$. Using the Lagrange multiplier method [13] gives us the optimal weights:

$$\alpha_k^{(b)2} = \frac{\frac{1}{\mathrm{MSE}_k^{(b)}}}{\sum_{b=1}^{B} \frac{1}{\mathrm{MSE}_k^{(b)}}}, \tag{A.11}$$

for $b = 1, \dots, B$. Given that all users use equal transmit power, normalized to 1, we have that $\mathrm{SNR}_k^{(b)} = 1/\mathrm{MSE}_k^{(b)}$, which yields:

$$\alpha_k^{(b)} = \frac{\mathrm{SNR}_k^{(b)}}{\sum_{b=1}^{B} \mathrm{SNR}_k^{(b)}} \tag{A.12}$$

for $b = 1, \dots, B$. The complete algorithm is summarized in Algorithm 1.

Algorithm 2 Bipartite graph construction from **H**.

```
Result: $\mathcal{H}$
Initialize: $\mathbf{H}$, $K$, $B$, $M_b$, $p_0$, $\mathcal{H} = \{0\}^{B \times K}$, $\mathcal{B} = \{1, \ldots, B\}$
for $k = 1$ to $K$ do
    1. reinitialize $p_k = 0$
    2. compute total cumulative power $P_k$
    3. compute per subarray power $P_k^{(b)}, b \in \mathcal{B}$
    while $p_k \leq p_0 \cdot P_k$ do
        1. find $b^* = \max_{b \in \mathcal{B}} P_k^{(b)}$
        2. $p_k = p_k + P_k^{(b^*)}$
        3. set $\mathcal{H}(b^*, k) = 1$
        4. $\mathcal{B} = \mathcal{B} \setminus b^*$
    end
end
```

3.2 Non-stationary regime

In this case, we assume that the users have a limited visibility region of the array, which is illustrated in Fig. A.1. Hence, the non-stationary regime of operation can be seen as a special case of the stationary one, implying that we can easily apply any of the receivers described in the previous subsection.

Nevertheless, we introduce a simple method, inspired from the concept of coded random access in slotted aloha IoT systems. They key idea operates as follows: given the non-stationary nature of the array, each user is predominantly present only on very limited number of subarrays, i.e., all of its power is concentrated over limited number of subarrays. As a result, the system becomes inherently *sparse*, implying that the subarrays where a user is not present should not be processed for that specific user. So, in principle, we can obtain a sparse bipartite graph, representing the connections of the users to the subarrays after which we can apply the principles of successive elimination of connections from the graphs as in coded random access. This further reduces the computational cost but since we are neglecting some portion of the signal energy, p_0, and treat it as interference at the remaining arrays, and we do not perform any soft fusion at the central processing unit, it is reasonable to expect that the performance of the method might be slightly degraded in some configuration regions of the system (i.e., specific values of K relative to B and $M_b, b = 1, \ldots, B$).

The most attractive feature of the proposed method is the fact that the bipartite graph can be constructed very simply, exploiting the sheer fact that the receiver has perfect CSI, i.e., it knows $\mathbf{H}$; in other words, by observing the k-th column, the receiver can determine which parts of the array the dominant part of the power of user k is allocated. This way, the receiver obtains a binary matrix $\mathcal{H} \in \{0, 1\}^{B \times K}$. We use $\mathcal{H}$ to construct a bipartite graph. Note that, at the beginning of the algorithm 3, the central unit runs Algorithm 2 according to the CSI and then sends the order of the detection to each subarray. This means that each subarray receives a schedule consisting of the list of users that should be detected by that subarray. This is the only centralized

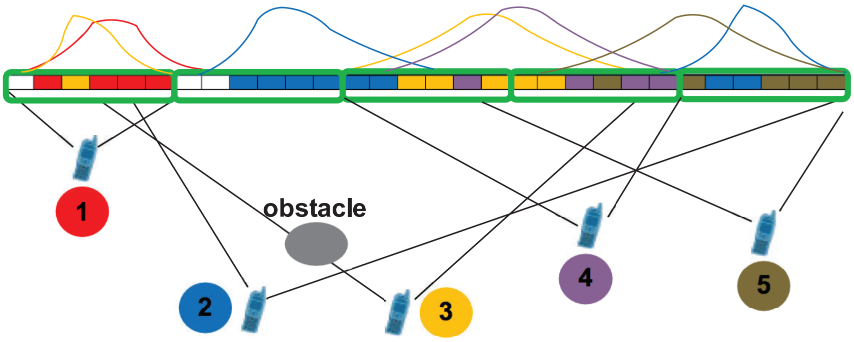


Fig. A.2: An example of linear M-MIMO array with different user visibility regions. Equivalent graph representation for this system is shown in Fig. A.3 (a).

broadcasting in this algorithm and the rest of it is decentralized.

The procedure is described in Algorithm 2. Moreover, an example of xMaMIMO with 5 users is illustrated in Fig. A.2 where the energy distribution of each user on antenna arrays is also presented. The equivalent bipartite graph representation of this setup is shown in Fig. A.3 (a). The extension of this binary graph to a weighted one, where the weights show the portion of user's total power, is left for future work.

The bipartite graph constructed this way is characterized by the following quantities: (i) a set $\mathcal{B}$ of B nodes representing subarray units, (ii) a set $\mathcal{K}$ of K user nodes and, (iii) a set $\mathcal{E}$ of edges, i.e. connection between users and subarrays. We use $\mathcal{G} = (\mathcal{B}, \mathcal{K}, \mathcal{E})$ to denote this graph. We also define the node degrees $S_b, b \in \mathcal{B}$ and $U_k, k \in \mathcal{K}$; the degrees give the number of edges connected to each of the nodes in $\mathcal{B}$ and $\mathcal{K}$, respectively. For instance in Fig. A.3(a), subarray $b = 1$ only receives signal from user $k = 1$ and user $k = 3$; therefore, its degree is $S_1 = 2$.

Once the graph has been constructed, we apply simple symbol detection strategy inspired from coded random access [10]. Hence, we search for subarrays with the lowest number of users; we detect the symbols of those users and subsequently remove them from the other subarrays. An illustrative example for the procedure is shown in Fig. A.3. We assume xMaMIMO system with $K = 5$ users and $B = 5$ subarrays with graph representation shown in Fig. A.3(a). After computing the degrees, we see that $S_2 = 1$. Thus, we start signal detection in subarray $b = 2$ for user $k = 2$. Then, we remove any other edges corresponding to user $k = 2$ from the graph. We repeat the procedure for S_5 and U_5 in Fig. A.3(b). In step (c), we have three nodes with same degree and similar conditions. We randomly choose S_3 and start a ZF detection within subarray $b = 3$ between users $k = 3$ and $k = 4$. After recovering both of them, we remove all edges from the graph corresponding to those users. Finally, in part (d) we have a singleton node that can be easily detected. Now, since all the users are detected, the algorithm terminates. The complete algorithm is summarized in Algorithm 3.

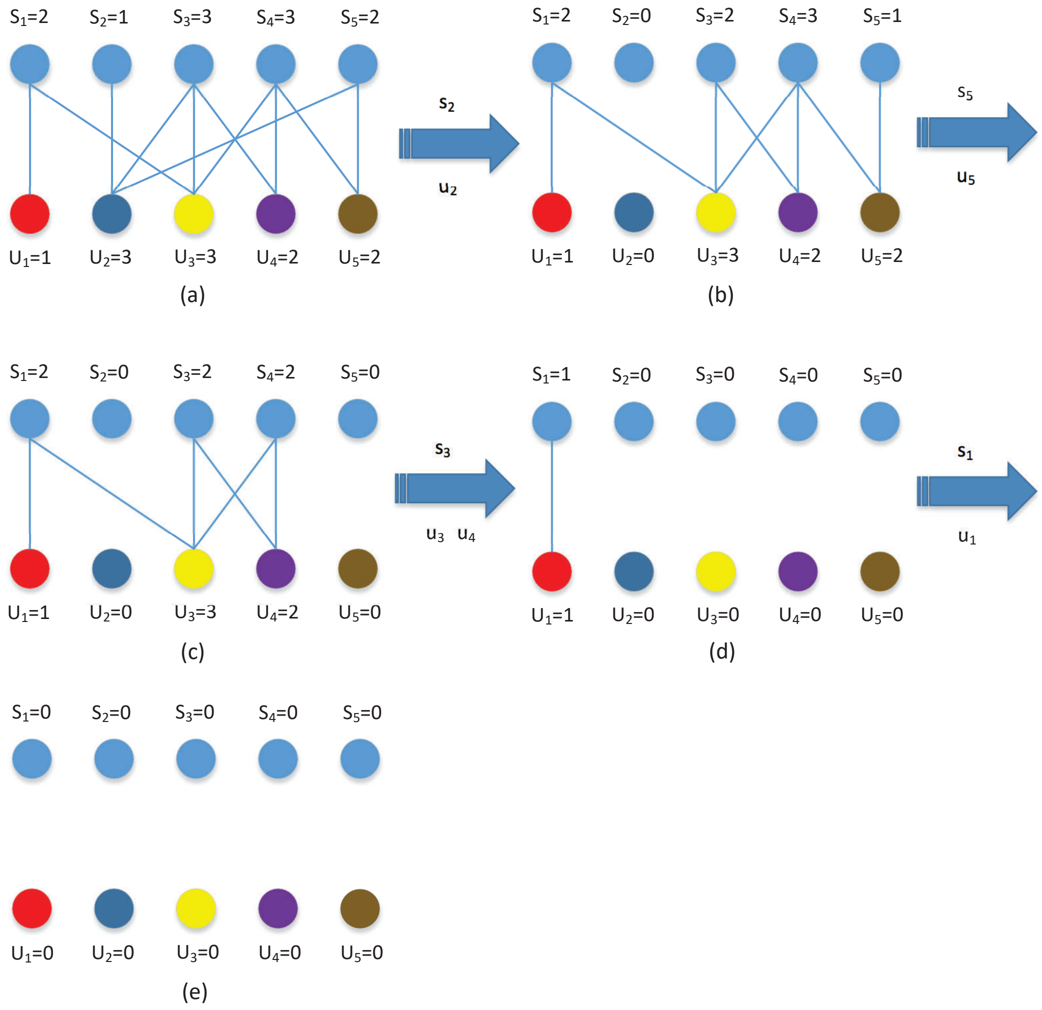

S_1=2 S_2=1 S_3=3 S_4=3 S_5=2
U_1=1 U_2=3 U_3=3 U_4=2 U_5=2
(a)
s_2
u_2
S_1=2 S_2=0 S_3=2 S_4=3 S_5=1
U_1=1 U_2=0 U_3=3 U_4=2 U_5=2
(b)
s_5
u_5
S_1=2 S_2=0 S_3=2 S_4=2 S_5=0
U_1=1 U_2=0 U_3=3 U_4=2 U_5=0
(c)
s_3
u_3 u_4
S_1=1 S_2=0 S_3=0 S_4=0 S_5=0
U_1=1 U_2=0 U_3=0 U_4=0 U_5=0
(d)
s_1
u_1
S_1=0 S_2=0 S_3=0 S_4=0 S_5=0
U_1=0 U_2=0 U_3=0 U_4=0 U_5=0
(e)

Fig. A.3: An example of the proposed detection model on a bipartite graph model with $K = 5$ users and $B = 5$ subarrays. (a): connections of the users to each array. The subarray with the lowest number of users is selected (S_2). The corresponding user symbols are detected and removed from the other subarrays. (b): the procedure is repeated for S_5 and U_5. (c): S_3 is randomly selected. ZF detection is used to decode user 3 and user 4. Their data is removed from the other subarrays. (d): the last user is detected.

Algorithm 3 Low complexity multiuser detection in non-stationary regime

```
Result: Estimates of $x_k, k = 1, \dots, K$
Initialize: $\mathbf{H}$, $K$, $B$, $M_b$, $\mathcal{B} = \{1, \dots, B\}$, $\mathcal{K} = \{1, \dots, K\}$
1. compute $\mathcal{H}$ via Algorithm 2
while $\mathcal{K} \neq \emptyset$ do
    1. compute node degrees $S_b, b \in \mathcal{B}$, $U_k, k \in \mathcal{K}$
    2. find $b^* = \min_{b \in \mathcal{B}} \{S_b\}$
    if $b^* = 1$(only user $k^*$in the subarray with minimal degree) then
        1. compute $\hat{x}_{k^*}^{(b^*)}$ via (A.7)
        2. broadcast $\hat{x}_{k^\dagger}$ so other subarrays remove it from $\mathbf{y}^{(b)}$ for $b \in \mathcal{B} \setminus b^*$
        3. $\mathcal{K} = \mathcal{K} \setminus k^*$
    if $b^* > 1$(multiple users $\mathcal{K}^* \subset \mathcal{K}$ in the subarray with minimal degree) then
        while $\mathcal{K}^* \neq \emptyset$ do
            1. sort the users according to $\mathrm{SNR}_{ZF,k}, k \in \mathcal{K}^*$ in (A.6)
            2. find $k^\dagger = \max_{k \in \mathcal{K}} \mathrm{SNR}_{ZF,k}$
            3. broadcast $\hat{x}_{k^\dagger}$ so all subarrays remove it from $\mathbf{y}^{(b)}$ for $b \in \mathcal{B}$
            4. $\mathcal{K}^* = \mathcal{K}^* \setminus k^\dagger$
        end
        6. $\mathcal{K} = \mathcal{K} \setminus \mathcal{K}^*$
end
```

4 Complexity, Convergence and Delay Analyses

In this section we first consider computation complexity comparison between the proposed algorithm and the linear data fusion using ZF detector. Convergence and delay characteristics of the proposed algorithms are discussed next.

4.1 Complexity of DLDF

In DLDF we have three phases for user detection which have the following complexities:

1. Data detection: Consists of ZF matrix inversion for all users in all of the subarrays with $M_b K$ elements, which have a complexity order of $BK(K)^3$.
2. SNR extraction: This function is also for all users in all of the subarrays containing $BK(M_b(K-1))$.
3. Soft fusion: The last part with BK matrix multiplications.

4.2 Complexity of Algorithm 3

We study this part with two extreme cases that could happen regarding the nature of non-stationarity.

Table A.1: Complexity comparison of the studied methods

Methods	Number of multiplications
ZF-DLDF	$BK(K)^3 + BK((K-1)) + BK$
Algorithm 3	Worst case: $K^4 + K(M_b(K-1))$
	Best case: $K^4B^{-3} + \frac{KM_bK}{B}$

Worst case

This case occurs when we have all of the users in all of the subarrays or when we have them in only one subarray. Thus, the algorithm sets $\min N_s = K$ and performs detection over only one subarray. Moreover SNR extraction is also done for all users in this subarray containing $K(M_b(K-1))$. Therefore, the complexity of this part is at most $K^4 + K(M_b(K-1))$ calculations.

Best case

This case happens when we have users evenly distributed between subarrays meaning that each subarray performs detection over $\frac{K}{B}$ users. Therefore the complexity of the detection part is

$$B\left[\left(\frac{K}{B}\right)^3 + \left(\frac{K}{B}-1\right)^3 + \cdots + 1\right] < \frac{K^4}{B^3}. \tag{A.13}$$

Also, for the ordering part we have $\frac{KM_bK}{B}$ computations. TABLE A.1 provides the complexity comparison between the aforementioned algorithms.

A rough comparison between the complexities of the Algorithms reveals that the complexity reduction of Algorithm 3 scales with B in the worst case and with B^4 in the best case. If B is of the order of 10, we see that significant computational power can be saved by employing Algorithm 3.

4.3 Convergence Analyses

The analyses for the convergence for Algorithm 3 over graphical model can be found in [10]. Moreover, considering this algorithm as an extension to the model in [10] by enabling continuous signal space, i.e. we recover a part of data at each step even on non-singleton nodes, it will converge even faster due to this claim.

4.4 Delay Characteristics

In order to highlight the trade-off between using our proposed algorithms instead of the centralized methods, we discuss delay properties of the algorithms. Algorithm 3 introduces some delay due to the sequential nature of the algorithm. For example, while a selected subarray b performs the detection of a given symbol, the other subarrays carrying the same symbol should wait for the input from subarray b to perform interference cancellation.

This waiting time vanishes when we have sparse channels since different subarrays become independent as they involve different users. Thus, they can work together in a parallel mode. A deeper analysis of the overall delay is left for future work.

5 Simulation Results

In this section we compare the Bit Error Rate (BER) performance of the proposed algorithm and DLDF. We assume a linear xMaMIMO configuration (see Fig. A.2) for our simulations setup. We use Monte-Carlo simulations to generate the channel realizations. Some of the fixed variables are: $\beta = 1$, $\gamma = 2$, array length $= 100$ m, power threshold for constructing the bipartite graph $p_0 = 0.9$ and we use an 8-PSK input constellation. Moreover, we assume that the user location has a random uniform distribution along the array. Note that user distribution and antenna array length have a direct impact on the large scale fading characteristics and therefore control the bipartite graph structure.

Fundamentally, for a given average user load per subarray, the BER performance is determined by the dimension of the visibility region seen by each user. Here, the users are at the same distance and uniformly distributed along the array, so that the average size of the visibility region is the same per user (except for users at the edges which have a non-significant impact for a large enough number of users). In the simulation, we study the following factors: a) the number of subarrays, b) the total number of users and c) the number of antenna per subarray.

First, we compare the BER of the detection methods for different numbers of subarrays, B, while the number of antennas is fixed, $M = 512$, starting from a number of 2 subarrays (256 antennas per subarray) and ending by a number of 32 subarrays (16 antennas per subarray). We observe:

- Algorithm 3 performs significantly better than the DLDF.
- Linear processing: when the number of antennas per subarray is asymptotically large, due to the law of large numbers, the processing decouples across the subarrays (this fact is supported by the rich scattering assumption). Subarray processing becomes equivalent to a centralized linear processing. We observe an almost stable performance level (corresponding to the centralized processing) until a number of antennas per subarray smaller then 32.
- Algorithm 3: the performance of algorithm 3 is degraded when the number of subarray is small. The reason is that the algorithm lacks degrees of freedom for the SIC mechanism to have its full effect. Performance saturates when the resolution offered by the number of subarrays reflects the non-stationarity patterns, more precisely when each subarray offers a stationary picture of the received signal.

In Fig. A.5, we compare the performance of Algorithm 3 and DLDF with respect to the number of users. The total array size is kept fixed. The array is comprised of B

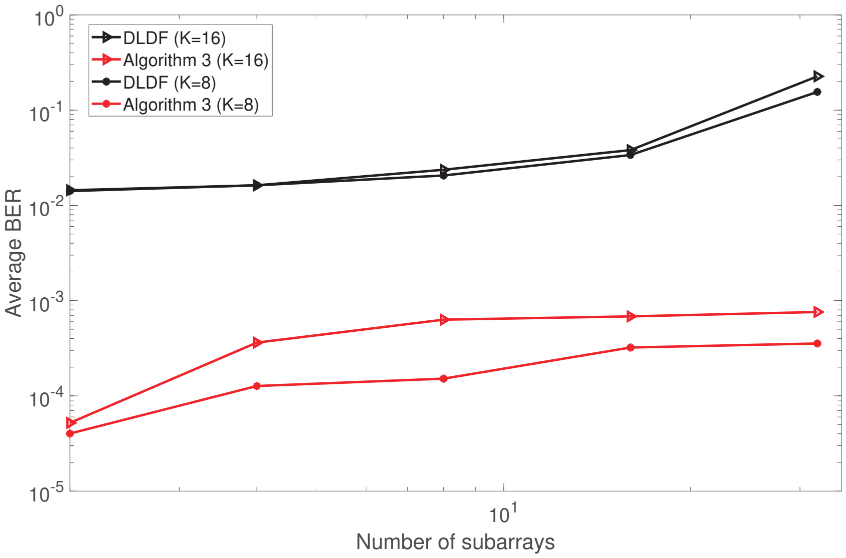


Fig. A.4: The effect of the number of subarrays B on the average BER of the detection systems. We set $M = 512$ and SNR$= 25$dB.

subarrays, each with a number of antennas $M_b = K$. Note that the subarrays are not adjacent in this simulation. We make the following observations:

- Again, algorithm 3 performs significantly better than the DLDF.
- As the number of antennas grows at the same speed as the number of users, the user load per subarray is maintained so that performance remains approximately constant.
- We observe an improvement of algorithm 3 as the granularity increases, i.e. the number of subarrays.

6 Conclusions

In a massive MIMO systems with extremely large arrays, users can effectively communicate only with a sub-part of the array called a visibility region. A receiver design should be adapted to this kind of non-stationary patterns with partially overlapping visibility regions. The receiver architectures proposed in this paper are based on subarray processing where part of the computational load is carried out. A central unit coordinates the operations at each subarray and proceeds to data fusion. We proposed a linear data fusion method, as well as a graph-based algorithm inspired from coded random access which uses low complexity and distributed scheme for the data detection. This method converts the propagation environment of the channel into a bipartite graph and detects the users in a novel scheme.

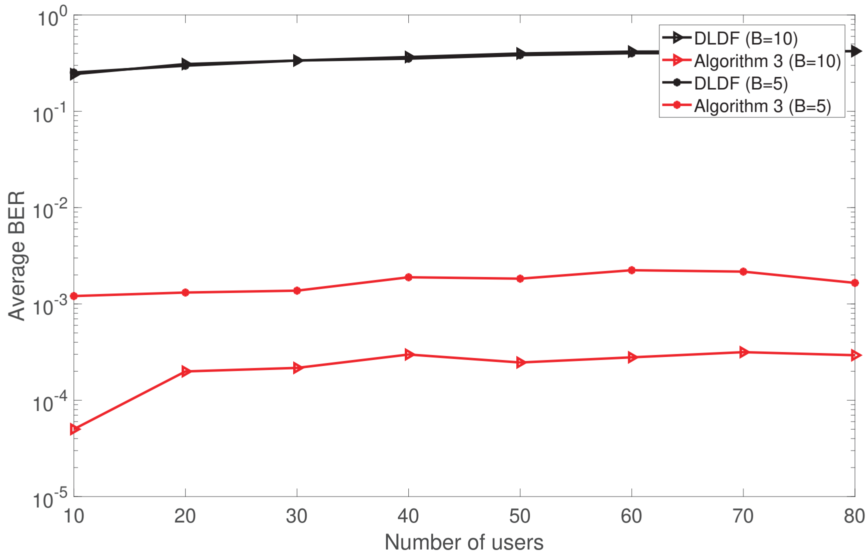


Fig. A.5: Performance comparison between Algorithm 3 and DLDF with respect to number of active users. SNR= 25dB and $K/M_b = 1$.

References

[1] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta and P. Popovski, "Five disruptive technology directions for 5G," in IEEE Communications Magazine, vol. 52, no. 2, pp. 74-80, Feb. 2014.

[2] E. G. Larsson, O. Edfors, F. Tufvesson and T. L. Marzetta, "Massive MIMO for next generation wireless systems," in IEEE Communications Magazine, vol. 52, no. 2, pp. 186-195, Feb. 2014.

[3] X. Gao and F. Tufvesson and O. Edfors, "Massive MIMO channels - Measurements and models" 2013 Asilomar Conference on Signals, Systems and Computers, pp. 280-284, Nov. 2013.

[4] A.O. Martinez, E. de Carvalho , J. Nielsen, "Towards very large aperture massive MIMO: A measurement based study", in 2014 IEEE Globecom Workshops (GC Wkshps), pp. 281-286, Dec 2014.

[5] S. Hu and F. Rusek and O. Edfors , "Beyond Massive MIMO: The Potential of Positioning With Large Intelligent Surfaces", IEEE Transactions on Signal Processing, vol. 66, no. 7, pp. 1761-1774, Apr. 2018.

[6] H. Q. Ngo, E. G. Larsson and T. L. Marzetta, "Energy and Spectral Efficiency of Very Large Multiuser MIMO Systems," in IEEE Transactions on Communications, vol. 61, no. 4, pp. 1436-1449, Apr. 2013.

[7] T. L. Marzetta, "Noncooperative Cellular Wireless with Unlimited Numbers of Base Station Antennas," in IEEE Transactions on Wireless Communications, vol. 9, no. 11, pp. 3590-3600, Nov. 2010.

[8] X. Li, S. Zhou, E. Björnson and J. Wang, "Capacity Analysis for Spatially Non-Wide Sense Stationary Uplink Massive MIMO Systems," in IEEE Transactions on Wireless Communications, vol. 14, no. 12, pp. 7044-7056, Dec. 2015.

[9] S. Hu, K. Chitti, F. Rusek and O. Edfors, "User Assignment with Distributed Large Intelligent Surface (LIS) Systems." arXiv preprint arXiv:1709.01696, 2017.

[10] G. Liva, "Graph-based analysis and optimization of contention resolution diversity slotted ALOHA", IEEE Transactions on Communications, vol. 59, no. 2, pp. 477-487, Nov. 2011.

[11] D. Tse and P.Viswanath,"Fundamentals of Wireless Communication", Cambridge University Press, New York, NY, USA, 2015.

[12] Tim Brown, Elisabeth De Carvalho and Persefoni Kyritsi, "Practical guide to MIMO radio channel: With MATLAB examples", John Wiley & Sons, 2012.

[13] S. Boyd and L. Vandenberghe, "Convex optimization", Cambridge university press, 2004.

Paper B

A Message Passing Based Receiver for Extra-Large Scale MIMO

Abolfazl Amiri, Carles Navarro Manchón, Elisabeth de Carvalho

The paper has been published in the
IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP) Proceedings pp. 564–568, 2019.

The layout has been revised.

Abstract

We consider a massive MIMO system where the array at the access point reaches a dimension that is much larger than the array in current systems. Transitioning to an extremely large dimension and hence large number of antennas implies a need to scale up the multi-antenna processing while maintaining a reasonable computational complexity. In this paper, we study the receiver of such an extra-large scale MIMO (XL-MIMO) system. We propose to base the reception on Variational Message Passing (VMP). The motivation is that the complexity of VMP scales (almost) linearly with the number of antennas and number of users, hence enabling low-complexity reception in crowd scenarios. Furthermore, VMP adapts to the non-stationarities of the MIMO channel that appear due to the large dimension of the array. Through numerical results, we show significant performance improvement and computational complexity reduction compared to a zero-forcing receiver.

***Keywords*—** Massive MIMO, Message passing, Extra-large scale MIMO, Crowd scenarios

1 Introduction

Motivated by the significant spectral efficiency gains brought by massive multiple-input multiple-output (MIMO) systems, the study of systems with even larger number of antenna elements is currently under investigation. In conventional massive MIMO [1], the arrays have a moderate size with an antenna spacing with half the wavelength. In extra-large scale MIMO (XL-MIMO), the focus is placed on increasing the dimension of the antenna array at the access point to capture additional spatial degrees of freedom [2].

Due to the large array dimension and according to the electromagnetic propagation effects, spatial non-wide sense stationary properties appear along the array in XL-MIMO (see Fig. B.1) [2]. This results in the appearance of *visibility regions (VR)*, which are the areas of the array where most of a given users' energy is concentrated. The non-stationary properties impact the performance of XL-MIMO and calls for adapted transceiver designs. Another challenge is that of the computational complexity of the receiver processing, especially when the system is serving a large number of users. Since the conditions for favorable propagation are not satisfied when the number of users is not much smaller than the number of antennas, simple receivers such as a matched filter do not offer good performance [3]. In addition, more advanced linear options such as zero-forcing (ZF) and linear minimum mean squared error (MMSE) receivers have prohibitive complexity due to large matrix inversions. Hence, there is a need for finding multi-user detection algorithms whose complexity scales well with the number of antenna elements and users, and whose performance is comparable to that of classical linear MIMO receivers.

We studied the receiver design for large scale MIMO systems in [4] where distributed units, called *sub-arrays*, detect users' signals by cooperation. We used a successive interference cancellation (SIC) based method between sub-arrays. Message passing (MP) has been applied to massive MIMO systems. In [5] authors develop low complexity MP methods using graphical models. Both channel estimation and data detection problems with one-bit quantization are solved with MP techniques in [6]. Recently, authors in [7] used expectation propagation (EP) to solve the symbol detection problem in XL-MIMO systems. They have exploited the sub-array structure to model their EP scheme. To the best of our knowledge, the case of heavily loaded XL-MIMO system, where the ratio between the numbers of BS antennas and active users is small (less than 5), is not studied in the literature. The main challenges in these scenarios are the extreme complexity of the conventional methods and huge degradation in the performance of the benchmark linear receivers.

In this paper we propose a multi-user symbol detection algorithm for XL-MIMO systems based on variational message passing (VMP). In the VMP framework, the a-posteriori probability of the symbols from all users is approximated by a fully-factorized distribution [8, 9], yielding an algorithm with complexity that scales linearly with the number of users and antenna elements, as no matrix inversions are involved. The proposed algorithm is initialized with a maximal ratio combiner (MRC), whose complexity also scales linearly with the system dimensions. In addition, since the variational framework operates with approximations of the posterior probability distributions of the unknown variables rather than with point estimates, it provides an inherent way of optimally fusing the information on a user's symbol obtained from the different antenna elements in the array. This property is especially useful in the presence of spatial non-stationarities and VRs, as occurring in XL-MIMO arrays.

Our complexity analysis shows that the number of multiplications for our proposed algorithm grows much slower with the size of the BS array than for the ZF, especially in crowded scenarios. We showcase the performance of the proposed receiver in an XL-MIMO system with a large user load, showing that it outperforms the ZF receiver with a significantly lower computational overhead.

2 System Model

In this section, we present the system model and introduce a channel model incorporating spatial non-stationarities. We denote the number of antennas and simultaneously active users by M and K, respectively. We assume narrow-band transmissions; $\mathbf{x}{\in}\mathbb{C}^K$ denotes the vector of complex user symbols, $\mathbf{H}{\in}\mathbb{C}^{M\times K}$ is the complex channel matrix and $\mathbf{n}{\sim}\mathcal{CN}(0,\sigma_n^2\mathbf{I}_M)$ is the AWGN ($\mathbf{I}_M$ denotes the identity matrix). We model the received baseband signal $\mathbf{y}\in\mathbb{C}^M$ as:

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}. \tag{B.1}$$

Denoting by $\mathbf{h}_k$ the kth column of $\mathbf{H}$, corresponding to user k, and adopting the channel model in [3]:

$$\mathbf{h}_k = \sqrt{\mathbf{w}_k} \odot \bar{\mathbf{h}}_k, \tag{B.2}$$

with $\odot$ denoting the element wise (Hadamard) products between two equal-size vectors. $\mathbf{w}_k$ captures the effect of large scale fading and has entries

$$w_{k,m} = \Omega_k s_{k,m}^{\nu}, \qquad m = 1, \dots, M \tag{B.3}$$

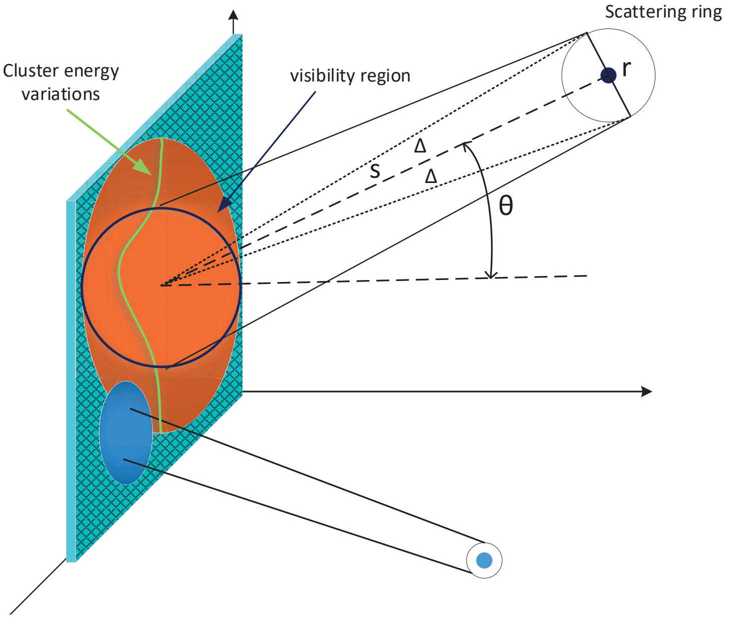


Fig. B.1: An illustrative example of visibility region and received power distribution over XL-MIMO array.

where $s_{k,m}$ is the distance between user k and BS antenna m, Ω_k is an attenuation coefficient and ν is the pathloss exponent [10].

$\bar{\mathbf{h}}_k \sim \mathcal{CN}(0, \mathbf{R}_k)$ models a non-line of sight fast fading scenario, with $\mathbf{R}_k$ being a symmetric positive semi-definite channel covariance matrix. Karhunen-Loeve expansion representation of the channel vectors $\bar{\mathbf{h}}_k$ are

$$\bar{\mathbf{h}}_k = \mathbf{U}_k \mathbf{\Lambda}_k^{\frac{1}{2}} \boldsymbol{\omega}_k, \tag{B.4}$$

where $\boldsymbol{\omega}_k \in \mathbb{C}^{\zeta \times 1} \sim \mathcal{CN}(0, \mathbf{I})$, $\mathbf{\Lambda}_k$ is an $\zeta \times \zeta$ diagonal matrix with dominant eigenvalues and $\mathbf{U}_k \in \mathbb{C}^{M \times \zeta}$ is the tall unitary matrix of the eigenvectors of $\mathbf{R}_k$ corresponding to the ζ dominant eigenvalues.

2.1 Visibility Regions and Correlation Model

In order to model the channel characteristics in the non-stationary conditions, we refer to the measurement-based data from [11]. There, more realistic scenarios exploiting the VRs over a relatively large array (7.35 meters) were applied. The authors modeled different properties for the VRs including VR centers and VR lengths which we denote with c_k and l_k, respectively for user k. The c_k's are modeled with a uniform random variable over the array, i.e. $c_k \sim \mathcal{U}(0, L)$, where L is the physical length of the XL-MIMO array. The length of the VR follows a log-normal distribution, $l_k \sim \mathcal{LN}(\mu_l, \sigma_l)$.

Exploiting the well-known *one-ring* model [12] to define $\mathbf{R}_k$, the correlation between the channel coefficients of antennas p and q is given by

$$[\mathbf{R}_k]_{p,q} = \frac{1}{2\Delta} \int_{-\Delta}^{\Delta} \exp\left(j\mathbf{f}(\alpha + \theta)(\mathbf{u}_p - \mathbf{u}_q)\right) \mathrm{d}\alpha, \tag{B.5}$$

where $\mathbf{f}(\omega) = -\frac{2\pi}{\lambda}(\cos(\omega), \sin(\omega))$ is the wave vector with carrier wavelength of λ and $\mathbf{u}_p, \mathbf{u}_q \in \mathbb{R}^2$ are the position vectors of the antennas p, q within the VR of user k, angle of arrival of ω and Δ is angular spread which is $\Delta \approx \arctan(\frac{r}{s})$, with r standing for the ring of scatterers radius [13]. Angle θ is the azimuth angle of user k with respect to antenna array (See Fig. B.1). When either of the antenna indices p, q is outside the VR for user k given by the non-stationary parameters, we have $[\mathbf{R}_k]_{p,q} = 0$. In this work, we use a uniform linear array (ULA) configuration and assume a uniform distribution for the scattering rings in front of the array.

3 Proposed Receiver Algorithm

In this section we describe our proposed symbol detection method which is based on variational message-passing (VMP). Due to space limitations, we cannot include the full details of the VMP method, and refer the readers to [8, 9] instead.

3.1 Variational Message Passing

We aim to detect the transmitted user symbols x_k (for user k) which take a value from the constellation set $\mathcal{A} = \{a_1, a_2, \cdots, a_{|\mathcal{A}|}\}$. Moreover, we also estimate as a nuisance variable the noise precision (i.e. inverse noise variance) λ_b, $b = 1, \ldots, M$ at each antenna port. The posterior probability density of these two variables factorizes as

$$p(x_1, \cdots, x_K, \lambda_1, \cdots, \lambda_M | y_1, \cdots y_M) \propto \\ \prod_{b=1}^{M} \underbrace{p(y_b | x_1, \cdots, x_K, \lambda_b)}_{f_{y_b}} \prod_{b=1}^{M} \underbrace{p(\lambda_b)}_{f_{\lambda_b}} \prod_{k=1}^{K} \underbrace{p(x_k)}_{f_{x_k}} \tag{B.6}$$

The factorization chosen in (B.6) can be visualized in the factor graph representation shown in Fig. B.2, where variable and factor nodes are illustrated with circles and squares, respectively.

In variational inference/message passing, we approximate the joint posterior of the variables in the system by a fully factorized auxiliary function of the form

$$q(\mathbf{x}, \lambda_1, \ldots, \lambda_M) = \prod_{k=1}^{K} q_{x_k}(x_k) \prod_{b=1}^{M} q_{\lambda_b}(\lambda_b) \tag{B.7}$$

The individual $q(\cdot)$ factors in the r.h.s. of (B.7) are then sequentially updated by minimizing the Kullback-Leibler divergence between the posterior probability function in (B.6) and the approximating auxiliary function with respect to one of the factors at a time. After convergence, they yield approximations of the posterior marginals of the system variables.

The variables λ_b model the variance of noise $\mathbf{n}$ and residual interference at antenna ports $b = 1, ..., M$. We set their prior as a Gamma distribution [14], which is the

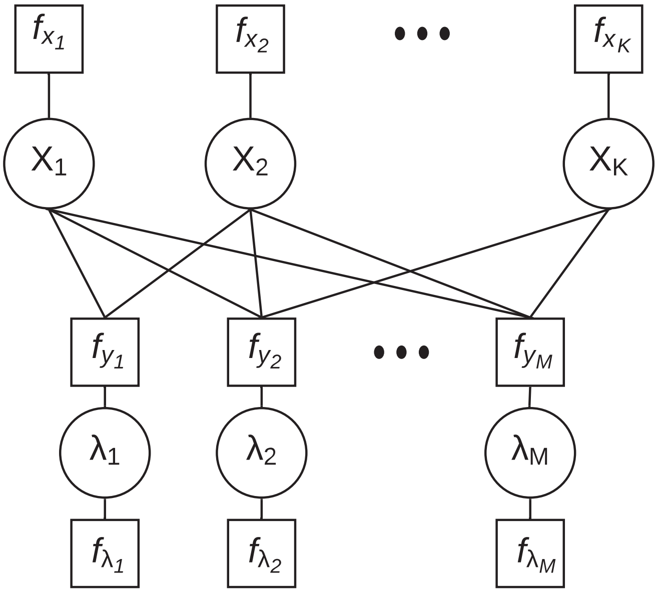


Fig. B.2: Factor graph representation of the system model.

conjugate prior for the precision of a Gaussian distribution with known mean. Hence, their pdfs read

$$f_{\lambda_b}(\lambda_b) \propto {\lambda_b}^{(\alpha_0-1)} \exp\left(-z_0\lambda_b\right) \quad \text{for}\, b \in \{1, \cdots, M\} \tag{B.8}$$

where α_0 and z_0 are respectively the shape and rate parameters. The factors f_{y_b} correspond to the pdf of the signal received at antenna ports b conditioned on the users' symbols and λ_b, reading

$$p(y_b|\mathbf{x}, \lambda_b) = \frac{\lambda_b}{\pi} \exp(-\lambda_b||y_b - \mathbf{H}_{[b,:]}\mathbf{x}||^2) \tag{B.9}$$

where $\mathbf{H}_{[b,:]}$ denotes bth row of $\mathbf{H}$, i.e. the channel for antenna element b. Finally, the prior distribution of the transmitted symbols are modeled as uniform over constellation set $\mathcal{A}$.

To begin with, and according to the definition of the VMP method, the message from factor node f_{y_b} to the variable node λ_b is

$$m_{f_{y_b} \longrightarrow \lambda_b}(\lambda_b) \propto \exp\left(\mathbb{E}_{\mathbf{x}}\{\ln\left(p(y_b|\mathbf{x}, \lambda_b)\right)\}\right) \tag{B.10}$$

where $\mathbb{E}_{\mathbf{x}}$ is the expectation with respect to the distribution given by $q_{\mathbf{x}}(\mathbf{x}) = \prod_{k=1}^{K} q_{x_k}(x_k)$.

After multiple simplifications, we reach to

$$m_{f_{y_b} \longrightarrow \lambda_b}(\lambda_b) \propto \lambda_b \exp(-\lambda_b Z) \tag{B.11}$$

where $Z = ||y_b - \sum_k \mathbf{h}_k \mu_{x_k}||^2 + \sum_k \sigma^2_{x_k} \mathbf{h}_k^H \mathbf{h}_k$ with μ_{x_k} and $\sigma^2_{x_k}$ standing for mean and variance of x_k, which is derived below. Now, we can calculate the approximate marginal probability distribution of λ_b at antenna port b by multiplying the messages entering the variable node λ_b as

$$q_{\lambda_b}(\lambda_b){=}f_{\lambda_b}{\times}m_{f_{y_b\to\lambda_b}}{=}\lambda_b^{\overbrace{\alpha_0}^{\alpha-1}}e^{-\lambda_b\overbrace{(z_0+Z)}^{\beta}} \tag{B.12}$$

Next, we consider the messages from each antenna to the symbol variables, $m_{f_{y_b}\longrightarrow x_k}$ which is

$$m_{f_{y_b\longrightarrow x_k}}{\propto}\,\mathcal{CN}\left(x_k;\frac{H_{b,k}}{|H_{b,k}|^2}(y_b{-}\sum_{k'\neq k}\mu_{x_{k'}}H_{b,k'}),\frac{1}{\mu_{\lambda_b}|H_{b,k}|^2}\right)$$

where μ_{λ_b} is the mean value of the λ_b variable which can be calculated from (B.12) as

$$\mu_{\lambda_b}|_{\substack{\alpha_0=0\\ z_0=0}}=\frac{\alpha}{\beta}=\frac{1}{|y_b{-}\sum_k H_{b,k}\mu_{x_k}|^2{+}\sum_k\sigma^2_{x_k}|H_{b,k}|^2} \tag{B.13}$$

with $H_{b,k}$ denoting the channel between antenna element b and user k. Finally, we calculate the marginal probability of the symbols of each user by multiplying messages from all the antenna elements and their prior, yielding

$$q_{x_k}(x_k)\propto\prod_{b=1}^{M}m_{f_{y_b}\longrightarrow x_k}(x_k)\times f_{x_k}(x_k). \tag{B.14}$$

From the resulting discrete distribution, we can compute the symbols mean and variance as $\mu_{x_k}=\sum_{x_k\in\mathcal{A}}x_k q(x_k)$ and μ_{x_k} and $\sigma^2_{x_k}=\sum_{x_k\in\mathcal{A}}|x_k|^2q(x_k)-|\mu_{x_k}|^2$.

Damping factor

In order to improve the convergence of our VMP-based scheme, we use a damping factor δ_{vmp} to smooth the updates of $q^{(t)}_{x_k}(x_k)$ at tth iteration as [15]

$$q^{(t)}_{x_k}(x_k)\Leftarrow\delta_{vmp}q^{(t)}_{x_k}(x_k)+(1-\delta_{vmp})q^{(t-1)}_{x_k}(x_k) \tag{B.15}$$

where the symbol "$\Leftarrow$" denotes the assignment and $\delta_{vmp}\in[0,1]$ performs a weighted average over the messages in the current and previous iterations.

3.2 MRC initialization

One of the important factors in the convergence rate of the message passing-based algorithms is the initialization. Here, we propose an initial MRC processing over the received signal and feed the soft information to the VMP algorithm for further processing.

In order to apply MRC to the received signal, we need to use $\mathbf{F}_{\text{MRC}}=\frac{\mathbf{h}_k^H}{||\mathbf{h}_k||^2}$ filter on (B.1) that yields to

$$x_k^{\text{MRC}}=\frac{\mathbf{h}_k^H}{||\mathbf{h}_k||^2}y=x_k+\sum_{k'\neq k}^{K}\frac{\mathbf{h}_k^H}{||\mathbf{h}_k||^2}\mathbf{h}_{k'}x_{k'}+\frac{\mathbf{h}_k^H}{||\mathbf{h}_k||^2}\mathbf{n} \tag{B.16}$$

Algorithm 4 VMP with MRC initialization.

```
Result: Symbol detection for all active users
Initialize: $M$, $K$, parameters in 2, $\mathcal{A}$, VMP iterations $\mathcal{I}$,$\delta_{vmp}$.
1. Generate channel matrix $\mathbf{H}$ using (B.2).
2. Calculate the initial MRC probabilities $q^0(x_k)$ using (B.17).
for $i = 1$ to $\mathcal{I}$ do
    3. Extract $\mu_x$ and $\sigma_x^2$ values from $q_{x_k}^{(i-1)}(x_k)$.
    4. Calculate the mean value of the precision parameter $\hat{\lambda}_b$ using (B.13) for
      all the antenna elements $b = \{1, \cdots, M\}$.
    5. Calculate symbol probabilities $q_{x_k}^{(i)}(x_k)$ using (B.14) for all the users
      $k = \{1, \cdots, K\}$.
    6. Update the symbol probabilities applying the damping factor in (B.15).
end
for $k = 1$ to $K$ do
    7. $\bar{x}_k = arg\max_i q_{x_k}^{(\mathcal{I})}(x_k = a_i | a_i \in \mathcal{A})$
end
```

Assuming the crowded scenario mode ($K \gg 1$), the second and third terms in x_k^{MRC} can be approximated as complex Gaussian random variable according to the central limit theory. Therefore, we set the initial marginal of x_k's as

$$q^0(x_k) = \mathcal{CN}\left(x_k; x_k^{\text{MRC}}, \frac{\sum_{k' \neq k}^{K} P_{x_{k'}} |\mathbf{h}_k^H \mathbf{h}_{k'}|^2 + ||\mathbf{h}_k||^2 \sigma_n^2}{||\mathbf{h}_k||^4}\right) \tag{B.17}$$

where $P_{x_k} = \mathbb{E}\{x_k x_k^H\}$ is user signal power.

3.3 Algorithm

Our proposed receiver is summarized in Algorithm 4, where VMP solver and the XL-MIMO system parameters are given as inputs and detected symbols are the outputs. After generating the true non-stationary MIMO channel and initial symbol probabilities by MRC, several loops in the VMP begin to exchange messages and update the variables. The first loop calculates the different parameters for the desired messages and updates them until a predefined number of iterations $\mathcal{I}$. Finally, the second loop chooses the most probable symbol for each of the users.

4 Simulation Results

In this section we evaluate the performance of the proposed algorithm and compare it with the other benchmark methods. We choose an ideal bound where perfect interference removal is done for each target user and then MRC is used for single-user detection in the interference-free channel. We call this bound "matched filter

Table B.1: Simulations parameters in detail.

Variable	Value	Variable	Value
M	512	K	256
$\mathcal{I}$	3	$\lvert\mathcal{A}\rvert$	4
$\mathbf{P}$	$\mathbf{I}$	r	Uniform$(5, 10)$
L	$29.51m$	ν	3
λ	2.6GHz	Antenna spacing	$\lambda/2$
(μ_l, σ_l)	$(2.25, 0.1)$	ζ	$M/4$
Ω	4	δ_{vmp}	0.45

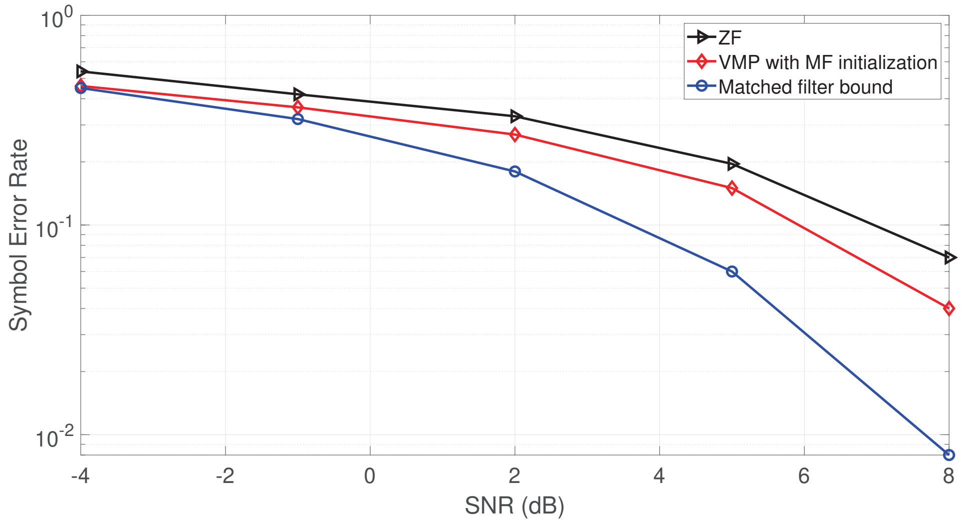


Fig. B.3: SER comparison of the different detection methods in a heavily loaded XL-MIMO system with $\frac{M}{K} = 2$

bound" [16]. All of the simulation parameters are shown in Table B.1. According to the numerical analyses VMP in our model converges at most at $\mathcal{I} = 3$. Fig. B.3 shows the SER comparison of the proposed method, ZF and the ideal single user bound. As it can be seen, the VMP based algorithm outperforms the ZF detector while keeping an acceptable gap with the ideal bound. As mentioned before, due to lack of the favorable propagation, ZF fails to work near-optimally.

4.1 Complexity Analyses

Here, we derive the computational complexity of the proposed method and the benchmark method zero-forcing (ZF). The complexity of ZF is [17]

$$C_{\text{ZF}} = \frac{K^3}{3} + MK^2 + MK \tag{B.18}$$

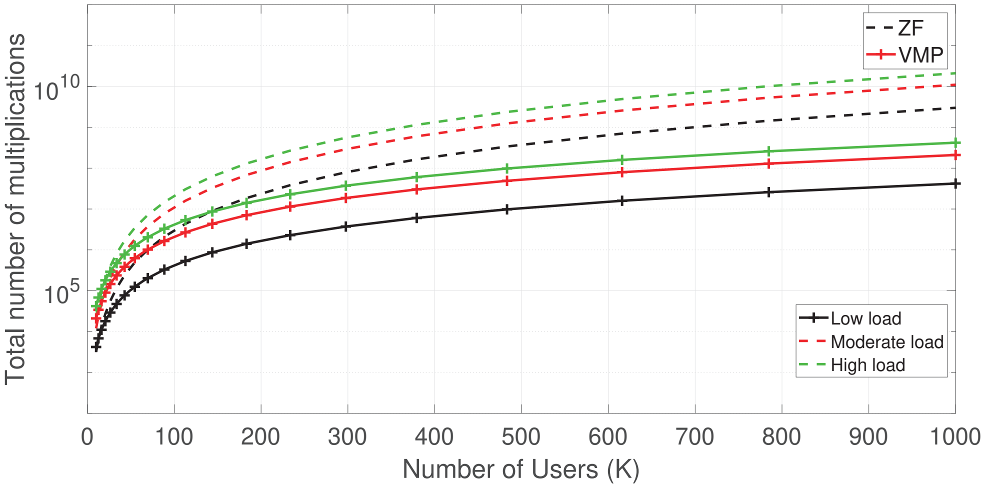


Fig. B.4: Complexity comparison of the different detection methods in three system load modes.

while the complexity of the VMP-based method is

$$C_{\text{VMP}} = \mathcal{I}(\underbrace{M(3+2K)}_{(I)} + \underbrace{MK|\mathcal{A}|}_{(II)}) + \underbrace{3MK}_{(III)} \tag{B.19}$$

where, in (I) we have 3 multiplications for updating μ_b, σ_b^2 and $\hat{\lambda}_b$ for each of the antennas plus 2 multiplications per user for deriving μ_x and σ_x^2. Then (II) stands for executing (B.14) and finally (III) is for the MRC initialization part in (B.17).

Fig. B.4 compares the complexity of these two methods in three different scenarios of high, moderate and low load regimes with M/K equal to 2, 10 and 20, respectively. The total number of the multiplications for VMP is always smaller than the ZF's and the gap grows as we approach to the crowded scenarios with much more users in the system.

5 Conclusions

In this work, we have studied the design of multi-user detection schemes for XL-MIMO systems. We have shown that VMP can be used to design a message-passing receiver with complexity that scales linearly with the number of users and antenna elements, thus making it suited for XL-MIMO systems with high system load. In addition, the detection performance surpasses that of a classical ZF detector, in spite of requiring fewer computations. Future research will address the inclusion of channel estimation together with data detection in the VMP receiver and the exploration of distributed implementations.

References

[1] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, “Scaling up mimo: Opportunities and challenges with very large arrays,” *arXiv preprint arXiv:1201.3210*, 2012.

[2] E. De Carvalho, A. Ali, A. Amiri, M. Angjelichinoski, and R. W. Heath Jr, “Non-stationarities in extra-large scale massive mimo,” *arXiv preprint arXiv:1903.03085*, 2019.

[3] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, “Energy and spectral efficiency of very large multiuser mimo systems,” *IEEE Transactions on Communications*, vol. 61, no. 4, pp. 1436–1449, April 2013.

[4] A. Amiri, M. Angjelichinoski, E. de Carvalho, and R. W. Heath, “Extremely large aperture massive mimo: Low complexity receiver architectures,” in *2018 IEEE Globecom Workshops (GC Wkshps)*, Dec 2018, pp. 1–6.

[5] P. Som, T. Datta, N. Srinidhi, A. Chockalingam, and B. S. Rajan, “Low-complexity detection in large-dimension mimo-isi channels using graphical models,” *IEEE journal of selected topics in signal processing*, vol. 5, no. 8, pp. 1497–1511, 2011.

[6] Z. Zhang, X. Cai, C. Li, C. Zhong, and H. Dai, “One-bit quantized massive mimo detection based on variational approximate message passing,” *IEEE Transactions on Signal Processing*, vol. 66, no. 9, pp. 2358–2373, 2017.

[7] H. Wang, A. Kosasih, C.-K. Wen, S. Jin, and W. Hardjawana, “Expectation propagation detector for extra-large scale massive mimo,” *arXiv preprint arXiv:1906.01921*, 2019.

[8] J. Dauwels, “On variational message passing on factor graphs,” in *2007 IEEE International Symposium on Information Theory*. IEEE, 2007, pp. 2546–2550.

[9] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.

[10] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. New York, NY, USA: Cambridge University Press, 2005.

[11] X. Gao, F. Tufvesson, and O. Edfors, “Massive mimo channels—measurements and models,” in *2013 Asilomar conference on signals, systems and computers*. IEEE, 2013, pp. 280–284.

[12] A. Adhikary, J. Nam, J. Ahn, and G. Caire, “Joint spatial division and multiplexing—the large-scale array regime,” *IEEE Transactions on Information Theory*, vol. 59, no. 10, pp. 6441–6463, Oct 2013.

[13] A. Rahmati, Y. Yapıcı, N. Rupasinghe, I. Guvenc, H. Dai, and A. Bhuyany, “Energy efficiency of rsma and noma in cellular-connected mmwave uav networks,” *arXiv preprint arXiv:1902.04721*, 2019.

[14] C. N. Manchón, G. E. Kirkelund, E. Riegler, L. P. Christensen, and B. H. Fleury, “Receiver architectures for mimo-ofdm based on a combined vmp-sp algorithm,” *arXiv preprint arXiv:1111.5848*, 2011.

[15] Y. Gao, H. Niu, and T. Kaiser, "Massive mimo detection based on belief propagation in spatially correlated channels," in *SCC 2017; 11th International ITG Conference on Systems, Communications and Coding*. VDE, 2017, pp. 1–6.

[16] C. Berrou, *Codes and Turbo Codes*. Springer Press, 2010.

[17] E. Björnson, L. Sanguinetti, J. Hoydis, and M. Debbah, "Optimal design of energy-efficient multi-user mimo systems: Is massive mimo the answer?" *IEEE Transactions on Wireless Communications*, vol. 14, no. 6, pp. 3059–3075, 2015.

Paper C

Distributed Receivers for Extra-Large Scale MIMO Arrays: A Message Passing Approach

Abolfazl Amiri, Sajad Rezaie, Carles Navarro Manchón,
Elisabeth de Carvalho

The paper is accepted for publication in the
Transactions on Wireless Communications (TWC)

Abstract

We study the design of receivers in extra-large scale MIMO (XL-MIMO) systems, i.e. systems in which the base station is equipped with an antenna array of extremely large dimensions. While XL-MIMO can significantly increase the system's spectral efficiency, they present two important challenges. One is the increased computational cost of multi-antenna processing. The second one is the presence of spatial non-stationarities in the channel response, which imply that the mean energy of a given user's signal varies across the array. Such non-stationarities limit the performance of the system. In this paper, we propose a distributed receiver for such an XL-MIMO system that can address both challenges. Based on variational message passing (VMP), we propose a set of receiver options providing a range of complexity-performance characteristics to adapt to different requirements. Furthermore, we distribute the processing into local processing units (LPU), that can perform most of the complex processing in parallel, before sharing their outcome with a central processing unit (CPU). Our designs are specifically tailored to exploit the spatial non-stationarities and require lower computations than linear receivers. Our simulation study, performed with a channel model accounting for the special characteristics of XL-MIMO channels, confirms the superior performance of our proposals compared to the state of the art methods.

Keywords— Massive MIMO, Message passing, Extra-large scale MIMO, Beyond 5G (B5G), Large intelligent surface, Spatial non-stationary, Complexity reduction

1 Introduction

Massive multiple-input multiple-output (MIMO) systems are known to have high spectral and energy efficiencies that make them a candidate for beyond fifth-generation (B5G) and 6G technologies [1–3]. Scaling up the number of antenna elements helps getting better performance, as it allows for spatially multiplexing a large number of users on the same time-frequency resources. Recently, the concept of extra-large scale MIMO (XL-MIMO) systems [4], or large intelligent surfaces (LIS) [5], has drawn attention among the researchers. Such systems can provide very high spatial resolutions leading to better quality of services for the mobile users.

One of the main obstacles limiting the possibility of increasing the dimensions of the MIMO array is the computational complexity cost. Most of the well-known conventional linear processing methods such as zero-forcing (ZF) and minimum mean squared error (MMSE) receivers have prohibitive complexity due to large matrix inversions when a large number of users is jointly served. Hence, there is a need for developing smarter multi-user detection algorithms that deal better with the higher number of users and antennas at the BS.

According to the electromagnetic propagation effects, in an array with very large dimensions spatial non-wide sense stationary properties appear [6]. In particular, the large dimension of the antenna array results in a mean energy received from a given user that may vary significantly across the array elements. On the one hand, the large separation between some of the array elements implies that their distance to the user of interest may be significantly different [4]. On the other hand, array elements that are distant from each other may experience a notably different propagation environment towards the user of interest [7]. These effects have been accounted for in proposed channel models by considering visibility regions (VRs) of a given user in the array. The VR for a given user is the subset of BS array elements that hold most of the user's received energy [6]. Presence of VRs limits the system performance compared to the conventional massive MIMO systems where VR sizes are bigger or equal to the array size [8]. Results in [9] also confirm system capacity reduction due to the existence of partially visible clusters in the massive MIMO channel. On the other hand, this property can be exploited to design smarter receivers that only consider the processing of the signals received inside the VRs [4] [10].

1.1 Literature review

The concept of low-complexity receivers for massive MIMO systems has gained a lot of attention recently. Various linear and non-linear techniques are used in the literature. The authors in [11] and [12] used a *daisy-chain* architecture and recursive methods based on approximating ZF for uplink detection and downlink precoding. The performance of these methods highly depends on the stationary conditions of the users' energy distribution over the antenna array. On the other hand, different low-complexity non-linear techniques such as message passing (MP) have been applied to massive MIMO systems as well. For instance, in [13] authors develop low complexity MP methods for MIMO inter-symbol-interference systems using graphical models. Both channel estimation and data detection problems with one-bit quantized massive MIMO are solved with variational approximate message passing (VAMP) in [14].

One of the major drawbacks of centralized processing methods in conventional massive MIMO receivers is that they lack the ability to parallelize the operations. Moreover, scaling up the dimensions of the array becomes very hard due to the high number of interconnections and heavy load on the central node. Therefore, various de-centralizing techniques have been proposed. One of them is *cell-free massive MIMO* that is trying to remove the cell boundaries and have a user-centric approach for the data-transmission [15] [16]. Unlike XL-MIMO systems, cell-free systems have considerable computational delay time [17]. For the XL-MIMO arrays, we designed a receiver in [4] using distributed units, called *sub-arrays*. First, the central processing unit (CPU) uses a method to assign users to the sub-arrays and then, the sub-arrays detect users' signals. Also, in [18], we used randomized Kaczmarz algorithm (rKA) to design a receiver with a heuristic approach to approximate the zero-forcing. All these methods work closely to the centralized linear approaches while having a performance gap with optimal solutions.

One of the ways to tackle the curse of dimensionality with the XL-MIMO arrays is to use algorithms that grow linearly with the number of the antennas at the BS. We proposed the idea of using MP for the XL-MIMO arrays to deal with its extreme dimensions [19]. Motivated by the complexity behaviour of the VMP (which scales almost linearly with the number of antennas and number of users), we managed to have a low-complexity reception in crowd scenarios. An expectation propagation (EP) based solution for symbol detection in XL-MIMO systems was presented in [20], where a sub-array structure is assumed to model the EP scheme. However, this method works fine with a small number of users.

1.2 Contributions

In this article, we propose a distributed receiver [1] structure based on variational message-passing and sequential interference cancellation for XL-MIMO systems. In our receiver, local processing units (LPU) perform in parallel soft user symbol detection by applying VMP to the signals received locally at each subarray. Then, they forward their soft symbol estimates to the CPU for the final decision. In this method, each of the LPUs can operate independently and in parallel. Moreover, SIC is used to improve the performance of the symbol detection. The CPU fuses information from all the LPUs, detects the strongest user and propagates the detected user's symbol to the LPUs. Then, each of the LPUs removes the signal contributions of the detected user. A secondary parameter called *noise precision* is estimated as well. This parameter provides a good measure of the quality of users' signals at each sub-array and thus helps to schedule the detection order. We propose multiple initialization strategies of the algorithm and investigate their convergence, performance and complexity characteristics.

The contributions of this work can be summarized as follows:

- We present a receiver architecture with flexible complexity-performance tradeoff, where we offer a set of receivers that can have low, moderate and high complexity. Depending on their level of complexity, the performance ranges from good to close to optimal. Their complexity also scales at a lower rate than the conventional central linear processing methods, such as the ZF, with regard to the number of users.
- Our receiver distributes symbol detection tasks between the CPU and the LPUs, making it possible to parallelize most of the computations.
- We introduce a generalised non-stationary channel model to capture realistic scenarios. We update the double-scattering channel model in [21] by introducing random spatial non-stationarities in it. This model is based on various measurements data and can model several scenarios and is a more accurate representation for XL-MIMO channels than typically used i.i.d models.
- We present various VR-aware receivers that account for the spatial non-stationarities of XL-MIMO channels. These receivers work in a more efficient way by obtaining much better performance than the centralized methods and working very closely to the normal-data fusion VMP method while keeping the complexity limited. To the best of our knowledge, this is the first work that exploits the non-stationarities to design a receiver obtaining close to optimal performance.

In our previous work [19], we proposed a centralized VMP architecture for the XL-MIMO arrays. We investigated a crowded scenario in a relatively simple channel model. This algorithm does not allow distributed processing at local units and has no complexity-performance flexibility. In contrast, in this paper we introduce a set of distributed receivers that work in co-operation in an iterative manner. We present a task scheduling technique between the central and local processing units

[1] The term 'Distributed receivers' is referring to the processing techniques that are used throughout the paper.

to enable low-complexity parallel processing. Moreover, we propose the VR-aware receivers that exploit channel non-stationarities to lower the complexity even more with a small cost in the performance. We study a general channel model and evaluate the performance in various scenarios with respect to different system parameters. We also add channel coding in our results and compare the error rate of our methods with [19] to show the improvement of the new receivers.

1.3 Paper structure

First, we start with discussing the channel model in an XL-MIMO system considering the random non-stationarities in Section 2. Then, we explain the principles of variational inference, our problem formulation, the distributed VMP scheme and different initialization techniques in Section 3. After designing the message passing structure, we aim to detect the transmitted symbols in Section 4. There, we discuss about different data fusion methods and symbol detection techniques taking place at the CPU. We conclude our paper with the simulations results in Section 5 and the conclusions section.

1.4 Notations

Capital calligraphic letters $\mathcal{A}$ denote finite sets. The cardinality of a set is denoted by $|\mathcal{A}|$. $\mathcal{X} \setminus \mathcal{Y}$ is set $\mathcal{X}$ from which set $\mathcal{Y}$ is excluded. Boldface small and capital letters stand for vector and matrix representations, respectively. $(.)^H$ is matrix conjugate transpose operator. $\mathbf{I}_s$ is an identity matrix of size $s \times s$. $f(x) \propto g(x)$ denotes that $f(x) = ag(x)$ for some positive constant a. $\mathbb{E}_x$ is the expectation operator over variable x and $\bar{\mathbf{x}}$ is the mean value of $\mathbf{x}$. $\mathrm{CN}(x; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ as the probability distribution function (pdf) of a multi-variate complex Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ over a variable x. $\mathcal{CN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a multi-variate complex Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. $\mathcal{U}(a, b)$ is a uniform distribution in $[a, b]$ interval.

2 System Model

We consider a narrow-band MIMO system where K single-antenna active users transmit in the uplink to a base station (BS) with M antenna elements. User symbols are denoted with the vector $\mathbf{x} \in \mathbb{C}^K$ with entries taking values from the complex constellation set $\mathcal{A} = \{a_1, a_2, \cdots, a_{|\mathcal{A}|}\}$. $\mathbf{H} = [\boldsymbol{h}_1, \cdots, \boldsymbol{h}_K] \in \mathbb{C}^{M \times K}$ is the channel matrix and has column vectors $\boldsymbol{h}_k$ each of which denote the channel for user k . At the BS, the noise is assumed to have circularly symmetric complex Gaussian distribution $\mathbf{n} \sim \mathcal{CN}(\mathbf{0}, \sigma_n^2 \mathbf{I}_M) \in \mathbb{C}^M$. We model the received baseband signal $\mathbf{y} \in \mathbb{C}^M$ across the whole array as follows:

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}. \tag{C.1}$$

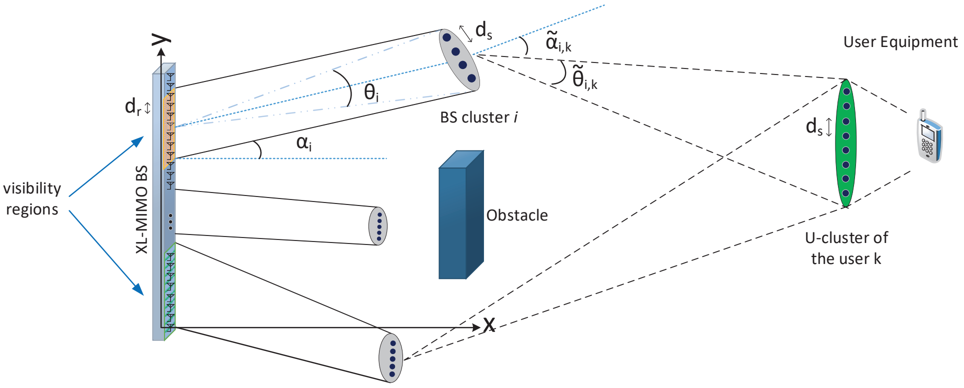


Fig. C.1: An illustrative example of the propagation model in a XL-MIMO array. Spatial non-stationarities are appearing along the array where most of users' energies are concentrated in the VRs. There are several local clusters at the BS side and one cluster per user at each user's side. The interaction between these clusters determines the characteristics of the non-stationarities.

The BS is made of a set of B sub-arrays each with $M_b = \frac{M}{B}$ antennas. Here, we define $\tilde{\mathbf{H}}_b \in \mathbb{C}^{M_b \times K}$ and $\mathbf{y}_b \in \mathbb{C}^{M_b}$ as the channel matrix and the received signal in the b-th sub-array for $b \in \{1, \cdots, B\}$, respectively.

2.1 Channel Model

Fig. C.1 shows the channel model for the XL-MIMO system accounting for the non-stationary properties of the propagation environment. We adopt a specific channel model called *double-scattering* model [21]. In this channel, correlation is allowed at both the transmitter and the receiver side. In the double scattering channel model, there are two types of scattering clusters in the propagation environment: the one located at the BS side called *BS-cluster* and one located at the user side called *U-cluster*. Signals emitted by the users first impinge on the U-cluster, which scatters them towards multiple BS-clusters directing the signal to the BS array.

As mentioned before, unlike conventional massive MIMO arrays, XL-MIMO arrays can have a large number of antennas spanning hundreds of wavelengths in space. By decomposing the propagation channel into scatterers, it is observed that the scatterers are not visible over the whole array. This causes variations in the received energy on the array and therefore the appearance of the spatial non-stationarities [6] [7]. The main difference between this model and the model in [21] is that we have several BS-clusters due to the large array size at the BS. Each of these clusters is seeing a subset of the antennas. Furthermore in our model, unlike the previous models proposed in the XL-MIMO literature, we distinguish between the correlation matrices at the receiver and the user side. This decomposition allows us to model wider ranges of dynamics in our MIMO channel. For instance, in Section 5.3, we evaluate the performance of our receivers in channels with different correlation characteristics at both user and the BS sides.

The channel between the user k and the BS is modeled as

$$\mathbf{h}_k = \left[\tilde{\boldsymbol{h}}_{1,k}, \cdots, \tilde{\boldsymbol{h}}_{C,k}\right] \boldsymbol{D}_k\, \mathbf{g}_k \in \mathbb{C}^{M\times 1},\ \forall\, k \in \{1, \cdots, K\} \tag{C.2}$$

where $\tilde{\boldsymbol{h}}_{i,k} \in \mathbb{C}^{M\times S_i}$ denotes the sub-channel for the ith BS-cluster with S_i scatterers, $\boldsymbol{D}_k \in \{0,1\}^{S'\times S}$ is in charge of assigning the visible BS-clusters to the U-cluster, and the entries of $\mathbf{g}_k \sim \mathcal{CN}(\mathbf{0}, \boldsymbol{I}_S) \in \mathbb{C}^{S\times 1}$ model the small-scale fading between user k and the S scatterers in its U-cluster. Moreover, we assume that there are C BS-clusters in the propagation channel and S' scatterers at the BS side visible to the U-cluster. We formulate the sub-channel of the ith BS-cluster and the U-cluster associated to user k as

$$\tilde{\boldsymbol{h}}_{i,k} = \mathbf{Y}_i \boldsymbol{\rho}_i^{\frac{1}{2}} \boldsymbol{R}_i^{\frac{1}{2}} \boldsymbol{G}_i \tilde{\boldsymbol{R}}_{i,k}^{\frac{1}{2}} \quad \in \mathbb{C}^{M\times S_i},\ \forall i \in \{1, \cdots, C\} \text{ and } k \in \{1, \cdots, K\} \tag{C.3}$$

where $\mathbf{Y}_i \in \{0,1\}^{M\times r_i}$ determines indices of the antennas at the BS that are visible to the i-th BS-cluster with r_i as the number of visible antennas. $\boldsymbol{\rho}_i \in \mathbb{C}^{r_i\times r_i}$ is the visibility gain matrix, $\boldsymbol{R}_i \in \mathbb{C}^{r_i\times r_i}$ and $\boldsymbol{G}_i \in \mathbb{C}^{r_i\times S_i}$ are the correlation matrix and the complex scattering amplitudes between the BS and the BS-cluster i, respectively. Also, $\tilde{\boldsymbol{R}}_{i,k} \in \mathbb{C}^{S_i\times S_i}$ is the correlation matrix between the i-th BS-cluster and the U-cluster for user k. In the following we will discuss each of the channel components in (C.2) and (C.3) in detail.

Cluster VR and power distribution

We have two types of VRs in our propagation channel: cluster VR and user VR. The cluster VR is defined as an antenna region on the BS array that is visible to a cluster. On the other hand, the user VR is a set of the clusters that are being seen by a user. The antenna region for the cluster VR has a random center c_i (indicating position of the center of the VR in meters) and a random length l_i consisting of r_i consecutive antennas on the BS in the interval $[c_i - \frac{l_i}{2}, c_i + \frac{l_i}{2}]$ with d_r as the BS antenna spacing. These antennas are belonging to a index set of $\mathcal{R}_i = \{a_1^i, \cdots, a_{r_i}^i\}$ where a_j^i is the j-th antenna element inside the i-th cluster VR starting from $a_1^i = \lfloor \frac{c_i - \frac{l_i}{2}}{d_r} \rfloor$ ($|\mathcal{R}_i| = r_i$). Note that for two-dimensional arrays, the cluster VR will be an area on the array. We have C cluster VRs that can overlap. Concerning the size of the array, the BS-clusters are all partially visible. This indicates that none of them can see all of the antennas at the BS or in other words, $|\mathcal{R}_i| \neq M\ \forall\, i \in \{1, \cdots, C\}$. We indicate the relation between the BS antennas and each of the BS-clusters with the antenna association matrix $\mathbf{Y}_i \in \{0,1\}^{M\times r_i}$ obtained by:

$$\mathbf{Y}_i = \begin{bmatrix} \mathbf{0}_{m_1^i\times r_i} \\ \boldsymbol{I}_{r_i} \\ \mathbf{0}_{m_2^i\times r_i} \end{bmatrix} \qquad,\ m_1^i = a_1^i - 1 \text{ and } m_2^i = M - a_{r_i}^i \tag{C.4}$$

where $\mathbf{0}_{m^i\times r_i}$s are zeros matrices and the identity matrix $\boldsymbol{I}_{r_i}$ starts from the a_1^i-th row, i.e. from the first antenna of the VR. For example, in a system with $M = 5$ antennas and a VR cluster covering antennas in $\mathcal{R}_1 = \{2, 3, 4\}$, this association matrix is

$$\mathbf{Y}_1 = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

The reason for using this matrix is to map the antenna elements into the individual sub-channels for each of the BS-clusters.

The energy distribution is not constant inside each of the cluster VRs. We refer to the measurements in [7] to model the variations in the received energy from cluster i within its VR. According to these measurements, energy peak happens at c_i and then it attenuates linearly, i.e with a constant slope ψ_i (dB/m), in a logarithmic scale per distance unit inside the VRs. We call the energy distribution *visibility gain* and can be calculated using a discrete function such as:

$$\rho_i[n] = \begin{cases} 10^{-\psi_i |c_i - d_r(n-1)|} & n \in \mathcal{R}_i \\ 0 & n \notin \mathcal{R}_i \end{cases} \tag{C.5}$$

where n is the index of the antenna elements inside the cluster VR. We assume a BS array with antennas located on the y-axis starting from the origin (see Fig. C.1). A small portion (less than 5%) of the reflected energy from the clusters is spread outside of the VR. For the sake of simplicity in our model, we assume that this energy is zero. Measurements in [7] suggest a uniform distribution for c_i along the BS array, a normal distribution for the slope of the gains ψ_i and a log-normal distribution for the cluster VR size l_i. The measured parameters for these distributions that are used for the simulations are presented in 5. Knowing the energy variations inside the VRs, we define the visibility gain matrix $\boldsymbol{\rho}_i = \mathrm{diag}(\rho_i[a^i_j] | a^i_j \in \mathcal{R}_i) \in \mathbb{C}^{r_i \times r_i}$ that stores the visibility gains for all of the antennas inside the i-th VR in its diagonal entries.

BS-Clusters

Aiming to model the correlation between the elements on the receiver side (at the BS antennas), we define the correlation matrix for the BS-cluster i $\mathbf{R}_i \in \mathbb{C}^{r_i \times r_i}$. This matrix can be calculated as the following where its (m, l) element is [21]:

$$[\mathbf{R}_i]_{m,l} = \frac{1}{S_i} \sum_{n=\frac{1-S_i}{2}}^{\frac{S_i-1}{2}} \exp\left(-2\pi j(m-l)d_r \cos(\pi/2 + \alpha_i + \frac{n\theta_i}{S_i - 1})\right) \quad \forall\, m, l \in \{1, \cdots r_i\} \tag{C.6}$$

where, θ_i is the angular spread and α_i is the azimuth angle between the BS and the i-th BS-cluster. We assume that the BS-clusters are located randomly in the $x - y$ plane and the x-axis is the referral line for all the azimuth angles in this paper. Moreover, the small-scale fading modelling the complex scattering amplitudes between the BS and the i-th BS-cluster is $\mathbf{G}_i \in \mathbb{C}^{r_i \times S_i}$ and each of its i.i.d entries follow a complex Gaussian distribution $\mathcal{CN}(\mathbf{0}, \boldsymbol{I}_{S_i})$.

In order to indicate the indices of the scatterers in all the BS-clusters, we use an index set. Without loss of generality, we assume that the scatterers of all the BS-clusters are indexed and stored in scattering set $\mathcal{S}' = \{1, 2, \cdots, S'\}$ with $S' = \sum_{i=1}^{C} S_i$.

U-Cluster k

This cluster models the scatterers around the user equipment (UE) such as buildings, cars and trees. We assume that each user only sees one U-cluster with S scatterers. These scatterers are viewed as an array of S virtual antennas that have an average spacing of d_s. For simplicity, we assume the same d_s size for both BS and U clusters. The correlation matrix between the i-th BS-cluster and the U-cluster for user k is $\tilde{\mathbf{R}}_{i,k} \in \mathbb{C}^{S_i \times S_i}$. The (m, l) element of this matrix is computed as:

$$[\tilde{\mathbf{R}}_{i,k}]_{m,l} = \frac{1}{S_i} \sum_{n=\frac{1-S_i}{2}}^{\frac{S_i-1}{2}} \exp\left(-2\pi j(m-l)d_s \cos(\pi/2 - \tilde{\alpha}_{i,k} + \frac{n\tilde{\theta}_{i,k}}{S_i - 1})\right) \quad \forall\, m, l \in \{1, \cdots S_i\} \tag{C.7}$$

where, $\tilde{\alpha}_{i,k}$ is the azimuth angle between the i-th BS-cluster and the U-cluster of user k and $\tilde{\theta}_{i,k}$ is the corresponding angular spread between them.

The visibility matrix

Due to the randomness and obstacles in the environment, only a subset of the C BS-clusters (See Fig. C.1) are visible to the U-cluster k. We denote the user VR by $\mathcal{V}_k$ for each user k containing the indices of the BS-clusters visible for the U-cluster k. We denote the scatterer visibility set for user k as $\mathcal{S}_k \subset \mathcal{S}'$ that stores all the indices of the scatterers of the BS-clusters in $\mathcal{V}_k$. Now, We can define the visibility matrix $\boldsymbol{D}_k \in \{0, 1\}^{S' \times S}$ and its m-th row is calculate as:

$$[\boldsymbol{D}_k]_{(m,:)} = \begin{cases} \mathbf{1}_{1\times S} & \text{if } m \in \mathcal{S}_k \\ \mathbf{0}_{1\times S} & \text{otherwise} \end{cases} \tag{C.8}$$

This matrix shows the visibility of the BS-cluster scatterers to the U-cluster. As an example, assume 3 BS-clusters each with 2 scatterers and a U-cluster with 5 scatterers and $\mathcal{V}_1 = \{1, 3\}$. Thus, the resulting scatterer visibility set is $\mathcal{S}_1 = \{1, 2, 5, 6\}$ and the visibility matrix is

$$\boldsymbol{D}_1 = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Finally, we can assemble the channel for each user k in (C.2) using different channel components from equations (C.3)-(C.8). The scatterer visibility set of user k, $\mathcal{S}_k$, is different and independent from the rest of the users. Thus, users can share some of the BS-clusters depending on these sets. Eventually, the complete channel matrix can be calculated as $\mathbf{H} = [\boldsymbol{h}_1, \cdots, \boldsymbol{h}_K]$.

3 Variational Message Passing

In this section, we introduce the basics of the VMP method before formulating the problem of estimating the transmitted symbols. Finally, we derive the messages for the VMP algorithm.

3.1 Variational Inference and Variational Message Passing

Let $p(\boldsymbol{z}, \boldsymbol{x})$ denote a joint probability density function (pdf), where $\boldsymbol{z} = \{z_1, \ldots, z_N\}$ denotes a set of unobserved variables and $\boldsymbol{x}$ a set of observed variables. Our goal is to find estimates of the variables in $\boldsymbol{z}$ from their marginal posterior pdfs $p(z_i|\boldsymbol{x})$. However, finding these pdfs is often too complex or intractable. Instead, we resort to computing a surrogate distribution $q(\boldsymbol{z})$ that approximates the posterior $p(\boldsymbol{z}|\boldsymbol{x})$ and from which marginals $q_i(z_i)$ can be easily found. This is obtained in variational inference by minimizing their Kullback-Leibler (KL) divergence, defined as

$$D(q||p) \triangleq \int q(\boldsymbol{z}) \log \frac{q(\boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{x})} dz. \tag{C.9}$$

To make the problem tractable, the surrogate function q is restricted to fulfill certain constraints. Typically, the mean-field approximation, which considers a fully factorized distribution of the form

$$q(\boldsymbol{z}) = \prod_{i=1}^{N} q_i(z_i) \tag{C.10}$$

is applied, in addition to normalization constraints $\int q_i(z_i) dz_i = 1$. With these constraints, a sequential minimization of the KL divergence with respect to each of the factors q_i is performed. It can be shown [22] that, at each step, the optimal factor q_i given all other factors $q_j, j \neq i$ is obtained by

$$q(z_i) \propto \exp\left(\mathbb{E}_{j\neq i}\{\ln p(\boldsymbol{z}, \boldsymbol{x})\}\right) \tag{C.11}$$

where the expectation is taken with respect to all approximate marginals $q_j, j \neq i$. The above update rule is applied alternately to the different factors until convergence is achieved.

This algorithm can also be formulated in terms of a message passing algorithm. Assume that the joint distribution factorizes as

$$p(\boldsymbol{z}, \boldsymbol{x}) = \prod_{a} f_a(\boldsymbol{z}_a, \boldsymbol{x}_a) \tag{C.12}$$

where $\boldsymbol{z}_a \subseteq \boldsymbol{z}$ and $\boldsymbol{x}_a \subseteq \boldsymbol{x}$ are subsets of the unobserved and observed variables. All of the $f_a(\boldsymbol{z}_a, \boldsymbol{x}_a)$s are the factors in the joint pdf and they depend on the statistical dependencies in the model. The factorization is not unique, as several factors can be combined together. Moreover, this factorization can be graphically represented as a factor graph which we will introduce it in the next subsection. The update in (C.11) can be expressed in terms of messages passed along the edges of the factor graph [22] as in

$$q_i(z_i) \propto \prod_{f_a \in \mathcal{N}(z_i)} m_{f_a \to z_i}(z_i) \tag{C.13}$$

where $\mathcal{N}(z_i)$ denotes the set of factors in (C.12) that contain variable z_i, and the messages read

$$m_{f_a \to z_i}(z_i) = \exp\left(\mathbb{E}_{j \neq i}\{\ln f_a(\boldsymbol{z}_a, \boldsymbol{x}_a)\}\right). \tag{C.14}$$

3.2 Probabilistic System Description

To apply the VMP inference described above, we first formulate a probabilistic model of the system. Our ultimate goal is to infer the values of the transmitted symbols x_k, $k = 1, \ldots, K$ and of the unknown noise precision (inverse of the noise variance) $\lambda = \frac{1}{\sigma_n^2}$. Ideally, this should be done from their joint posterior distribution, which reads

$$p(\mathbf{x}, \lambda|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{x}, \lambda)p(\mathbf{x})p(\lambda) \tag{C.15}$$

where, due to the white Gaussian noise, $p(\mathbf{y}|\mathbf{x}, \lambda) = \mathrm{CN}(\mathbf{y}; \mathbf{H}\mathbf{x}, \frac{1}{\lambda}\mathbf{I}_M)$. The prior symbol probability mass function (pmf) reads $p(\mathbf{x}) = \prod_k p(x_k)$, with $p(x_k)$ being uniform over the constellation set $\mathcal{A}$, and we assume the noise precision to have a non-informative Gamma prior.

As we are interested in performing as much of the processing locally at each of the BS sub-arrays, however, we formulate instead B similar models, one for each of the sub-arrays:

$$p(\mathbf{x}^b, \lambda_b|\mathbf{y}_b) \propto \underbrace{p(\mathbf{y}_b|x_1^b, \cdots, x_K^b, \lambda_b)}_{f_{\mathbf{y}_b}} \underbrace{p(\lambda_b)}_{f_{\lambda_b}} \prod_{k=1}^{K} \underbrace{p(x_k^b)}_{f_{x_k^b}}, \qquad b = 1, \ldots, B. \tag{C.16}$$

where the variables $\mathbf{x}^b = [x_1^b, \ldots, x_K^b]^T$ and λ_b denote respectively the transmitted symbols and noise precision observed by the b-th BS sub-array,
$f_{\mathbf{y}_b}(\mathbf{x}^b, \lambda_b) = \mathrm{CN}(\mathbf{y}_b; \tilde{\mathbf{H}}_b\mathbf{x}^b, \frac{1}{\lambda_b}\mathbf{I}_{M_b})$, $f_{x_k^b}(x_k^b) = \frac{1}{|\mathcal{A}|}\mathbf{1}(x_k^b \in \mathcal{A})$, and $f_{\lambda_b} \propto 1/\lambda_b$ [2]. Although, clearly, $\mathbf{x}^b$ and λ_b represent the same random variables for the different sub-arrays $b = 1, \ldots, B$, we treat them separately here such that each sub-array can, in their local processing phase, obtain independent estimates of them based on solely their received signals $\mathbf{y}_b$. After each local processing phase, their respective estimates are fused in a CPU and distributed back to the sub-arrays. The factor graphs illustrating the models in (C.16) and their linking with the CPU are depicted in Fig. C.2.

[2] This choice of prior corresponds to an improper, noninformative Gamma prior distribution with shape and rate parameters approaching zero.

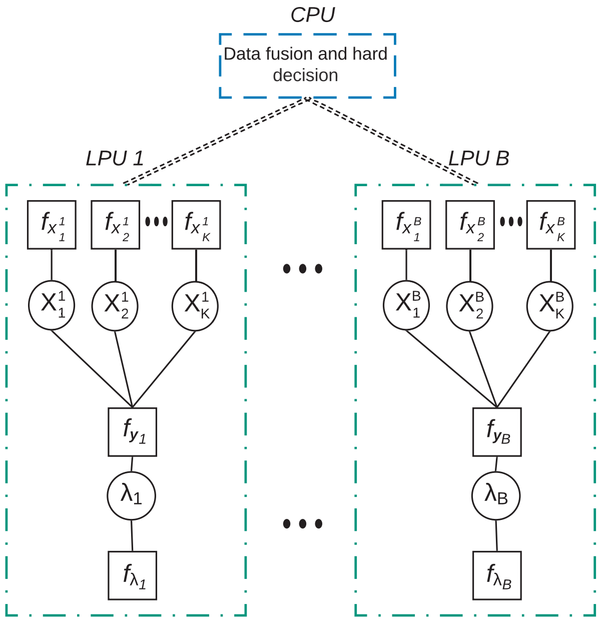


Fig. C.2: Factor graph representation of the local processing units and the central unit. Local estimations obtained by the local units are sent to the central unit for the data fusion and detection.

3.3 VMP at the Local Processing Units

We proceed in this section to describing the VMP algorithm run at each of the LPUs at the sub-arrays. The b-th LPU aims at approximating the posterior in (C.16) by using the approximate distribution

$$q_b(\mathbf{x}^b, \lambda_b) = q_{\lambda_b}(\lambda_b) \prod_{k=1}^{K} q_{x_k^b}(x_k^b) \tag{C.17}$$

where the naïve mean-field approximation [23] is applied. At each of the local processing rounds, an initial setting for the factors $q_{x_k^b}(x_k^b)$, $k = 1, \ldots, K$ is available, with different initialization strategies discussed in Section 3.4. In the first step, the message from factor node $f_{\mathbf{y}_b}$ to the variable node λ_b is calculated as [24] [22]

$$m_{f_{\mathbf{y}_b \longrightarrow \lambda_b}}(\lambda_b) \propto \exp\left(\mathbb{E}_{\mathbf{x}^b}\{\ln\left(f_{\mathbf{y}_b}(\mathbf{x}^b, \lambda_b)\right)\}\right) \propto {\lambda_b}^{M_b} \exp(-\lambda_b Z_b) \tag{C.18}$$

where $\mathbb{E}_{\mathbf{x}^b}\{\cdot\}$ denotes the expectation with respect to the initial distribution $q_{\mathbf{x}^b}(\mathbf{x}^b) = \prod_{k=1}^{K} q_{x_k^b}(x_k^b)$, and $Z_b = ||\mathbf{y}_b - \sum_k \tilde{\mathbf{h}}_{b,k}\bar{x}_k^b||^2 + \sum_k \sigma^2_{x_k^b}\tilde{\mathbf{h}}_{b,k}^H\tilde{\mathbf{h}}_{b,k}$. In this expression, $\tilde{\mathbf{h}}_{b,k}$ denotes the kth column of $\tilde{\mathbf{H}}_b$ while $\bar{x}_k^b = \sum_{s\in\mathcal{A}} s q_{x_k^b}(s)$ and $\sigma^2_{x_k^b} = \sum_{s\in\mathcal{A}} |s|^2 q_{x_k^b}(s) - |\bar{x}_k^b|^2$ are the mean and variance of x_k^b with respect to $q_{x_k^b}(x_k^b)$. The approximate marginal distribution $q_{\lambda_b}(\lambda_b)$ is then obtained by multiplying the messages entering the variable node λ_b as

$$q_{\lambda_b}(\lambda_b){=}f_{\lambda_b}(\lambda_b)\times m_{f_{\mathbf{y}_b\rightarrow\lambda_b}}(\lambda_b) = \lambda_b^{\overbrace{M_b-1}^{\alpha-1}} e^{-\lambda_b \overbrace{(Z_b)}^{\beta}} \tag{C.19}$$

which correspondes to a Gamma distribution with mean

$$\bar{\lambda}_b = \frac{\alpha}{\beta} = \frac{M_b}{Z_b}. \tag{C.20}$$

Next, the LPU computes the messages from factor node $f_{\mathbf{y}_b}$ to the variable nodes x_k^b, which result in

$$m_{f_{\mathbf{y}_b\longrightarrow x_k^b}} \propto \exp\left(\mathbb{E}_{\lambda_b,\mathbf{x}^b_{\backslash k}}\{\ln\left(f_{\mathbf{y}_b}(\mathbf{x}^b,\lambda_b)\right)\}\right) \propto \mathcal{CN}\left(x_k^b;\frac{\tilde{\mathbf{h}}_{b,k}^H}{||\tilde{\mathbf{h}}_{b,k}||^2}(\mathbf{y}_b-\sum_{k'\neq k}\bar{x}_{k'}^b\tilde{\mathbf{h}}_{b,k'}),\frac{1}{\bar{\lambda}_b||\tilde{\mathbf{h}}_{b,k}||^2}\right) \tag{C.21}$$

where, similarly as in (C.18), $\mathbb{E}_{\lambda_b,\mathbf{x}^b_{\backslash k}}\{\cdot\}$ denotes the expectation with respect to $q_{\lambda_b}(\lambda_b)$ and $\prod_{k'\neq k} q_{x_{k'}^b}(x_{k'}^b)$ [24] [22].

To finalize, the approximate marginals of the symbols of each user at the sub-array b are obtained by multiplying these messages with their local priors, yielding

$$q_{x_k}^b(x_k^b) \propto m_{f_{\mathbf{y}_b}\longrightarrow x_k^b}(x_k^b) \times f_{x_k^b}(x_k^b). \tag{C.22}$$

3.4 Initialization Options

As mentioned above, VMP requires initial approximate symbol distributions $q^0_{x_k^b}(x_k^b)$ to begin its operation, which we review next.

Type of the initialization

The simplest option is to initialize the algorithm with a uniform distribution where all the symbols are, a priori, equiprobable for all users. In this case, the initial distributions are set as $q^0_{x_k^b}(x_k^b) = \frac{1}{|\mathcal{A}|}\mathbf{1}(x_k^b \in \mathcal{A})$, $\forall k \in \{1,\cdots,K\}$ and $\forall b \in \{1,\cdots,B\}$. This method has no computational complexity, but typically results in slow convergence of the VMP algorithm.

The performance and convergence speed of the local processing can be improved by using linear processing techniques to set the initial symbol distributions. As a first option, we consider maximum ratio combining (MRC) over all the BS array. Applying the MRC for user k to the received signal in (C.1) yields

$$\hat{x}_k^{\text{MRC}} = \frac{\mathbf{h}_k^H}{||\mathbf{h}_k||^2}\mathbf{y} = x_k + \sum_{k'\neq k}^{K}\frac{\mathbf{h}_k^H}{||\mathbf{h}_k||^2}\mathbf{h}_{k'}x_{k'} + \frac{\mathbf{h}_k^H}{||\mathbf{h}_k||^2}\mathbf{n} \tag{C.23}$$

Assuming a large number of users ($K \gg 1$), the sum of the second and third terms in (C.23) can be approximated as a complex Gaussian random variable according to the central limit theory. The initial approximate marginals of the symbols x_k's are therefore set as proportional to a Gaussian pdf, restricted to the symbol alphabet $\mathcal{A}$, i.e.

$$q^0_{x^b_k}(x^b_k) \propto \mathrm{CN}\left(x^b_k; \hat{x}^{\mathrm{MRC}}_k, \frac{\sum_{k' \neq k}^K P_{x_{k'}} |\mathbf{h}_k^H \mathbf{h}_{k'}|^2 + ||\mathbf{h}_k||^2 \sigma_n^2}{||\mathbf{h}_k||^4}\right) \mathbf{1}(x^b_k \in \mathcal{A}), \quad \forall b \in \{1, \ldots, B\}. \tag{C.24}$$

where $P_{x_k} = \mathbb{E}\{x_k x_k^*\}$ is user signal power. This initialization introduces a complexity load of $3MK$ multiplications. There is also another possibility to apply the MRC initialization locally and at each of the LPUs. For this type, local channel vectors $\tilde{\boldsymbol{h}}_{b,k}$ and $\tilde{\boldsymbol{h}}_{b,k'}$ and local received signal $\boldsymbol{y}_b$ are used in (C.23) for each sub-array b. Then, local estimates are calculated using (C.24) for all of the B sub-arrays.

Another candidate for the initialization is the ZF method. The ZF receiver filter for user k, denoted by $\mathbf{F}_{\mathrm{ZF},k}$, reads

$$\mathbf{F}_{\mathrm{ZF},k}[\mathbf{H}] = \frac{\mathbf{h}_k^H \mathbf{P}^{\perp}_{\bar{\mathbf{H}}_k}}{\mathbf{h}_k^H \mathbf{P}^{\perp}_{\bar{\mathbf{H}}_k} \mathbf{h}_k}, \tag{C.25}$$

with $\mathbf{P}^{\perp}_{\bar{\mathbf{H}}_k} = \mathbf{I} - \bar{\mathbf{H}}_k (\bar{\mathbf{H}}_k^H \bar{\mathbf{H}}_k)^{-1} \bar{\mathbf{H}}_k^H$ [25]; $\bar{\mathbf{H}}_k$ is obtained from $\mathbf{H}$ by removing its k^{th} column $\mathbf{h}_k$. The resulting estimates after application of the ZF filter are

$$\hat{x}^{\mathrm{ZF}}_k = \mathbf{F}_{\mathrm{ZF},k} \mathbf{y} = x_k + \mathbf{F}_{\mathrm{ZF},k} \mathbf{n} \tag{C.26}$$

and they have mean equal to the transmitted symbol x_k and a variance given by

$$\sigma^2_{x_k^{\mathrm{ZF}}} = \sigma_n^2 \left(\mathbf{h}_k^H \mathbf{P}^{\perp}_{\bar{\mathbf{H}}_k} \mathbf{h}_k \right)^{-1}. \tag{C.27}$$

Similarly as for MRC initialization, we approximate the inital marginals as

$$q^0_{x^b_k}(x^b_k) \propto \mathrm{CN}\left(x^b_k; \hat{x}^{\mathrm{ZF}}_k, \sigma^2_{x_k^{\mathrm{ZF}}}\right) \mathbf{1}(x^b_k \in \mathcal{A}), \qquad \forall b \in \{1, \ldots, B\}. \tag{C.28}$$

A last option is to perform similar ZF initialization but applied locally at each of the LPUs. In this case, the ZF filter for the bth sub-array is computed analogously to (C.25) but using channel matrix $\tilde{\mathbf{H}}_b$ instead of $\mathbf{H}$. After this, an approximate marginal similar to that in (C.28) is calculated for each of the B sub-arrays.

Strategy

In this subsection, we present two different modes to initialize the VMP method. The first option is to initialize the VMP just a single time. We call this mode *One-time* initialization. The second option is to initialize the algorithm multiple times. This mode is done to help stabilizing the outputs within the consecutive iterations of the VMP method. This mode becomes more interesting when we initialize the VMP at each step of the interference cancellation detection. There, after each step of the interference removal, the linear pre-processing used for the initialization will perform more accurately and improve the performance of the overall scheme. With this, we finalize the description of the processing of the LPUs of each sub-array, which is summarized in Algorithm 5.

```
Algorithm 5 VMP at each of the LPUs.
Result: Local symbol estimates for all active users
Initialize: $M$, $K$, $\mathbf{y}$, sub-array index $b$ , parameters in Sec. 2, $\mathcal{A}$, VMP iterations $\mathcal{I}$.
1. Get the corresponding channel matrix for sub-array $b$ from $\tilde{\mathbf{H}}_b$ that is generated
   using (C.2).
2. Choose one of the initialization methods in (3.4) set the initial probabilities as
   $q^0(x_k)$.
for $i = 1$ to $\mathcal{I}$ do
    3. Extract $\bar{x}_k^b$ and $\sigma^2_{x_k^b}$ values from $q_{x_k}^{(i-1)}(x_k)$.
    4. Calculate the mean value of the precision parameter $\bar{\lambda}_b$ using (C.20) for the
       sub-array.
    5. Calculate symbol probabilities $q_{x_k}^{(i)}(x_k)$ using (C.22) for all the users $k =
       \{1, \cdots, K\}$.
end
6. Finalize the local information to be sent to the CPU $q_{x_k}^b(x_k)$=$q_{x_k}^{(i)}(x_k)$.
```

4 Data Fusion and Symbol Detection

In this section, we detail how the results of the local VMP processing performed by the LPUs at each of the sub-arrays are combined at the CPU to yield the final symbol estimates. The overall receiver process is illustrated in the block diagram in Fig. C.3. The operations are divided between the LPUs and the CPU while offering each of them several options. At the LPUs illustrated in the left-hand side of the diagram, VMP processing is performed as discussed in Section 3, including the different initialization options. On the right-hand side of the diagram, the CPU is illustrated as having two basic tasks: the fusion of the symbol estimates provided by the different LPUs, and the eventual detection of the symbols by using the fused information. Four different options are studied for the data fusion process, and two options are considered for detection: non-iterative, and SIC based. In the latter case of SIC based detection, several iterations of local and central processing are performed before the symbols of all users are detected, which is illustrated in the diagram by the feedback connection between the CPU and the LPU. In the following we introduce each of the options.

4.1 Data Fusion at the CPU

After the local processing at each LPU is performed as described in Section 3 and Algorithm 5, the local approximate marginals from all the sub-arrays are sent to the CPU. The CPU fuses the received information to get an overall estimate of each user’s symbol. The data fusion is done based on a sub-array data fusion binary matrix

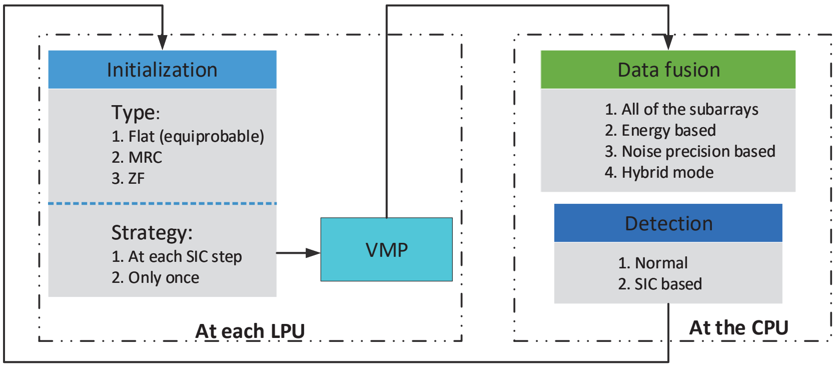


Fig. C.3: A summary of the algorithms for the VMP based symbol detection in the XL-MIMO system.

defined as

$$\mathbf{V} = \begin{pmatrix} v_{1,1} & v_{1,2} & \cdots & v_{1,K} \\ \vdots & \vdots & \ddots & \vdots \\ v_{B,1} & v_{B,2} & \cdots & v_{B,K} \end{pmatrix} \in \{0,1\}^{B\times K}. \tag{C.29}$$

If $v_{b,k} = 1$ then the local estimate x_k^b of sub-array b will contribute to the global estimate of x_k, otherwise if $v_{b,k} = 0$ it will not be fused. The data fusion is done by multiplying the estimates coming from each of the sub-arrays. Thus, the global estimates can be calculated as

$$q_{x_k}(x_k) \propto \prod_{b\in\mathcal{F}_k} q_{x_k^b}(x_k) \tag{C.30}$$

where $\mathcal{F}_k = \{b \in \{1,\ldots,B\} | v_{b,k} = 1\}$, i.e., for the kth user symbol the product is only taken over those sub-arrays b whose entry $v_{b,k} = 1$. Depending on the values of $\mathbf{V}$, different data fusion strategies can be selected. The simplest choice of the data fusion is when the CPU fuses data from all of the sub-arrays, i.e. $\mathbf{V} = \mathbf{1}_{B\times K}$ (a matrix of all ones).

Other selection options exploiting the non-stationary spatial structure of the received signals over the array are discussed next, with the goal of finding a data fusion matrix $\mathbf{V}$ which yields an advantageous complexity-performance trade-off.

Power based data fusion

One of the main differences of an XL-MIMO system with a conventional one is the variations of received users' energy along the array. For instance, measurements in [7] confirm that the energy from each user is not evenly distributed along all the elements of the array. Therefore, this property can be exploited to reduce the complexity of the receivers by processing only the parts of the array with the highest energy. Furthermore, results in [6] confirm that processing parts of the XL-MIMO array that contain a significant amount of energy is enough to get almost the same spectral efficiency

Algorithm 6 sub-array data fusion matrix **V** construction.

```
Result: V
Initialize: $\mathbf{H}$, $K$, $B$, $B_{\max}$, $M_b$, $p_0$, $\boldsymbol{\lambda}$, $\mathbf{V} = \{0\}^{B\times K}$, $\mathcal{B} = \{1,\ldots,B\}$ , $data\, fusion\, type$
 if $data\, fusion\, type$ ==PWR then
    for $k = 1$ to $K$ do
        1. reinitialize $p_k = 0$
        2. compute total cumulative power $P_k = ||\mathbf{h}_k||_2^2$
        3. compute per sub-array power $P_k^{(b)} = ||\tilde{\mathbf{H}}_{b,k}||_2^2, b \in \mathcal{B}$
        while $p_k \leq p_0 \cdot P_k$ do
            1. find $b^* = \max_{b\in\mathcal{B}} P_k^{(b)}$
            2. $p_k = p_k + P_k^{(b^*)}$
            3. set $\mathbf{V}(b^*, k) = 1$
            4. $\mathcal{B} = \mathcal{B} \setminus b^*$
        end
    end
else if $data\, fusion\, type$ ==NOP then
    1. sort $\boldsymbol{\lambda}$'s elements decreasingly ($\mathbf{b} = \arg\mathrm{sort}(\boldsymbol{\lambda})$)
    2. choose the first $B_{\max}$ elements of the sorted indices vector $\mathbf{b}$ as $\mathbf{b}^* = \mathbf{b}(1 : B_{\max})$
    3. set $\mathbf{V}(\mathbf{b}^*, :) = 1$
else if $data\, fusion\, type$ ==HYB then
    1. compute per sub-array power $P_k^{(b)}, b \in \mathcal{B}$ and $\mathbf{P} = \{p_{b,k} | p_{b,k} = P_k^{(b)}\ \forall k, b\}$
    2. compute the hybrid measure matrix from (C.31)
    for $k = 1$ to $K$ do
        3. sort elements of the $k$-th column of $\boldsymbol{\Gamma}$ decreasingly ($\mathbf{b_k} = \arg\mathrm{sort}(\boldsymbol{\Gamma}(:, k))$)
        4. choose the first $B_{\max}$ elements of $\mathbf{b_k}$ as $\mathbf{b_k}^* = \mathbf{b_k}(1 : B_{\max})$
        5. set $\mathbf{V}(\mathbf{b_k}^*, k) = 1$
    end
else
    All sub-arrays fusion : $\mathbf{V} = \mathbf{1}_{B\times K}$
end
```

(SE) as processing all the elements of the array. Here, we use the same principle by assigning the users to the sub-arrays that contain a certain ratio of the user's energy, for instance 80%, as we introduced in [4]. This method is noted by *PWR* mode in Algorithm 6.

Noise precision based data fusion

As mentioned in Section 3, the noise precision parameter λ_b is a measure of residual interference and AWGN variance per each sub-array. Therefore, more reliable sub-arrays for signal detection are those with the highest $\bar{\lambda}_b$ values. In order to limit the complexity, we restrict the algorithm to choose the top $B_{\max}$ sub-arrays. This method is included as *NOP* mode in Algorithm 6, where $\boldsymbol{\lambda} = \{\bar{\lambda}_b | \forall b \in \mathcal{B}\}$ and $\mathcal{B} = \{1, \cdots, B\}$ is the set of all sub-arrays.

Hybrid data fusion

Inspired by (C.21) and the variance of the Gaussian distribution, another possible metric is to choose the ones with the lowest variance in (C.21). In order to calculate this metric, we define a *hybrid measure matrix* $\boldsymbol{\Gamma} \in \mathbb{R}^{B \times K}$ as

$$\boldsymbol{\Gamma} = diag(\boldsymbol{\lambda})\mathbf{P} \tag{C.31}$$

where, $\mathbf{P} = \{p_{b,k} | p_{b,k} = P_k^{(b)} \; \forall k, b\}$ is the energy of each user in each of the sub-arrays. Then, based on the number of sub-arrays to be processed $B_{\max}$, we select the top $B_{\max}$ sub-arrays for each of the users. These sub-array subsets determine the data fusion candidates for each of the user symbols. This method is listed by *HYB* mode in Algorithm 6.

4.2 Symbol Detection Strategies

After the data fusion process, CPU can decide how to detect the users' symbols. In the following, we discuss two different approaches for the detection in the CPU.

Non-iterative Data Fusion and Detection

In this case, the CPU fuses all the local estimates and no further processing is done over the fused information. Then, it uses the global estimate for each of the users' transmitted symbol, $q_{x_k}(x_k)$, and detects $\hat{x}_k$ as the constellation point from the set $\mathcal{A}$ with largest approximate marginal , i.e. $\hat{x}_k = \arg\max_{a \in \mathcal{A}} q_{x_k}(a) \quad \forall k$.

SIC Data-Fusion and Detection

One of the effective ways to boost the receiver performance is to use SIC. This type of detector works sequentially and at each step detects the strongest user (or layer) and then removes its effect from the received signal. This operation reduces the interference successively and therefore improves the probability of successful detection of the subsequent symbols. One of the main factors that determine the performance

Algorithm 7 Data fusion and the SIC at the CPU.

```
Result: Central symbol estimates using the SIC detection
Initialize: $\mathbf{y}$, $K$, $\mathcal{A}$, detected symbols set $\mathcal{S} = \phi$.
1. Define user set $\mathcal{K} \triangleq \{1, \cdots, K\}$
for $i = 1$ to $K$ do
    2. Run Algorithm 5 to get the local estimates $q^b_{x_k}(x_k)$ for all the LPUs $b \in \{1, \cdots, B\}$.
    3. Fuse the local estimates to get $q_{x_k}(x_k)$ from (C.30) for all the users in $\mathcal{K}$.
    4. Choose the strongest user $k^*$ to detect with the LR measure using (C.32).
    5. Detect the transmitted symbol for $k^*$ as $\tilde{x}_{k^*}$ and include it to the detected symbol set $\mathcal{S} \leftarrow \mathcal{S} \cup \tilde{x}_{k^*}$.
    6. Cancel the interference caused by $k^*$ by: $\mathbf{y} \leftarrow \mathbf{y} - \tilde{x}_{k^*}\mathbf{h}_{k^*}$ and remove $\mathbf{h}_{k^*}$ from $\mathbf{H}$
    7. Fix the prior for $k^*$ as $q_{x_{k^*}}(x_{k^*}) = \delta(x_{k^*} - \tilde{x}_{k^*})$
    8. $\mathcal{K} \leftarrow \mathcal{K} \setminus k^*$
end
9. Output: $\mathcal{S}$.
```

of the SIC detector is the user (or layer) ordering method. Instead, we propose a new metric called likelihood ratio (LR) metric, based on the ratio of probabilities between the top two most likely symbols. In order to define the symbol certainty, first, we sort the symbol probabilities provided by the approximate marginals $q_{x_k}(x_k)$ of each user k as $p_1^{(k)} \geq p_2^{(k)} \geq \cdots p_{|\mathcal{A}|}^{(k)}$. Next, we define the symbol certainty as

$$\Delta_k \triangleq \frac{p_1^{(k)}}{p_2^{(k)}}, \qquad k = 1, \ldots, K. \tag{C.32}$$

Finally, the LR metric which chooses the strongest user as $k^* = \arg\max_k \Delta_k$. Algorithm 7[3] represents the SIC mechanism and cooperation between the CPU and the LPUs.

5 Performance Evaluation

In this section, we present numerical results for the performance of the proposed algorithms. We begin by describing the used simulation model and the selected benchmarks. We follow by analysing the computational complexity of the methods and end by illustrating the performance of the proposed receivers with respect to the benchmarks.

Generating the channel and simulation parameters

Intending to generate the channel model in (C.2), and based the on simulation parameters in Table C.1, first we generate the VRs. After assigning n_b random clusters

[3] $\delta(\cdot)$ is the Dirac delta function.

Table C.1: Simulations Parameters

Variable	Value	Variable	Value	Variable	Value
M	256	K	32	d_r	0.0578 m
d_S	5 m	$\mathcal{I}$	1	$\lvert\mathcal{A}\rvert$	4 (QPSK)
C	20	l_i	$\mathrm{LN}(0.7, 0.2)$	c_i	$\mathcal{U}(0, M)$
ψ_i	$\mathrm{N}(-0.21, 0.8)$	S_i	31	$\alpha_i, \tilde{\alpha}_{i,k}$	$\mathcal{U}(-\frac{\pi}{2}, \frac{\pi}{2})$
$\tilde{\theta}_{i,k}$	$3\pi/4$	θ_i	$7\pi/8$	B	4
$B_{\max}$	3	p_0	0.75	n_b	4

to each user k and forming $\mathcal{V}_k$, the correlation matrices are calculated using (C.6) and (C.7).

Note that, the fast fading parameters are generated at each channel realization while the correlation matrices are updated every 50 realizations to model the long term statistics of the channel. We have implemented channel coding using low density parity check (LDPC) codes. We use the code in the 5G New Radio standard [26, Section 5.2] with the following parameters: LDPC base graph 2 is used with block length $= 20$, $K_b = 6$ and $Z_c = 2$ resulting in an initial coding rate of $1/5$. Rate matching is used to increase the code rate to 0.5.

5.1 Benchmarks

Ideal matched filter bound

We choose matched filter bound which is the case when the effect of all the other users are ideally canceled and the target user's signal is detected by MRC. This single-user detection in the interference-free channel gives the best achievable BER [27].

Central linear processes

To compare the performance of our proposed receiver with centralized linear processing methods, MRC and ZF benchmarks are implemented. These are obtained by respectively applying (C.23) and (C.25).

Expectation propagation method from [20]

Aiming to have a benchmark method for a message passing based scheme, we implemented the EP algorithm presented in [20]. This method works in a distributed manner where the sub-arrays are exchanging their local estimates and a final decision is taken in the central node. It is worth mentioning that the complexity of this method is higher than our method due to the matrix inversions and singular value decompositions required. However, here we only consider the error rate results regardless of the complexity.

Table C.2: Complexity analyses for Alg. 5

Step	# of multiplications	Remarks
1	None	–
2	$3M_bK$	For the MRC initialization
3	$2K$	Two operations for each user
4	$2K + M_b$	Two summations + $l2$ norm for each user
6	$K\|\mathcal{A}\|$	$\|\mathcal{A}\|$ constellation points for each user
–	$\mathcal{I}(K(4 + \|\mathcal{A}\|) + M_b) + 3M_bK$	Total number of multiplications

Centralized VMP method from [19]

In order to have a VMP based receiver benchmark, we compare our proposed receivers' performance with the centralized VMP method that we proposed in [19].

5.2 Complexity Analyses

In this subsection we analyse the complexity of the aforementioned methods. The complexity for the central ZF and the central MRC are [28]

$$C_{\text{ZF}} = \frac{K^3}{3} + MK^2 + MK \qquad \text{(C.33)}$$

$$C_{\text{MRC}} = 3MK. \qquad \text{(C.34)}$$

In order to calculate the complexity of the VMP method in both of the algorithms, we start with Algorithm 5 and analyse each of the steps separately. Their complexity is reported in Table C.2. Next, the total complexity of Algorithm 7 is calculated using the complexity values obtained for VMP processing at each sub-array. To begin with, we discuss the following remarks regarding this algorithm:

- **Remark 1**: The number of VMP iterations $\mathcal{I}$ is one of the important parameters in the complexity-convergence performance of the VMP method. We tested different values for $\mathcal{I}$ and found that the VMP converges at $\mathcal{I} = 1$ and there is no need to repeat the operations. Thus, $\mathcal{I} = 1$ is assumed for all simulations and analyses henceforth.
- **Remark 2**: At each SIC iteration, the number of undetected users decreases by 1. Therefore, we have to consider a variable complexity for step 2 of Alg. 7 due to the size of $\mathcal{K}$. This can be done easily by factorizing K from the expressions in Table C.2 as $C_{\text{VMP}} \approx Kf(M_b, \mathcal{A})$ and a summation over different values of K. For instance, algorithm starts with K users, then in the second round with $K-1$ users and so on. The total complexity can be approximated as $\sum_{k=K}^{1} kf(M_b, \mathcal{A}) \approx K^2 f(M_b, \mathcal{A})/2$.
- **Remark 3**: For the VR based VMP methods the complexity depends on $B_{\max}$ and p_0 values. Thus, with a rough comparison, the complexity will scale with factor of $B_{\max}/B$ and p_0 for noise precision and power based data fusion methods, respectively. For instance, considering 75% power threshold or having $B_{\max} = 3$ in a system with $B = 4$ will approximately reduce the total complexity by 25%.

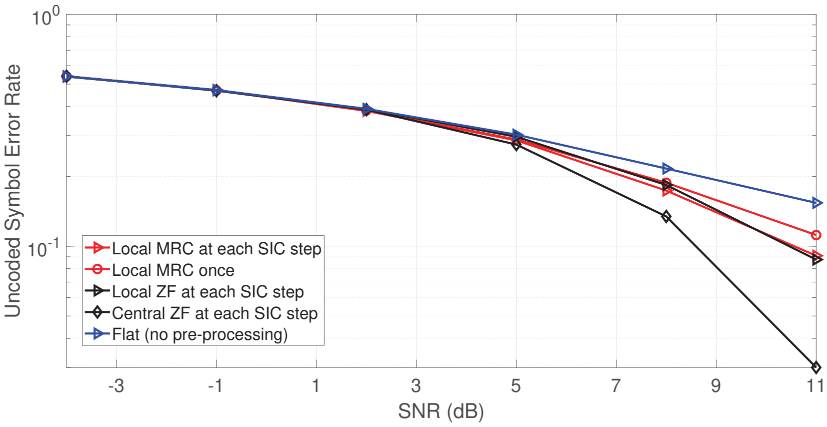


Fig. C.4: Different initialization techniques and their effect on the probability of error. In this simulations we used a lowly-correlated channel having the parameters in Table. C.1 with $M = 128$, $K = 16$, $B = 2$.

Thus, using last row of Table II, Remark 1 and 2, the complexity after doing SIC at each LPU is $C_{\text{LPU-SIC}} = \sum_{k=K}^{1} k(4 + 3M_b + |\mathcal{A}|) + M_b \approx K^2/2(4 + 3M_b + |\mathcal{A}|) + M_bK$. This operation is done in B subarrays, thus, $C_{\text{all-LPU-SIC}} \approx BK^2/2(4 + 3M_b + |\mathcal{A}|) + BM_bK$. Finally, the data fusion at the CPU that has k multiplications per each SIC step, resulting in approximately $K^2/2$ additional multiplications. Therefore, total complexity of the algorithm is $C_{\text{VMP-SIC}} \approx K^2/2(4B + 3BM_b + B|\mathcal{A}| + 1) + BM_bK$ and knowing $BM_b = M$ we can simplify it to

$$C_{\text{SIC-VMP}} \approx \frac{K^2}{2}(3M + B(|\mathcal{A}| + 4) + 1) + MK \tag{C.35}$$

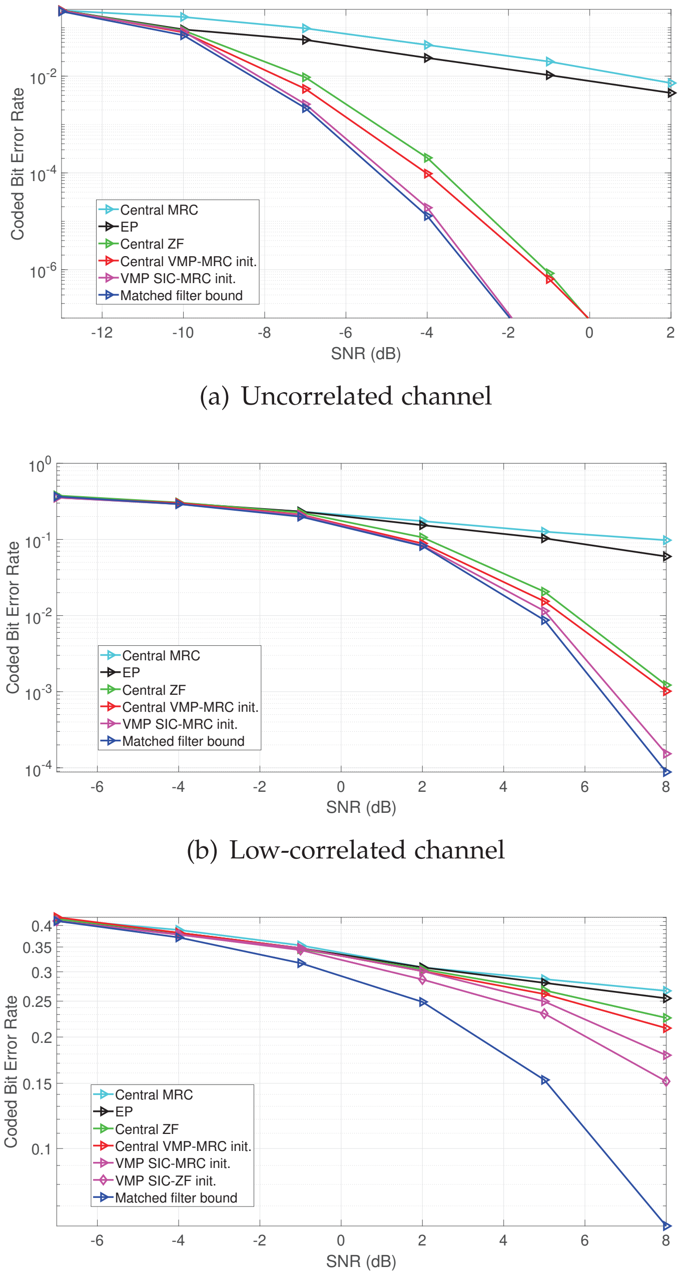
which is a second-order function of K. We will compare the numerical evaluations of the expressions we derived in this subsection later in Sec. 5.[4]

Fig. C.4 shows the different performances depending of the initialization types discussed in 3.4. As expected, the best performance is when we initialize at each step of the SIC operation. Due to the similar performance of the ZF and the MRC methods, it is more favorable to use the MRC mode for its less complexity. Thus, from now on we use the MRC at each step of the SIC as our default initialization method in all of the VMP performance evaluations.

5.3 Simulation Results

We start with comparing the coded bit error rate (BER) of the detection methods for different values of the pre-processing SNR=$\frac{P}{\sigma_n^2}$ where P is the expected transmitting power of the users. The data fusion matrix is $\mathbf{V} = \mathbf{1}_{B\times K}$ except for Fig. C.8 for the VR-aware methods. We consider three types of channels according to the correlation matrices introduced in (C.3) as

[4] The complexity burden of calculating VR-aware metrics is negligible. For instance, in the *PWR* mode, the value of user energies are calculated within the messages of the VMP method. Therefore, the CPU only needs to order them and choose the first B_{max} of them with the complexity of $\mathcal{O}(B\log B)$. The same complexity order holds for the other modes as well.



(a) Uncorrelated channel

(b) Low-correlated channel

(c) High-correlated channel

Fig. C.5: Coded BER of different methods vs pre-processing SNR for different correlation scenarios. Simulation parameters are detailed in Table I.

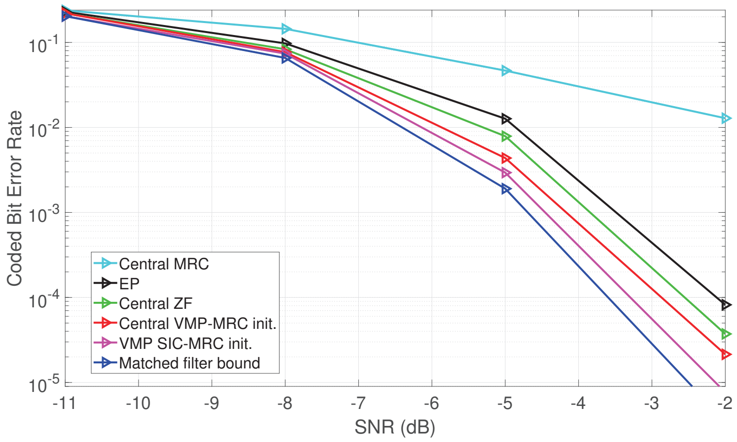


Fig. C.6: Coded BER of different methods vs pre-processing SNR. $M = 100$, $K = 8$ and $B = 4$.

1. Uncorrelated channel with $\mathbf{H} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$.
2. Lowly correlated channel with $\theta_i = 7\pi/8$, yielding $\boldsymbol{R}_i$ very close to the identity matrix.
3. Highly correlated channel with $\theta_i = 3\pi/4$.

In Fig. C.5 we compare the performance of the aforementioned methods for the three types of channels. First, in an uncorrelated Rayleigh fading channel, our VMP based method works very close to the lower bound and clearly outperforms the benchmarks. For the case of the lowly correlated channel, our method still performs very close to the matched filter bound. However, for the highly correlated channel, the performance gain over linear methods is small. One way to boost the performance is to add a central ZF initialization to our VMP receiver, which significantly increases its complexity. As expected, this ZF initialization is showing better performance than the local MRC initialization one. The results show the superior performance of the proposed distributed VMP method where it performs very close to the ideal bound. Having a SIC mechanism in combination with VMP processing provides the best performance in high SNR regimes where the effect of error propagation becomes negligible. Another observation from Fig. C.5 is the large performance degradation caused by correlated and non-stationary channels compared to an uncorrelated one. This is due to the channel capacity reduction that has been shown in [21] for the correlated channels and in [9] for the non-stationary channels. The reason for the poor performance of the EP based method is because of the relatively high ratio of the users and antennas $\frac{M}{K} \lessapprox 10$. Furthermore, the complex correlated channel model of (C.2) impairs the receiver and makes it perform only slightly better than the centralized MRC. The EP based method works significantly better in the unrealistic i.i.d. channel model with a smaller number of users [20], as shown in Fig. C.6. Here, we compare the methods in an i.i.d channel and a lower system load $\frac{M}{K} = 12.5$. As can be seen, the EP based method performs much closer to the central ZF curve. For the rest of the simulations we use the lowly correlated channel model.

In Fig. C.7, we compare the performance of the receivers with respect to the number of users K. The total array size is kept fixed at $M = 256$. We can observe that

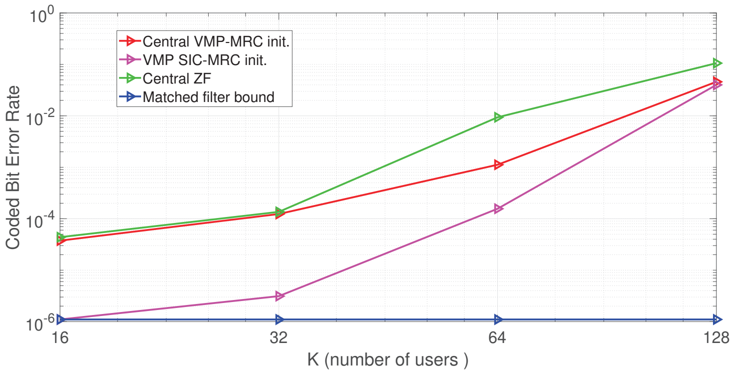


Fig. C.7: Effect of number of the users in performance of the detection methods. The VMP based method degrades slower than the benchmarks with increasing K. (SNR= 10 dB, $M = 256$ and $B = 4$)

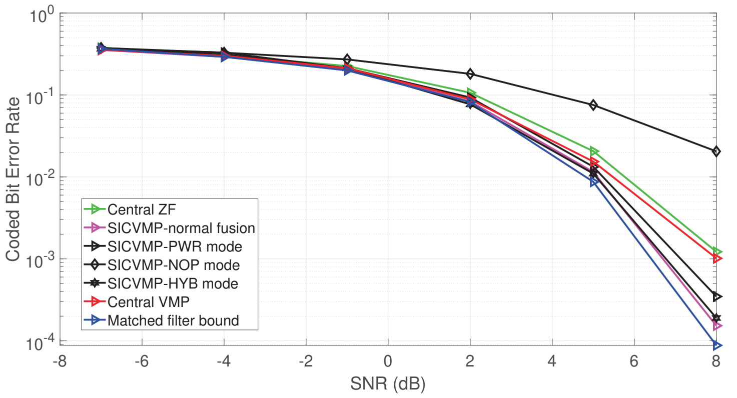


Fig. C.8: The VR based methods and their performance compared to the benchmarks and also the VMP with full sub-array data fusion. $B_{\text{max}} = 3$ and $p_0 = 0.75$. Other parameters are from Table I.

as we add more user in the system and go towards crowded scenarios, the VMP method, which is using the MRC initialization, degrades at a slower rate than the ZF receiver. The reason is the fact that the linear receivers fail to operate properly when the number of users becomes comparable with the number of the antennas at the BS ($\frac{M}{K} < 10$). With a larger number of users present, the probability that a user's channel is approximately orthogonal to that of all other users decreases, with large degradation of the "favorable propagation" conditions usually present in massive MIMO channels. Although the performance also degrades considerably for the VMP receiver, the degradation is less than for the ZF.

The BER of the different VR based modes which restrict data fusion to 75% of the sub-arrays according to section 4.1 is illustrated in Fig. C.8. The VMP method with the fusion of data from all sub-arrays, i.e. with $\mathbf{V} = \mathbf{1}_{B\times K}$, provides the best performance as expected. The power-based data fusion technique (PWR) provides better performance than the ZF receiver, while the noise precision based (NOP) receiver performs poorly. The receiver with hybrid (HYB) is the best among VR-based methods, as it approaches the performance of full data fusion while only requiring approximately 75%

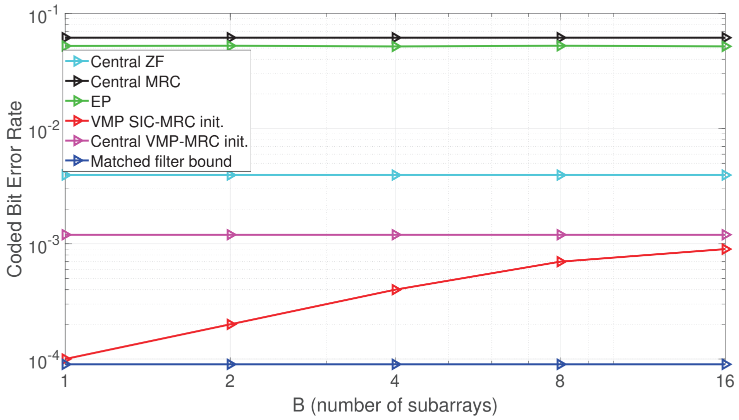


Fig. C.9: Effect of number of the sub-arrays B on the BER behaviour of the distributed schemes. The size of the array M is fixed to 128. (SNR= 5 dB, $K = 8$)

of its complexity.[5] (The complexity reduction is scaled with $\frac{B_{\text{max}}}{B}$[6]) These results illustrate that exploiting non-stationary properties helps to obtain a receiver with lower computational complexity and almost the same performance of normal data fusion.

In Fig. C.9 we analyze the performance of the VMP receiver as the processing is distributed among different number sub-arrays. With an array of fixed size $M = 128$, the performance at an SNR of 5 dBs is analyzed for different number of subarrays B. The central ZF and MRC receivers and the matched filter bound are unaffected by B. Predictably, distributing the processing among more local units leads to an increase in the BER, although the performance is still better than that of centralized receivers. Note that in the rightmost point of this figure, with $B = 16$, the number of antennas per sub-array is $M_b = M/B = K = 8$ which is an extreme case for a massive MIMO system, but the VMP receiver still operates with acceptable performance.

We finalize by showcasing the complexity of the assessed receivers in Fig. C.10. We consider two system load regimes: one with $M/K = 10$ for a moderate load, and $M/K = 5$ in a crowded scenario. As it can be observed, the complexity of the VMP method when it fuses all the sub-arrays is higher than the ZF method. As we discussed before, this is a trade-off point where we get a near to optimal performance while spending more computational resources. Moreover, we can see that the complexity of a VR based method with hybrid fusion mode is close to the ZF while it still provides better BER output. These results illustrate that the proposed VMP receiver can be tuned to trade-off performance and computational complexity depending on the system requirements and operating conditions.

[5] The term "approximately" is subarray ordering the CPU and only activate the first B_{max} of the LPUs.

[6] For example, for the case of *power-based* mode, after the channel state information (CSI) obtaining phase, the CPU can choose maximum B_{max} LPUs for each of the users to be in charge of that user, and the rest of LPUs will not run the VMP algorithm for that user.

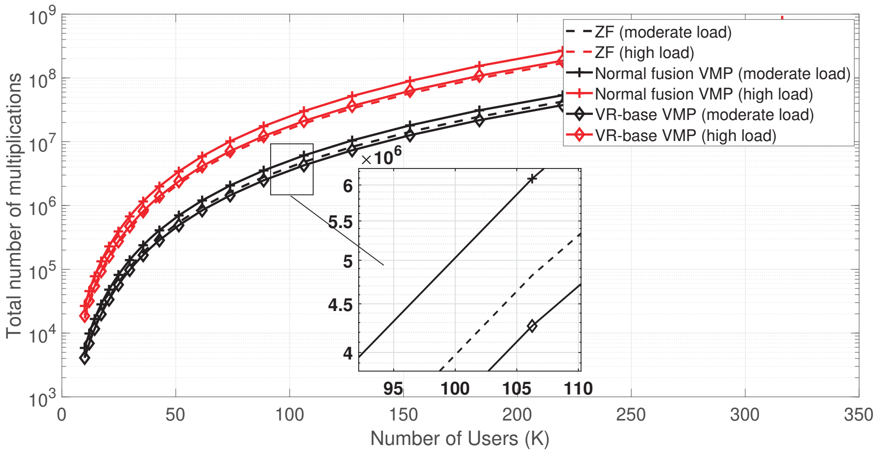


Fig. C.10: Number of complex multiplications for the VMP based methods and the central ZF for different load conditions. For the VR-based VMP we assume to have $\frac{B_{\max}}{B} = 0.75$, $M = 256$ and $B = 4$.

6 Conclusions

We propose a distributed receiver structure based on VMP that outperforms conventional massive MIMO receivers, especially when operating in spatial non-stationary channels. Numerical results show that the receivers implementing the proposed algorithm perform close to a genie-aided receiver (matched filter bound), even in highly correlated channel conditions. One of the key components of our method is the internal SIC mechanism which takes advantage of the energy variations over the VRs of different users. This interference cancellation improves the receiver performance for the users with overlapping VRs. Unlike the conventional linear receivers, our VR-based methods use information about the non-stationarities to limit the complexity without performance degradation. Our design is versatile in several respects: the distributed manner of all the processing tasks makes it easier for practical deployments of the XL-MIMO systems. The variety of options for initialization, data fusion and detection methods gives us several control parameters allowing for trading off computational complexity for performance and vice-versa in the receiver design for different applications. Our future research will address lower complexity receivers for higher modulations schemes, extending the detectors using machine learning techniques to optimize the data fusion and the SIC at the CPU.

References

[1] E. Björnson, L. Sanguinetti, H. Wymeersch, J. Hoydis, and T. L. Marzetta, "Massive MIMO is a reality—What is next?: 5 promising research directions for antenna arrays," *Digital Signal Processing: A Review Journal*, vol. 94, pp. 3–20, 2019.

[2] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Energy and spectral efficiency of very large multiuser mimo systems," *IEEE Transactions on Communications*,

vol. 61, no. 4, pp. 1436–1449, April 2013.

[3] P. Singh, H. B. Mishra, A. K. Jagannatham, K. Vasudevan, and L. Hanzo, “Uplink sum-rate and power scaling laws for multi-user massive mimo-fbmc systems,” *IEEE Transactions on Communications*, vol. 68, no. 1, pp. 161–176, 2019.

[4] A. Amiri, M. Angjelichinoski, E. de Carvalho, and R. W. Heath, “Extremely large aperture massive mimo: Low complexity receiver architectures,” in *2018 IEEE Globecom Workshops (GC Wkshps)*, Dec 2018, pp. 1–6.

[5] S. Hu, F. Rusek, and O. Edfors, “Beyond massive mimo: The potential of data transmission with large intelligent surfaces,” *IEEE Transactions on Signal Processing*, vol. 66, no. 10, pp. 2746–2758, 2018.

[6] E. De Carvalho, A. Ali, A. Amiri, M. Angjelichinoski, and R. W. Heath Jr, “Non-stationarities in extra-large scale massive mimo,” *arXiv preprint arXiv:1903.03085*, 2019.

[7] X. Gao, F. Tufvesson, and O. Edfors, “Massive mimo channels—measurements and models,” in *2013 Asilomar conference on signals, systems and computers*. IEEE, 2013, pp. 280–284.

[8] A. Ali, E. de Carvalho, and R. W. Heath, “Linear receivers in non-stationary massive mimo channels with visibility regions,” *IEEE Wireless Communications Letters*, 2019.

[9] X. Li, S. Zhou, E. Björnson, and J. Wang, “Capacity analysis for spatially non-wide sense stationary uplink massive mimo systems,” *IEEE Transactions on wireless communications*, vol. 14, no. 12, pp. 7044–7056, 2015.

[10] A. Amiri, C. Navarro Manchón, and E. de Carvalho, “Deep learning based spatial user mapping on extra large mimo arrays,” *arXiv preprint arXiv:2002.00474*, 2020.

[11] J. R. Sanchez, F. Rusek, O. Edfors, M. Sarajlic, and L. Liu, “Decentralized massive mimo processing exploring daisy-chain architecture and recursive algorithms,” *arXiv preprint arXiv:1905.03160*, 2019.

[12] M. Sarajlić, F. Rusek, J. R. Sánchez, L. Liu, and O. Edfors, “Fully decentralized approximate zero-forcing precoding for massive mimo systems,” *IEEE Wireless Communications Letters*, vol. 8, no. 3, pp. 773–776, 2019.

[13] P. Som, T. Datta, N. Srinidhi, A. Chockalingam, and B. S. Rajan, “Low-complexity detection in large-dimension mimo-isi channels using graphical models,” *IEEE journal of selected topics in signal processing*, vol. 5, no. 8, pp. 1497–1511, 2011.

[14] Z. Zhang, X. Cai, C. Li, C. Zhong, and H. Dai, “One-bit quantized massive mimo detection based on variational approximate message passing,” *IEEE Transactions on Signal Processing*, vol. 66, no. 9, pp. 2358–2373, 2017.

[15] G. Interdonato, E. Björnson, H. Q. Ngo, P. Frenger, and E. G. Larsson, “Ubiquitous cell-free massive mimo communications,” *EURASIP Journal on Wireless Communications and Networking*, vol. 2019, no. 1, p. 197, 2019.

[16] E. Björnson and L. Sanguinetti, “Making cell-free massive mimo competitive with mmse processing and centralized implementation,” *IEEE Transactions on Wireless Communications*, 2019.

[17] T. T. Vu, D. T. Ngo, N. H. Tran, H. Q. Ngo, M. N. Dao, and R. H. Middleton, “Cell-free massive mimo for wireless federated learning,” *arXiv preprint arXiv:1909.12567*, 2019.

[18] V. C. Rodrigues, A. Amiri, T. Abrao, E. de Carvalho, and P. Popovski, “Low-complexity distributed xl-mimo for multiuser detection,” *arXiv preprint arXiv:2001.11879*, 2020.

[19] A. Amiri, C. Navarro Manchón, and E. de Carvalho, “A message passing based receiver for extra-large scale mimo,” *arXiv preprint arXiv:1912.04131*, 2019.

[20] H. Wang, A. Kosasih, C.-K. Wen, S. Jin, and W. Hardjawana, “Expectation propagation detector for extra-large scale massive mimo,” *arXiv preprint arXiv:1906.01921*, 2019.

[21] D. Gesbert, H. Bolcskei, D. A. Gore, and A. J. Paulraj, “Outdoor mimo wireless channels: Models and performance prediction,” *IEEE Transactions on Communications*, vol. 50, no. 12, pp. 1926–1934, 2002.

[22] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.

[23] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, “An introduction to variational methods for graphical models,” *Machine learning*, vol. 37, no. 2, pp. 183–233, 1999.

[24] C. Navarro Manchón, "Advanced signal processing for mimo-ofdm receivers,” Ph.D. dissertation, Aalborg University, 2011.

[25] T. Brown, P. Kyritsi, and E. De Carvalho, *Practical guide to MIMO radio channel: With MATLAB examples*. John Wiley & Sons, 2012.

[26] “3GPP TS 38.212. NR; Multiplexing and Channel Coding.” Technical Specification Group Radio Access Network, Standard, Jul. 2018.

[27] W. Burchill and C. Leung, “Matched filter bound for ofdm on rayleigh fading channels,” *Electronics Letters*, vol. 31, no. 20, pp. 1716–1717, 1995.

[28] E. Björnson, L. Sanguinetti, J. Hoydis, and M. Debbah, “Optimal design of energy-efficient multi-user mimo systems: Is massive mimo the answer?” *IEEE Transactions on Wireless Communications*, vol. 14, no. 6, pp. 3059–3075, 2015.

Paper D

Uncoordinated and Decentralized Processing in Extra-Large MIMO Arrays

Abolfazl Amiri, Carles Navarro Manchón, Elisabeth de Carvalho

The paper is accepted for publication in the
Wireless Communications Letters (WCL)

Abstract

We propose a decentralized receiver for extra-large multiple-input multiple-output (XL-MIMO) arrays. Our method operates with no central processing unit (CPU) and all the signal detection tasks are done in distributed nodes. We exploit a combined message-passing framework to design an uncoordinated detection scheme that overcomes three major challenges in the XL-MIMO systems: computational complexity, scalability and non-stationarities in user energy distribution. Our numerical evaluations show a significant performance improvement compared to benchmark distributed methods while operating very close to the centralized receivers.

Keywords— Massive MIMO, Message-passing, decentralized receivers, XL-MIMO

1 Introduction

Beyond fifth-generation (B5G) multi-input multiple-output (MIMO) systems will rely on antenna arrays with an extreme number of elements that provide a very high spatial resolution. Recently, different variants of such systems such as extra-large scale MIMO (XL-MIMO) systems [1], or large intelligent surfaces (LIS) have been introduced. An XL-MIMO array has hundreds of elements where its physical size is in the range of tens of meters in sub-6 GHz frequency bands. These technologies offer a boost in the system's spectral efficiency thanks to their ability to jointly serve a large number of users. However, the large array dimensions bring about three important challenges: an increase in the computational complexity of the receiver processing, difficulties in the scalability of the array architecture, and the emergence of spatial non-stationarities (NS) of the received signal across the array.

Most of the conventional receiver designs, e.g. zero-forcing (ZF), rely on a central processing unit (CPU) that inverts large matrices [11]. Such receivers need a vast amount of computational capacity to operate and are not suitable for our XL-MIMO system. Distributed receiver techniques such as [5] try to divide the computations between several nodes, but still, need a CPU to supervise all the data transmission steps.

Having a central node for processing all the transmitting signals requires a dedicated link between all the antenna elements and the CPU. Managing these interconnections is costly and hinders adding more antennas next to the existing array deployment. Therefore, scaling up the array size becomes challenging. Different hierarchical processing methods are used in [6–8] that aim to divide the processing tasks between the CPU and local units at the *sub-arrays*. However, a connection between each of these sub-arrays and the CPU is still necessary and limits the scalability of the receivers.

The last challenge is the presence of spatial NS of the channel gains imposing variable mean energy of a given user's signal along the array, thereby creating visibility regions (VR) [1]. A VR is a subset of the antennas that hold most of a user's received energy and limits the performance of linear receivers [5]. Therefore, there is a need for more efficient techniques that work regardless of the properties of the wireless channel and deliver an acceptable performance.

Most of the literature focuses on either lowering the computational complexity or dealing with the effect of NS. Authors in [9–11] propose fully decentralized ZF approximators that work without a CPU. However, these methods have a high processing delay and their performance is highly dependent on the even distribution of users' energy on the array. [8] and [12] use expectation propagation to distribute the symbol detection between a central node and local units. A Gaussian message passing technique is used for an overloaded MIMO system in [4]. In [6], we presented an NS-aware receiver that works in a hierarchical way. Yet all of these methods rely on a central node with a high processing delay.

In this paper, we propose a decentralized receiver that works without a CPU for user symbol detection. Our main focus is to decentralize the processing, and have a scalable receiver architecture that can be augmented by adding more subarrays without increasing the computational complexity and data-exchange load of the individual local processing units (LPUs). Our method leverages an approximate inference framework based on a combination of belief propagation (BP) and variational message-passing (VMP). We use distributed nodes called LPU that work in parallel to calculate the local symbol estimates. These nodes are only allowed to exchange information with their neighbors, making the receiver scalable and easy to deploy. Moreover, LPUs use a local successive interference cancellation (SIC) scheme to boost the symbol detection performance. Our numerical results show a significant improvement compared to other decentralized techniques while obtaining almost the same performance as the centralized benchmark methods.

Notations: Boldface lowercase and uppercase letters represents vectors and matrices, respectively. Set cardinality is denoted by $|\mathcal{I}|$; the relative complement of i in a set $\mathcal{I}$ is denoted as $\mathcal{I} \setminus i$. Superscripts $(\cdot)^T$ and $(\cdot)^H$ show transposition and Hermitian, respectively. The probability density function (pdf) of a multivariate complex Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ and its distribution are denoted by $\mathrm{CN}(\cdot;\boldsymbol{\mu},\boldsymbol{\Sigma})$ and $\mathcal{CN}(\boldsymbol{\mu},\boldsymbol{\Sigma})$, respectively. $\mathcal{U}(a,b)$ shows a uniform distribution in $[a,b]$. $\mathbf{I}_M$ denotes the identity matrix of size M and $\mathbf{1}_M$ is a vector with M 1s. We use $f(x) \propto g(x)$ when $f(x) = cg(x)$ for some positive constant c.

2 System Model

We assume a fully-digital narrow-band MIMO system with K single-antenna users and a base station (BS) with M antenna elements placed in a uniform linear array (ULA)[1] in the uplink transmission. User symbols are denoted with the vector $\mathbf{x} \in \mathbb{C}^K$ with entries taking values from the complex constellation set $\mathcal{M} = \{a_1, a_2, \cdots, a_{|\mathcal{M}|}\}$. $\mathbf{H} = [\boldsymbol{h}_1, \cdots, \boldsymbol{h}_K] \in \mathbb{C}^{M \times K}$ is the channel matrix with column vectors $\boldsymbol{h}_k$ each of which denotes user k's channel. We assume perfect channel knowledge is available at the BS. The noise at the BS is assumed to have circularly symmetric complex Gaussian distribution $\mathbf{n} \sim \mathcal{CN}(0, \sigma_n^2 \mathbf{I}_M) \in \mathbb{C}^M$ and σ_n^2 is noise variance. The received baseband signal $\mathbf{y} \in \mathbb{C}^M$ across the whole array is

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}. \tag{D.1}$$

[1] The extension to planar array topologies follows similar principles described in the paper.

The BS is made of a set of B sub-arrays located next to each other in one dimension, each controlled by an LPU and having $M_b = \frac{M}{B}$ antennas . We denote $\tilde{\mathbf{H}}_b \in \mathbb{C}^{M_b \times K}$ and $\mathbf{y}_b \in \mathbb{C}^{M_b}$ as the channel matrix and the received signal in the bth sub-array for $b \in \{1, \cdots, B\}$, respectively. An optimal receiver aims to find the symbols that maximize the posterior pdfs $p(\mathbf{x}|\mathbf{y})$. However, this problem has a combinatorial behaviour with $|\mathcal{M}|^K$ options for each instance of $\mathbf{x}$ and is unfeasible. Instead, we resort to inferring approximate marginals using our message passing algorithm that requires only $|\mathcal{M}| \times K$ evaluations instead.

3 Proposed Receiver Algorithms

In this section, we describe the unified message-passing algorithm that combines the BP and mean-field approximation (MF) approaches [16] and its application to derive our proposed decentralized receiver algorithm.

3.1 Preliminaries

Let $p(\boldsymbol{z})$ denote the pdf of a random vector $\boldsymbol{z} \triangleq (z_i | i \in \mathcal{I})^T$ with the set $\mathcal{I}$ indexing all the random variables in its entries. The combined BP-MF inference method is used to calculate approximate marginals $q_i(z_i)$ that are commonly called *beliefs*. To apply the BP-MF framework, first a factorization of $p(\boldsymbol{z})$ of the form

$$p(\boldsymbol{z}) = \prod_{a \in \mathcal{A}_{\text{MF}}} f_a(\boldsymbol{z}_a) \prod_{c \in \mathcal{A}_{\text{BP}}} f_c(\boldsymbol{z}_c), \tag{D.2}$$

should be selected, in which the different factors are classified as either belonging to the BP part (factors f_a, $a \in \mathcal{A}_{\text{BP}}$ and messages m^{BP}) or the MF part (factors f_c, $c \in \mathcal{A}_{\text{MF}}$ and messages m^{MF}). The vectors $\boldsymbol{z}_a$ denote vectors containing all the random variables that are argument of a given factor f_a. In addition, the sets $\mathcal{N}(a) \subseteq \mathcal{I}$ and $\mathcal{N}(i) \subseteq \mathcal{A}$ are defined to be the set of indices of all variables that are arguments of factors f_a and all factors that have variable z_i as an argument, respectively. With the above factorization, an algorithm is formulated that exchanges information, called *messages*, between factors in (D.2) and the variables $z_i, i \in \mathcal{I}$ iteratively. The messages are computed according to

$$m^{\text{BP}}_{a \to i}(z_i) = \int \prod_{j \in \mathcal{N}(i) \setminus i} \mathrm{d}z_j n_{j \to a}(z_j) f_a(\boldsymbol{z}_a), \forall a \in \mathcal{A}_{\text{BP}}, \tag{D.3a}$$

$$m^{\text{MF}}_{a \to i}(z_i) = \exp\left(\int \prod_{j \in \mathcal{N}(i) \setminus i} \mathrm{d}z_j n_{j \to a}(z_j) \ln f_a(\boldsymbol{z}_a) \right), \forall a \in \mathcal{A}_{\text{MF}}, \tag{D.3b}$$

$$n_{i \to a}(z_i) \propto \prod_{c \in \mathcal{A}_{\text{BP}} \cap \mathcal{N}(i) \setminus a} m^{\text{BP}}_{c \to i}(z_i) \prod_{c \in \mathcal{A}_{\text{MF}} \cap \mathcal{N}(i)} m^{\text{MF}}_{c \to i}(z_i), \forall a \in \mathcal{A} \tag{D.3c}$$

for all $i \in \mathcal{N}(a)$, and the messages in (D.3c) are normalized so that they integrate to unity, resembling a valid pdf. From inspecting (D.3a), we see that messages to a factor in the BP part are computed using the sum-product (SP) algorithm, while the messages to an MF factor in (D.3b) are computed using the VMP rule [16]. At any point during the algorithms, each of the variables' beliefs in the system can be recovered as

$$q_i(z_i) \propto \prod_{c \in \mathcal{A}_{\mathrm{BP}} \cap \mathcal{N}(i)} m^{\mathrm{BP}}_{c \to i}(z_i) \prod_{c \in \mathcal{A}_{\mathrm{MF}} \cap \mathcal{N}(i)} m^{\mathrm{MF}}_{c \to i}(z_i), \tag{D.4}$$

where the beliefs are normalized to behave as proper pdfs.

3.2 Probabilistic System Description

We seek to detect the transmitted user symbols x_k, $k = 1, \ldots, K$ and estimate the noise precision $\lambda = \frac{1}{\sigma_n^2}$ as a nuisance variable. Our focus is to perform such processing locally at each of the BS sub-arrays. The posterior probability density of these two variables is factorized as

$$\begin{aligned} p(\mathbf{x}^1, \cdots, \mathbf{x}^B, \boldsymbol{\lambda} | \mathbf{y}_1, \cdots, \mathbf{y}_B) &\propto p(\mathbf{y}_1 | \mathbf{x}^1, \lambda_1) p(\lambda_1) \prod_{k=1}^{K} [\underbrace{p(x_k^1)}_{f_{x_k}}] \\ &\times \prod_{b=2}^{B} [\, \underbrace{p(\mathbf{y}_b | \mathbf{x}^b, \lambda_b)}_{f_{\mathbf{y}_b}} \underbrace{p(\lambda_b)}_{f_{\lambda_b}} \prod_{k=1}^{K} \underbrace{p(x_k^b | x_k^{b-1})}_{f_{\mathrm{E}_k^b}} \,] \end{aligned} \tag{D.5}$$

where the variables λ_b and $\mathbf{x}^b = [x_1^b, \ldots, x_K^b]^T$ represent noise precision and the transmitted symbols observed by the bth BS sub-array, respectively. Also, $f_{\mathbf{y}_b}(\mathbf{x}^b, \lambda_b) = \mathrm{CN}(\mathbf{y}_b; \tilde{\mathbf{H}}_b \mathbf{x}^b, \frac{1}{\lambda_b} \mathbf{I}_{M_b})$ and $f_{\lambda_b} \propto 1/\lambda_b$ [2]. $f_{x_k} = p(x_k) = \frac{1}{|\mathcal{M}|} \mathbf{1}_{|\mathcal{M}|}(x_k \in \mathcal{M})$ and $p(x_k^b | x_k^{b \pm 1})$ are equality constraints that we will describe in 3.4. Note that (D.5) is a direct application of (D.2) where $\prod_{k=1}^{K} p(x_k^b | x_k^{b-1})$ shows the factors for the BP part and the rest of the terms are the MF factors. Clearly, $\mathbf{x}^b$ and λ_b model the same random variables for the different sub-arrays $b = 1, \ldots, B$. However, we treat them separately such that each sub-array can get independent estimates of them. We impose equality constraints on the symbols x_k at different sub-arrays since it is important that different sub-arrays converge to common detected symbols by exchanging messages on their respective local estimates[3]. The graphical representation of the model (D.5) is depicted in Fig. D.1, where Tanner-style factor graphs [16] are used.

3.3 MF at the Local Processing Units

At each of the LPUs, the processing is done using the VMP algorithm which is a message-passing interpretation of MF inference. Thus, the bth LPU aims at approximating the posterior distribution of $\mathbf{x}^b$ and λ_b by using the approximate distribution $q_b(\mathbf{x}^b, \lambda_b) = q_{\lambda_b}(\lambda_b) \prod_{k=1}^{K} q_{x_k^b}(x_k^b)$, where the naïve MF approximation is applied. To

[2] This choice of prior corresponds to an improper, non-informative Gamma prior distribution with shape and rate parameters approaching zero.

[3] We do not use the equality constraint for the noise precision since it is a nuisance variable and, consequently, it is not critical for the algorithm if different sub-arrays yield slightly different estimates for it.

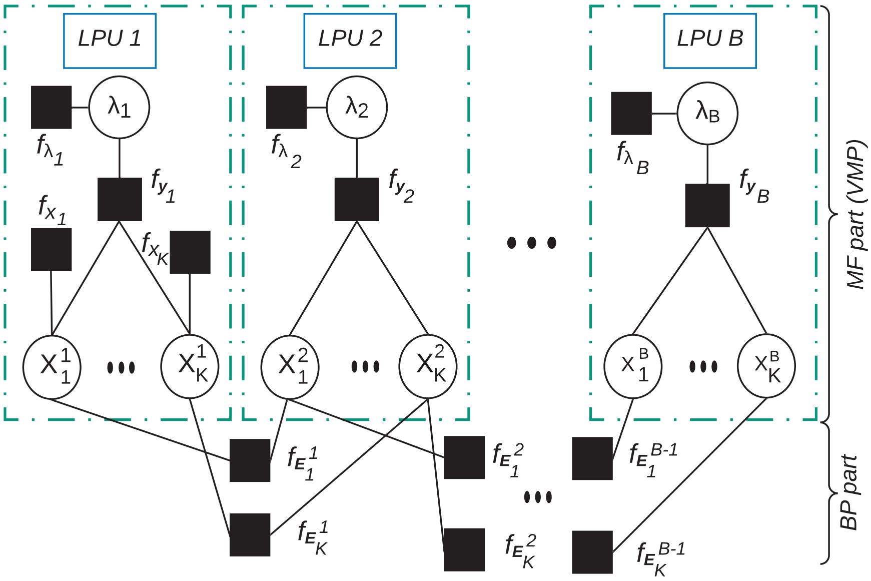


Fig. D.1: Factor graph representation of the local processing units. Local estimations are being exchanged between the sub-arrays.

begin with, using (D.3b), the message from factor node $f_{\mathbf{y}_b}$ to the variable node λ_b is calculated as [6]

$$m^{\mathrm{MF}}_{f_{\mathbf{y}_b}\longrightarrow\lambda_b}(\lambda_b) \propto {\lambda_b}^{M_b}\exp(-\lambda_b Z_b) \tag{D.6}$$

where $Z_b = ||\mathbf{y}_b - \sum_k \tilde{\mathbf{h}}_{b,k}\bar{x}^b_k||^2 + \sum_k \sigma^2_{x^b_k}\tilde{\mathbf{h}}^H_{b,k}\tilde{\mathbf{h}}_{b,k}$ and $\tilde{\mathbf{h}}_{b,k}$ denotes the kth column of $\tilde{\mathbf{H}}_b$ while $\bar{x}^b_k = \sum_{s\in\mathcal{M}} s q_{x^b_k}(s)$ and $\sigma^2_{x^b_k} = \sum_{s\in\mathcal{M}} |s|^2 q_{x^b_k}(s) - |\bar{x}^b_k|^2$ are the mean and variance of x^b_k with respect to $q_{x^b_k}(x^b_k)$. Next, the LPU calculates the approximate marginal distribution $q_{\lambda_b}(\lambda_b)$ by multiplying the messages entering the variable node λ_b as

$$q_{\lambda_b}(\lambda_b) = f_{\lambda_b}(\lambda_b) \times m^{\mathrm{MF}}_{f_{\mathbf{y}_b}\rightarrow\lambda_b}(\lambda_b) = {\lambda_b}^{M_b-1}e^{-\lambda_b Z_b} \tag{D.7}$$

which corresponds to a Gamma distribution with mean $\bar{\lambda}_b = \frac{M_b}{Z_b}$. Afterwards, the LPU computes the messages from factor node $f_{\mathbf{y}_b}$ to the variable nodes $x^b_{k'}$, yielding [6]

$$m^{\mathrm{MF}}_{f_{\mathbf{y}_b}\rightarrow x^b_k} \propto \mathcal{CN}\Bigg(x^b_k; \frac{\tilde{\mathbf{h}}^H_{b,k}}{||\tilde{\mathbf{h}}_{b,k}||^2}(\mathbf{y}_b - \sum_{k'\neq k}\bar{x}^b_{k'}\tilde{\mathbf{h}}_{b,k'}), \frac{1}{\bar{\lambda}_b||\tilde{\mathbf{h}}_{b,k}||^2}\Bigg). \tag{D.8}$$

To conclude the MF part, the approximate marginals of the symbols of each user at the LPU b are obtained by multiplying the above messages with their local priors and incoming messages from the neighboring LPUs, resulting in

$$q_{x^b_k}(x^b_k) \propto m^{\mathrm{MF}}_{f_{\mathbf{y}_b}\rightarrow x^b_k}(x^b_k)\, m^{\mathrm{BP}}_{f_{\mathrm{E}^b_k}\rightarrow x^b_k}(x^b_k)\, m^{\mathrm{BP}}_{f_{\mathrm{E}^{b-1}_k}\rightarrow x^b_k}(x^b_k), \tag{D.9}$$

where, $m^{\text{BP}}_{f_{\text{E}_k^b} \to x_k^b}(x_k^b)$ is the message from the LPU $b+1$ to LPU b. Also, for $b = 1$ the message coming from LPU $b-1$ is replaced by $f_{x_k^b}(x_k^b)$, while for $b = B$ there is no message from LPU $b+1$. Basically, the equation shows that, at each LPU, the local estimate of the symbol distribution is obtained from a combination of the LPU's observed signal and the information received from the adjacent LPUs. Since adjacent LPUs estimates include, in turn, information from their respective adjacent LPUs, this mechanism ensures that after sufficient iterations of the algorithm the different LPUs converge to consistent estimates. We describe the details for these messages in the next subsection.

3.4 BP between the sub-arrays

In this subsection, we discuss the exchange of the BP messages taking place between adjacent LPUs. The fact that each LPU exchanges messages only with their neighbors, and independently of how many sub-arrays are there in total, makes our proposed solution scalable. To begin with, we define the equality factor nodes as

$$f_{\text{E}_k^b}(x_k^b, x_k^{b+1}) = \delta(x_k^b - x_k^{b+1}), \tag{D.10}$$

where $\delta(\cdot)$ is Kronecker Delta function ensuring consistency of the local estimates at LPUs b and $b+1$. Thus, we can use this function to calculate the incoming messages to the bth LPU from both right and left side as:

$$m^{\text{BP}}_{f_{\text{E}_k^b} \to x_k^b}(x_k^b) = n_{x_k^{b+1} \to f_{\text{E}_k^b}}(x_k^b),\ 1 \le b \le B-1 \tag{D.11}$$

$$m^{\text{BP}}_{f_{\text{E}_k^{b-1}} \to x_k^b}(x_k^b) = n_{x_k^{b-1} \to f_{\text{E}_k^{b-1}}}(x_k^b),\ 2 \le b \le B \tag{D.12}$$

representing the incoming messages from the right and the left side, respectively[4]. Using (D.3) for the BP messages, the outgoing messages of the LPUs are computed as:

$$n_{x_k^b \to f_{\text{E}_k^b}}(x_k^b) \propto q_{x_k^b}(x_k^b) / m^{\text{BP}}_{f_{\text{E}_k^b} \to x_k^b}(x_k^b),\ 1 \le b \le B-1 \tag{D.13}$$

$$n_{x_k^b \to f_{\text{E}_k^{b-1}}}(x_k^b) \propto q_{x_k^b}(x_k^b) / m^{\text{BP}}_{f_{\text{E}_k^{b-1}} \to x_k^b}(x_k^b),\ 2 \le b \le B \tag{D.14}$$

with $q_{x_k^b}(x_k^b)$ given by (D.9).

3.5 Local SIC boosting

One of the key points in the XL-MIMO non-stationary channels is uneven user energy distribution between the sub-arrays. This phenomenon can be utilized to manage the inter-user interference; symbols of the strong users in one sub-array can be detected and other sub-arrays can use this information to cancel the interference from those users.

[4] We assume that the LPUs are located and ordered from left to right.

Here, unlike conventional SIC receivers, there is no central unit to decide which users should be detected at each SIC step. Thus, we introduce a local SIC mechanism that works at each of the LPUs. Upon each update of $q_{x^b_k}(x^b_k)$, a likelihood ratio (LR) [6] is compared with a predefined threshold Γ_{thr} to find the strong users. The LR metric is defined as $\Gamma_k \triangleq \frac{p^k_1}{p^k_2}, \forall k$, where $p^k_1 \geq \cdots \geq p^k_{|\mathcal{M}|}$ are sorted symbol probabilities provided by $q_{x_k}(x_k),\ x_k \in \mathcal{M}$ for each user k.

When the LR metric exceeds the threshold, i.e. when $\Gamma_k > \Gamma_{thr}$, the belief of the corresponding symbol is set to $q_{x^b_k}(x^b_k) = \delta(x^b_k - a^*)$, where a^* denotes the symbol in $\mathcal{M}$ with largest probability in the approximate marginal. Note that this restriction of the belief to a delta function corresponds to a hard decision on the symbol x^b_k. When this belief is propagated to neighboring sub-arrays via the messages in (D.13) and (D.14), it leads to these sub-arrays also adopting the hard decision. Analogous messages are progressively propagated to the neighboring sub-arrays, eventually yielding the same hard symbol decision for all sub-arrays.

3.6 The algorithm

The proposed combined MF-BP receiver mechanism is demonstrated in Algorithm 8. It is composed of three main elements: 1. **Local symbol estimation**, which is done by the VMP at each LPU (steps 3-6, Sec III.C). 2. **The SIC detector**, that is activated if the LR metric is satisfied (steps 7-8, Sec III.D). 3. **Exchange of local estimates**, which takes care of the message exchanges between the LPUs using the BP (steps 2 and 9, Sec III.E).

We use a maximal ratio combining (MRC) initialization technique [6] for the local estimates. Therefore, the initial approximate marginal of the symbol x_k in LPU b is

$$q^0_{x^b_k}(x^b_k) \propto \mathrm{CN}\big(x^b_k; \hat{x}^b_k, \frac{\sum_{k' \neq k}^{K} P_{x_{k'}} \frac{|\tilde{\mathbf{h}}^H_{b,k}\tilde{\mathbf{h}}_{b,k'}|^2}{||\tilde{\mathbf{h}}_{b,k}||^2} + \sigma^2_n}{||\tilde{\mathbf{h}}_{b,k}||^2}\big) \tag{D.15}$$

which is restricted to the symbol alphabet $\mathcal{M}$. Here, $\hat{x}^b_k = \frac{\tilde{\mathbf{h}}^H_{b,k}}{||\tilde{\mathbf{h}}_{b,k}||^2}\mathbf{y}_b$ and P_{x_k} is the user signal power. Also, the operations in steps 9 and 10 of Algorithm 1 make sure that all the users that do not satisfy the LR condition by the end of the algorithm be detected without SIC. Note that the algorithm is applicable to receivers operating with a signal model as in (D.1), regardless of the channel properties.

4 Performance Evaluation

4.1 Channel Model

We adopt the channel model presented in [13], where the effect of VRs have been applied to the *one-ring* model [14]. The channel for each user, $\mathbf{h}_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}_k)$, models the small-scale fading of the channel, with channel covariance matrix $\mathbf{R}_k$. The entries of $\mathbf{R}_k$ are defined as $[\mathbf{R}_k]_{p,q} = \frac{1}{2\Delta}\int_{-\Delta}^{\Delta} \exp\big(j\mathbf{f}^T(\alpha+\theta)(\mathbf{u}_p - \mathbf{u}_q)\big)\mathrm{d}\alpha$, representing the correlation between the channel coefficients of antennas p and q. Here,

Algorithm 8 Proposed combined MF-BP receiver.

```
Result: Symbol estimates for all active users
Initialize: $M$, $K$, $\mathbf{y}$, $B$, $\Gamma_{thr}$, $\mathcal{M}$, VMP iterations $\mathcal{J}$, total iterations $\mathcal{T}$, local detected user
  sets $\mathcal{S} = \{(\mathcal{S}_1 = \phi), \cdots, (\mathcal{S}_B = \phi)\}$.
for $j = 1$ to $\mathcal{T}$ do
    1. Initialize local user sets as $\mathcal{K}_b \triangleq \{1, \cdots, K\}, \forall b$.
    for $b = 1$ to $B$ do
        2. Extract the messages to the $b$th LPU from (D.12) and (D.11).
        3. Use (D.15) to set the initial probabilities as $q^0(x_k)$.
        for $i = 1$ to $\mathcal{J}$ do
            4. Extract $\bar{x}_k^b$ and $\sigma^2_{x_k^b}$ values from $q_{x_k}^{(i-1)}(x_k)$.
            5. Calculate $\bar{\lambda}_b$ and find symbol probabilities $q_{x_k}^{(i)}(x_k^b)$ using (D.9) for all
              the users $k = \{1, \cdots, K\}$.
        end
        6. Set the LPU's estimate as $q_{x_k}(x_k^b) = q_{x_k}^{(i)}(x_k^b)$.
        for $k \in \mathcal{K}_b$ do
            if $\Gamma_k \geq \Gamma_{thr}$ then
                7. Detect the transmitted symbol for $k$ as $\tilde{x}_k$ and $\mathcal{S}_b = \mathcal{S}_b \cup \tilde{x}_k$, $\mathcal{K}_b =$
                  $\mathcal{K}_b \setminus k$ and cancel the interference, fix the prior as $q_{x_k^b}(x_k^b) = \delta(x_k^b -$
                  $\tilde{x}_k)$.
            else if $j = \mathcal{T}$ and $\Gamma_k < \Gamma_{thr}$ then
                8. Detect the user $k$ without SIC; $\tilde{x}_k = \arg\max q_{x_k^b}(x_k^b)$, $\mathcal{S}_b = \mathcal{S}_b \cup \tilde{x}_k$
                  and $\mathcal{K}_b = \mathcal{K}_b \setminus k$.
        end
        9. Compute the right and left outgoing messages in (D.13) and (D.14), respec-
          tively.
    end
end
10. Choose any of the LPUs to get the detected symbol set.
```

$\mathbf{f}(\omega) = -\frac{2\pi}{\lambda}\left[\cos(\omega), \sin(\omega)\right]^T$ is the wave vector with carrier wavelength of λ and angle of arrival of ω and $\mathbf{u}_p, \mathbf{u}_q \in \mathbb{R}^2$ are the position vectors of the antennas p, q within the VR of user k and Δ is angular spread. Angle θ is the azimuth angle of user k with respect to the antenna array. We have $[\mathbf{R}_k]_{p,q} = 0$ when either of the antenna indices p, q is outside the VR for user k. We use a uniform distribution for the center of the VRs within the array and the length of the VRs follows a lognormal(μ_l, σ_l^2) distribution [15].

4.2 Benchmarks

We have implemented five different benchmarks to compare with our proposed algorithm. We choose *matched filter single-user bound* which is the case when the effect of all interfering users is ideally cancelled and the target user's signal is detected by the MRC [13]. Centralized ZF as a linear and central VMP [13] and central minimum mean square error (MMSE-SIC) as non-linear methods are presented. Moreover, we implemented the daisy-chain algorithm from [11] as a benchmark for the distributed receivers and the hierarchical VMP (H-VMP) receiver in [6] as a mixed method that has both local and central processing units.

4.3 Complexity and Exchanged Messages

We have calculated the computational complexity of the VMP method in one LPU in [6], which is $C^{\mathrm{MF}} = \mathcal{J}(K(4+|\mathcal{M}|)+M_b)+3M_bK$ and $\mathcal{J}$ is the number of iterations in the VMP algorithm. The additional complexity by the BP part is $2\mathcal{M}$ multiplications for computing (D.14) and (D.13) for each user and LPU[5]. Moreover, the SIC part requires BK multiplications to check the LR metric. Since activation of the SIC part is not deterministic, we have a variable complexity expression that varies between the best and worst complexity cases. The worst case happens when non of the users satisfy the LR condition and $C_{\mathrm{tot}} = \mathcal{T}\left(BC^{\mathrm{MF}} + K[B(2\mathcal{M}+1)-2\mathcal{M}]\right)$ with $\mathcal{T}$ denoting the number of total BP iterations. The best case is when all of the users are detected in one iteration of the main loop with K/B users detected at each LPU resulting in $C_{\mathrm{tot}} = BC^{\mathrm{MF}} + K(2\mathcal{M}(B-1)+1)$. The complexities for the daisy-chain, the ZF, the MMSE-SIC, the centralized VMP [13] and the hierarchical VMP [6] are $C_{\mathrm{DC}} = M(K+2)$, $C_{\mathrm{ZF}} = \frac{K^3}{3} + MK^2 + MK$, $C_{\text{MMSE-SIC}} = K(K+1)^2(\frac{K}{12}+M)$, $C_{\mathrm{VMP}} = \mathcal{J}(M(3+2K)+MK|\mathcal{M}|)+3MK$ and $C_{\text{H-VMP}} = \frac{K^2}{2}(3M+B(|\mathcal{M}|+4)+1)+MK$, respectively. With a simple comparison, we can see that the complexity grows with a slower slope with respect to both K and M in C_{tot} than the central methods. However, it is more complex than the daisy-chain method which is a trade-off to get a better performance and lower processing delay.

At each iteration of our algorithm, LPUs send $2K$ messages in each direction (two complex numbers $\hat{x}_k^b$, $\sigma^2_{x_k^b}$ for each user). Thus, the total number of exchanged messages (EM) in the algorithm is $4K\mathcal{T}(B\text{-}1)$ which is much lower than the EM for H-VMP method, which is $4BK^2$.

4.4 Simulation results

Here, we present the numerical results evaluating the performance of our proposed method, as well as the benchmarks. Simulation parameters are as follows: QPSK modulation, uniform linear array (ULA) with $M = 300$, $K = 40$, $B = 5$, $\Delta = \frac{\pi}{10}$, $\theta \sim \mathcal{U}(\frac{-\pi}{2}, \frac{\pi}{2})$, $\mathcal{J} = 2$, $\mathcal{T} = 10$, $\Gamma_{thr} = 10^3$, 8×10^4 channel realizations and correlation matrices are updated every 50 realizations. Centers of the VRs are uniformly distributed across the array and $\mu_l = 0.7$ and $\sigma_l^2 = 0.2$.

In Fig. D.2, we compare the symbol error rate (SER) of the proposed method and the benchmarks. The SNR is calculated as SNR$= \frac{1}{\sigma_n^2}$ at the BS side, where we assumed unit received power for all the users over the BS array. Moreover, for our MF-BP method, we include the symbol detection results of the first and $\frac{B}{2}$th LPUs to show the convergence of the local estimates. The performance of the MF-BP method is very close to the central processing techniques and outperforms the daisy-chain

[5] There is only one BP message for the first and the last LPUs.

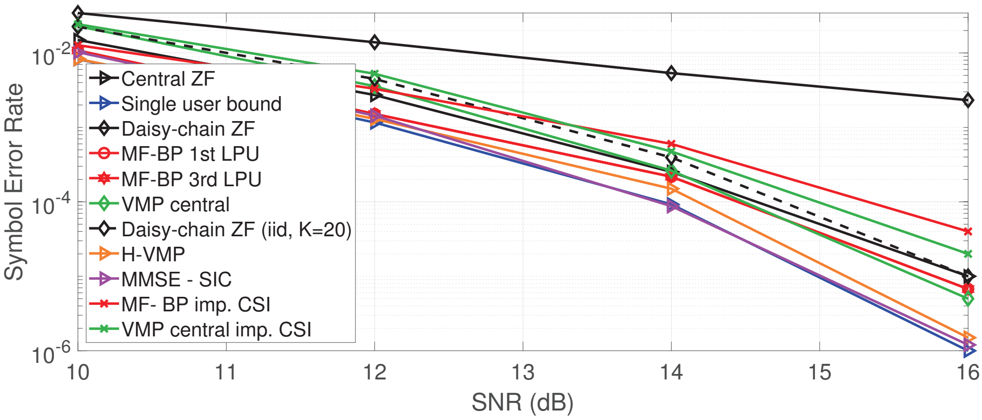


Fig. D.2: SER comparison of the proposed MF-BP method with other benchmarks. ($M = 300$, $K = 40$, $\Delta = \frac{\pi}{10}$)

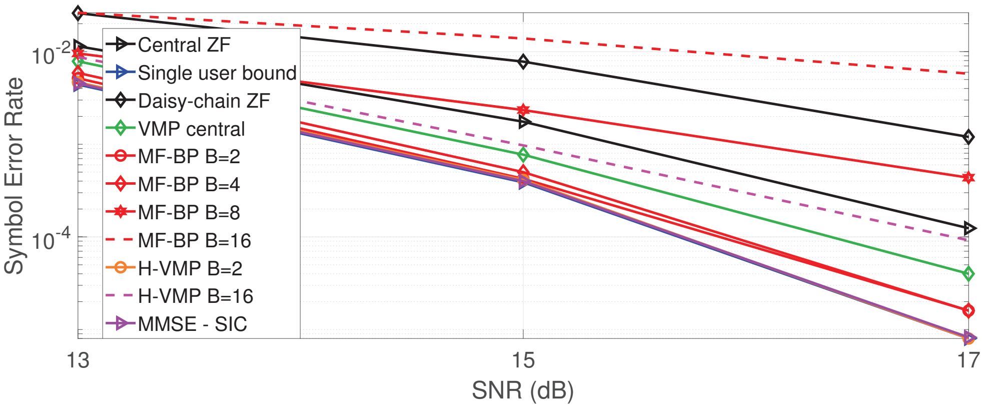


Fig. D.3: The effect of number of LPUs on the SER of the proposed MF-BP method. The curves for the single user bound and the H-VMP $B = 2$ are superposed. ($M = 128$, $K = 25$, $\Delta = \frac{\pi}{5}$)

technique. The reason for the degraded SER of the daisy-chain receiver is due to the channel NS and a high system load $\frac{M}{K} < 10$ that makes it hard for the algorithm to completely cancel the inter-user interference. The dashed line shows the performance of the daisy-chain method in a non-correlated channel with $\frac{M}{K} = \frac{300}{20} > 10$ and confirms the statement above. The H-VMP is outperforming all the methods since it uses a more complex receiver with a central SIC that repeats the detection process K times. Besides, we show the effect of imperfect CSI, where the estimated channel is $\hat{\mathbf{H}} = \sqrt{1-\tau_h}\mathbf{H} + \tau_h\mathbf{Z}$ and τ_h determines CSI accuracy and $\mathbf{Z}$ is modelling Gaussian measurement noise. We assume $\tau_h = 0.1$. As expected, non-accurate CSI is increasing the error rate. One way to alleviate this effect is to include channel estimation in the VMP part which is left for future work.

The effect of number of distributed units, i.e. LPUs, on the SER of the MF-BP scheme is shown in Fig. D.3. The performance of our method is better than the central VMP algorithm for $B = 2, 4$. Also, for these two values of B, the algorithm is giving similar results even though there are fewer antennas per each LPU in the latter case, i.e. $M_b|_{b=4} = \frac{1}{2}M_b|_{b=2}$. This is the result of having more SIC detectors in $B = 4$ that compensate for the lower spatial resolution compared to the $B = 2$ case. On the other hand, as B increases, the number of antennas per LPU reduces (e.g. $M_b = 8$ for $B = 16$) while the number of users is still high ($K = 25$) and the quality of local estimates weakens and results in a poor outcome.

5 Conclusions

We introduce a fully decentralized receiver for the XL-MIMO array that only relies on the LPUs for the user symbol detection. This receiver is scalable and can be deployed easily with minimum inter-connections between the sub-arrays. The ability to operate in parallel is minimizing the processing delay experienced by other benchmark techniques. Moreover, the size of exchanged messages between the LPUs, i.e. communication overhead, is limited and allowing the use of inexpensive backhaul links. Future works will focus on integrating channel estimation and coding within the local units.

References

[1] E. D. Carvalho, A. Ali, A. Amiri, M. Angjelichinoski, and R. W. Heath, “Non-stationarities in extra-large-scale massive mimo,” *IEEE Wireless Communications*, vol. 27, no. 4, pp. 74–80, 2020.

[2] S. Hu, F. Rusek, and O. Edfors, “Beyond massive mimo: The potential of positioning with large intelligent surfaces,” *IEEE Transactions on Signal Processing*, vol. 66, no. 7, pp. 1761–1774, April 2018.

[3] E. Björnson, L. Sanguinetti, J. Hoydis, and M. Debbah, “Optimal design of energy-efficient multi-user mimo systems: Is massive mimo the answer?” *IEEE Transactions on Wireless Communications*, vol. 14, no. 6, pp. 3059–3075, 2015.

[4] L. Liu, C. Yuen, Y. L. Guan, Y. Li, and C. Huang, “Gaussian message passing for overloaded massive MIMO-NOMA,” *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 210–226, 2019.

[5] A. Amiri, M. Angjelichinoski, E. de Carvalho, and R. W. Heath, “Extremely large aperture massive mimo: Low complexity receiver architectures,” in *2018 IEEE Globecom Workshops*, Dec 2018, pp. 1–6.

[6] A. Amiri, S. Rezaie, C. N. Manchon, and E. de Carvalho, “Distributed receivers for extra-large scale mimo arrays: A message passing approach,” *arXiv preprint arXiv:2007.06930*, 2020.

[7] V. C. Rodrigues, A. Amiri, T. Abrao, E. de Carvalho, and P. Popovski, “Low-complexity distributed xl-mimo for multiuser detection,” *arXiv preprint arXiv:2001.11879*, 2020.

[8] H. Wang, A. Kosasih, C.-K. Wen, S. Jin, and W. Hardjawana, “Expectation propagation detector for extra-large scale massive mimo,” *arXiv preprint arXiv:1906.01921*, 2019.

[9] J. R. Sanchez, F. Rusek, O. Edfors, M. Sarajlic, and L. Liu, “Decentralized massive mimo processing exploring daisy-chain architecture and recursive algorithms,” *arXiv preprint arXiv:1905.03160*, 2019.

[10] C. Zhang, Y. Jing, Y. Huang, and L. Yang, “Performance analysis for massive mimo downlink with low complexity approximate zero-forcing precoding,” *IEEE Transactions on Communications*, vol. 66, no. 9, pp. 3848–3864, 2018.

[11] M. Sarajlić, F. Rusek, J. R. Sánchez, L. Liu, and O. Edfors, “Fully decentralized approximate zero-forcing precoding for massive mimo systems,” *IEEE Wireless Communications Letters*, vol. 8, no. 3, pp. 773–776, 2019.

[12] Z. Zhang, H. Li, Y. Dong, X. Wang, and X. Dai, “Decentralized signal detection via expectation propagation algorithm for uplink massive mimo systems,” *IEEE Transactions on Vehicular Technology*, vol. 69, no. 10, pp. 11 233–11 240, 2020.

[13] A. Amiri, C. N. Manchón, and E. de Carvalho, “A message passing based receiver for extra-large scale mimo,” in *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, 2019, pp. 564–568.

[14] A. Adhikary, J. Nam, J. Ahn, and G. Caire, “Joint spatial division and multiplexing—the large-scale array regime,” *IEEE Transactions on Information Theory*, vol. 59, no. 10, pp. 6441–6463, Oct 2013.

[15] X. Gao, F. Tufvesson, and O. Edfors, “Massive mimo channels—measurements and models,” in *2013 Asilomar conference on signals, systems and computers*. IEEE, 2013, pp. 280–284.

[16] E. Riegler, G. E. Kirkelund, C. N. Manchón, M.-A. Badiu, and B. H. Fleury, “Merging belief propagation and the mean field approximation: A free energy approach,” *IEEE Transactions on Information Theory*, vol. 59, no. 1, pp. 588–602, 2012.

Paper E

Low-Complexity Distributed XL-MIMO for Multiuser Detection

Victor Croisfelt Rodrigues, Abolfazl Amiri, Taufik Abrão,
Elisabeth de Carvalho, Petar Popovski

The paper has been published in the
IEEE International Conference on Communications Workshops (ICC Workshops) Proceedings pp. 1–6, 2020.

The layout has been revised.

Abstract

In this paper, the zero-forcing and regularized zero-forcing schemes operating in crowded extra-large MIMO (XL-MIMO) scenarios with a fixed number of sub-arrays have been emulated using the randomized Kaczmarz algorithm (rKA). For that, non-stationary properties have been deployed through the concept of visibility regions when considering two different power normalization methods of non-stationary channels. We address the randomness design of rKA based on the exploitation of spatial non-stationary properties. Numerical results show that, in general, the proposed rKA-based combiner applicable to XL-MIMO systems can considerably decrease computational complexity of the signal detector by paying with small performance losses.

1 Introduction

Motivated by higher area throughput that extremely large arrays can offer [1], recent notable research efforts are being carried out to improve the scalability of the so-called extra-large MIMO (XL-MIMO) systems. Due mainly to increased spatial resolution and the emergence of non-stationary channels, this new vision is currently materializing as an important beyond 5G technology and being considered as a distinct operating regime of Massive MIMO (M-MIMO) [2]. With physical large arrays, spatial non-stationarity and inherent high array dimensions under user crowded scenarios have significant harmful impacts on the performance and computational complexity of linear receive combining techniques, which are traditionally used in M-MIMO systems [3]. This calls for different manners of performing receive combining in XL-MIMO systems, which try to exploit non-stationarities and seek a good trade-off between performance and computational complexity when a large number of users are served.

Taking into account crowded scenarios and the desire for low cost base stations (BSs), several low-complexity linear detection algorithms that attempt to relax the computation of known linear receive combining criteria have been proposed in recent years for canonical M-MIMO; such as [4, 5] to cite a few. These works, however, do not consider non-stationary channels that appear when antenna arrays are scaled up, as is the case of XL-MIMO. Meanwhile, the authors in [6] propose a variational message passing (VMP) based symbol detection method for XL-MIMO and under crowded scenarios. Although the proposed method outperforms linear receivers, the algorithm demands the optimization of a damping factor, which accelerates the convergence of the algorithm, but unfortunately translates into undesired additional complexity. In addition to that, its complexity depends on the modulation order used to transmit user messages, making the comparison with linear receivers cumbersome. To the best of our knowledge, few are the works that study low-complexity linear receive combining techniques under the presented scenario of interest.

Contributions: Inspired by the promising results obtained for M-MIMO [5, 7, 8], this work proposes the application of the randomized Kaczmarz algorithm (rKA) as a way to circumvent the high-dimensional matrix inversion that comes with zero-forcing (ZF) and regularized zero-forcing (RZF) schemes when these are applied to recover the signal estimates of a crowded XL-MIMO scenario [2]. The contributions are listed as follows: (i) extension of rKA to resemble the performance of ZF and RZF schemes for a XL-MIMO system with a fixed number of sub-arrays; (ii) consideration of non-stationary properties through the concept of visibility regions (VRs) when taking into account two different power normalization methods of non-stationary channels [9]; (iii) exploitation of non-stationary features in the randomness design of rKA; (iv) complexity analysis considering the different random variants of the proposed algorithm.

Some valuable features of the algorithm are as follows. *Simplicity:* the only tuning parameter needed to be set is the number of iterations at each sub-array. The others stem from network design choices and environment characteristics, which obviously affect the convergence of the algorithm, as discussed in [5], [7], and [8]. In opposite to that, this also means that a convergence analysis is sufficient to characterize the efficiency of the algorithm in achieving its goal. *Graceful degradation:* given the computational constraints for any BS, we can flexibly trade off the number of iterations with the performance.

2 System Model

In this section, we describe the uplink transmission phase of a XL-MIMO BS equipped with M antennas that is serving K single-antennas users. The users are using the same time-frequency resources and simultaneously transmitting data to the BS, where narrowband transmissions are considered. From now on, BS is supposed to know the channel state information (CSI) perfectly. This communication setup is shown in Fig. E.1. Let S be the number of fixed sub-arrays that splits an M antenna array into disjoint groups of $M^{(s)} = M/S$ antennas, where $\sum_{s=1}^{S} M^{(s)} = M$ and each group has its own local processing unit for signal detection. A central unit is considered responsible for performing a data fusion operation that combines the soft information received by each sub-array [10]. Furthermore, to ensure the benefits of M-MIMO, it is assumed that $M^{(s)} \geq K$. Thus, sub-array s receives the following baseband signal:

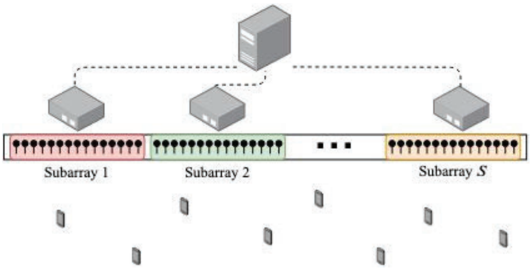


Fig. E.1: XL-MIMO BS with fixed sub-arrays.

$$\mathbf{y}^{(s)} = \sqrt{p}\mathbf{H}^{(s)}\mathbf{x} + \mathbf{n}^{(s)}, \tag{E.1}$$

where p is the uplink transmit power equal to for all users, $\mathbf{H}^{(s)} \in \mathbb{C}^{M^{(s)} \times K} = [\mathbf{h}_1^{(s)}, \dots, \mathbf{h}_K^{(s)}]$ is the channel matrix of sub-array s, $\mathbf{x} \in \mathbb{C}^K$ is a vector that contains the K complex symbols messages with normalized power, and $\mathbf{n}^{(s)} \in \mathbb{C}^{M^{(s)}} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \sigma^2\mathbf{I})$

is a white noise vector. Noise vectors are considered to be independent over s. The $M^{(s)} \times 1$ channel vector with the channel coefficients of user k to $M^{(s)}$ antennas of sub-array s is modeled as [6]

$$\mathbf{h}_k^{(s)} = \sqrt{\mathbf{w}_k^{(s)}} \odot \bar{\mathbf{h}}_k^{(s)}, \tag{E.2}$$

where $\mathbf{w}_k^{(s)}$ embodies large-scale fading effects; path-loss is modeled as $\mathbf{w}_k^{(s)} = \Omega(\mathbf{d}_k^{(s)})^{-\nu}$, where Ω is the path-loss attenuation coefficient, $\mathbf{d}_k^{(s)} \in \mathbb{R}^{M^{(s)}}$ is a vector of the distances between user k and each antenna of sub-array s, and ν is the path-loss exponent. Channel effects resulting from small-scale fading are embraced by $\bar{\mathbf{h}}_k^{(s)} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathbf{\Theta}_k^{(s)})$, where $\mathbf{\Theta}_k^{(s)} \in \mathbb{R}^{M^{(s)} \times M^{(s)}}$ is the sub-array channel covariance matrix that takes into account non-stationarity and spatial channel correlation effects. The overall channel covariance matrix of the antenna array is then $\mathbf{\Theta}_k \in \mathbb{R}^{M \times M} = \text{blkdiag}(\mathbf{\Theta}_k^{(1)}, \ldots, \mathbf{\Theta}_k^{(S)})$ and

$$\mathbf{\Theta}_k = \mathbf{D}_k^{\frac{1}{2}} \mathbf{R}_k \mathbf{D}_k^{\frac{1}{2}}, \tag{E.3}$$

where $\mathbf{R}_k \in \mathbb{R}^{M \times M}$ is a symmetric positive semi-definite matrix that captures spatial channel correlation effects and $\mathbf{D}_k \in \{0,1\}^{M \times M}$ is a diagonal, indicator matrix that embraces non-stationary modeled through the VR concept.

2.1 Visibility Regions (VRs)

The VRs describe the portion of the array being "viewed" by each user, i.e., where most portion of users' energy is concentrated. In particular, we adopt the model described in [6], wherein each user has a VR identified by two main properties: its center and its length. Thus, VR centers are modeled as $c_k \sim \mathcal{U}(0, L)$, where L is the XL-MIMO antenna array physical length, whereas VR lengths $l_k \sim \mathcal{LN}(\mu_l, \sigma_l)$.

Let D_k denote the number of active antennas that are serving user k, which is defined as the sum of array antennas within the physical region delimited by $[c_k - l_k, c_k + l_k]$. Hence, the diagonal matrix $\mathbf{D}_k$, introduced in (E.3), has D_k non-zero diagonal elements. In the sequel, two different power normalizing schemes for the non-stationary channels are revisited [9].

Normalization 1. Stationary and non-stationary channels have the same norm, i.e., $\text{tr}\,(\mathbf{\Theta}_k) = \text{tr}(\mathbf{R}_k) = M\ \forall k$. This is achieved by $\mathbf{D}_k = \text{diag}([\mathbf{0}, (M/D_k)^{1/2}\mathbf{1}_{D_k}, \mathbf{0}]^T)$.

Normalization 2. Non-stationary channels have norm (in general) less than or equal to stationary ones. In this case, $\text{tr}\,(\mathbf{\Theta}_k) = D_k\ \forall k$ and $\mathbf{D}_k = \text{diag}([\mathbf{0}, \mathbf{1}_{D_k}, \mathbf{0}]^T)$.

2.2 Signal-to-Interference-Plus-Noise Ratio (SINR)

Considering that the data symbols of each user are i.i.d. and Gaussian distributed, the instantaneous uplink SINR $\gamma_k^{(s)}$ of user k regarding sub-array s can be defined as [1]:

$$\gamma_k^{(s)} = \frac{p|(\mathbf{v}_k^{(s)})^H \mathbf{h}_k^{(s)}|^2}{p\sum_{i=1, i\neq k}^{K} |(\mathbf{v}_k^{(s)})^H \mathbf{h}_i^{(s)}|^2 + \sigma^2 ||\mathbf{v}_k^{(s)}||^2}, \tag{E.4}$$

where $\mathbf{v}_k^{(s)} \in \mathbb{C}^{M^{(s)}}$ is the receive combining vector of sub-array s. Recall that the objective of this work boils down to obtain an efficient way to compute $\mathbf{v}_k^{(s)}$ in terms of performance-complexity trade-off.

3 Randomized Kaczmarz Signal Detection

The rKA is an iterative algorithm that solves systems of linear equations (SLEs) and has been recently applied to efficiently tackle the problem of relaxing linear signal processing schemes in the context of M-MIMO. This procedure was first presented in [5] and deepened in [7, 8]. The randomization in rKA is related to the order in which the SLE equations are being selected when solved. Adapted and novel random selection methods that exploit non-stationary effects are discussed here.

Each BS with fixed sub-array dimensions is interested in detecting the users's transmitted symbols. In the context of M-MIMO, ZF and RZF are two widely used schemes that, for the sake of argument, can be applied over each sub-array, yielding in the following symbol estimates for $\xi = 0$ (ZF) or $\xi \neq 0$ (RZF):

$$\hat{\mathbf{x}}^{(s)} = (\mathbf{V}^{(s)})^H \mathbf{y}^{(s)} = [(\mathbf{H}^{(s)})^H \mathbf{H}^{(s)} + \xi \mathbf{I}_K]^{-1} (\mathbf{H}^{(s)})^H \mathbf{y}^{(s)}, \tag{E.5}$$

where $\mathbf{V}^{(s)} \in \mathbb{C}^{M^{(s)} \times K} = [\mathbf{v}_k^{(s)} 1, \ldots, \mathbf{v}_k^{(s)} K]$ is the receive combining matrix associated with sub-array s and $\xi = \frac{1}{\text{SNR}} = \frac{\sigma^2}{p}$.

The problems with adopting the procedure described in (E.5) for extremely large arrays are the increased computational cost of the matrix inversion in crowded scenarios and its inherent scalability with the growing number of antennas and sub-arrays. To circumvent this high computational complexity and alleviate/decrease the hardware cost of each sub-array's processing unit, our proposal is to obtain the symbol estimates at each sub-array by still relying on the ZF and RZF methodologies, but instead of using the classical computation form in (E.5), we apply the rKA to obtain them. The main idea behind this is to realize that (E.5) can be posed as the following optimization problem [5]:

$$\arg \min_{\varrho^{(s)} \in \mathbb{C}^K} \|\mathbf{H}^{(s)} \varrho^{(s)} - \mathbf{y}^{(s)}\|_2^2 + \xi \|\varrho^{(s)}\|_2^2, \tag{E.6}$$

which can be compactly written as

$$\arg \min_{\varrho^{(s)} \in \mathbb{C}^K} \|\mathbf{B}^{(s)} \varrho^{(s)} - \mathbf{y}_0^{(s)}\|_2^2, \tag{E.7}$$

where $\varrho^{(s)}$ represents the K symbol estimates at sub-array s, $\mathbf{B}^{(s)} \in \mathbb{C}^{(M^{(s)}+K) \times K} = [\mathbf{H}^{(s)}; \sqrt{\xi} \mathbf{I}_K]$, while $\mathbf{y}_0^{(s)} \in \mathbb{C}^{M^{(s)}+K} = [\mathbf{y}^{(s)}; \mathbf{0}]$. The symbol estimate vector in (E.5) becomes

$$\hat{\mathbf{x}}^{(s)} = [(\mathbf{B}^{(s)})^H \mathbf{B}^{(s)}]^{-1} (\mathbf{B}^{(s)})^H \mathbf{y}_0^{(s)}. \tag{E.8}$$

3.1 Signal Estimates for Each sub-array via rKA

To derive the rKA-based signal detection schemes at each sub-array s, the key idea is to solve the optimization problem by finding the solution of the SLE $\mathbf{B}^{(s)}\boldsymbol{\varrho}^{(s)} = \mathbf{y}_0^{(s)}$ via rKA, considering that each sub-array is an independent MIMO system. However, due to the presence of arbitrary noise in the receive signal, it is possible to observe that this SLE is inconsistent, i.e., if the rKA is applied to solve this system, a high level of residual error would be obtained. To solve this problem, the authors of [5] proposed a suitable transformation over the above SLE to remove the inconsistency, by solving the SLE in two steps: Step.1. Estimation of $\mathbf{y}_0^{(s)}$ as

$$\hat{\mathbf{y}}_0^{(s)} = \mathbf{B}^{(s)}\hat{\mathbf{x}}^{(s)} \overset{(a)}{=} \mathbf{B}^{(s)}([\mathbf{B}^{(s)}]^H\mathbf{B}^{(s)})^{-1}(\mathbf{H}^{(s)})^H\mathbf{y}^{(s)}, \tag{E.9}$$

where in (a) we used (E.8). Note that $\hat{\mathbf{y}}_0^{(s)}$ lies in the subspace spanned by the columns of $\mathbf{B}^{(s)}$. Thus, the following SLE can be obtained:

$$\begin{aligned}(\mathbf{B}^{(s)})^H\hat{\mathbf{y}}_0^{(s)} &= ([\mathbf{B}^{(s)}]^H\mathbf{B}^{(s)})([\mathbf{B}^{(s)}]^H\mathbf{B}^{(s)})^{-1}(\mathbf{H}^{(s)})^H\mathbf{y}^{(s)} \\ &= (\mathbf{H}^{(s)})^H\mathbf{y}^{(s)},\end{aligned} \tag{E.10}$$

which can be written as:

$$(\mathbf{B}^{(s)})^H\mathbf{w}^{(s)} = \mathbf{b}^{(s)}, \tag{E.11}$$

where $\mathbf{w}^{(s)} \in \mathbb{C}^{M^{(s)}+K}$ plays the role of $\hat{\mathbf{y}}_0^{(s)}$ as an unknown vector, while $\mathbf{b}^{(s)} = (\mathbf{H}^{(s)})^H\mathbf{y}^{(s)}$. This SLE outputs $\hat{\mathbf{y}}_0^{(s)}$ and represents the first step to obtain the signal estimates.

Step.2. Without loss of generality, lets assume that $\hat{\mathbf{y}}_0^{(s)}$ can be recovered through the solution of (E.11) via rKA. With $\hat{\mathbf{y}}_0^{(s)}$, the SLE in (E.9) can be solved to obtain the estimates of the symbols transmitted by the users. This second SLE does not need to be solved directly, since, when recovering $\hat{\mathbf{y}}_0^{(s)}$, we can already obtain $\hat{\mathbf{x}}^{(s)}$ via the solution of $(\mathbf{B}^{(s)})^H\mathbf{w}^{(s)} = \mathbf{b}^{(s)}$ by considering the K last components of $\mathbf{w}$ divided by $\sqrt{\xi}$, where $\mathbf{b}^{(s)} = (\mathbf{H}^{(s)})^H\mathbf{y}^{(s)}$.

3.2 Receive Combining Matrix for Each sub-array via rKA

For scenarios where the channel coherence block is large, it turns out that the procedure described above is not computationally efficient, since we have to compute it to get estimates of $\hat{\mathbf{x}}^{(s)}$ at each complex-valued sample of the coherence block. A better way would be to have a method that computes $\mathbf{V}^{(s)}$ only once, and then use this information to compute all the signal estimates concerning a given coherence block[1]. The key to finding a way to get an estimate of the receive combining matrix $\hat{\mathbf{V}}^{(s)}$ is to note that a scaled version of the K receive combining vectors can be acquired when we have K different SLEs of the form $(\mathbf{B}^{(s)})^H\mathbf{w}_i^{(s)} = \mathbf{e}_i$, where $\mathbf{e}_i$ is the ith canonical basis, i.e., a vector comprised of zeros with a single value one in the ith position, for

[1] This procedure, however, would not be adequate in cases where channel responses fluctuate rapidly.

$i = 1, 2, \dots, K$. It can be argued that this SLE results in a scaled estimate of the receive combining vector $\mathbf{v}_k^{(s)} i$ of user i (see further details in Section V of [5]). As a result, if this SLE is solved for each user i, we can obtain an estimate of $\hat{\mathbf{V}}^{(s)}$, which can be used to get the symbol estimates $\hat{\mathbf{x}}^{(s)}$. These observations yield in the procedure summarized in Algorithm 9. Note that the K rKAs carried out by a sub-array s can be executed in parallel in a commodity hardware, i.e., they are independent, their randomness may or may not be shared[2], and the processing can be distributed over cheap, not-so-powerful computing units.

3.3 Algorithm Features and Data Fusion

The main differences of Algorithm 9 for XL-MIMO in comparison to its analogous counterpart for M-MIMO are: (i) the algorithm does not need to run over users that do not have sufficient (or any) power present at sub-array s, see step 5; this comes from the non-stationary nature of extremely large arrays which implies that users are only being served by a limited number of sub-arrays defined by D_k, and (ii) each sub-array's distributed unit needs to execute the algorithm possibly with a different number of iterations $T^{(s)}$ for a central unit to get all symbol estimates $\hat{\mathbf{x}}^{(s)} = (\hat{\mathbf{V}}^{(s)})^H \mathbf{y}^{(s)}$ for $s = 1, \dots, S$; then, the central unit applies a final data fusion step over these estimates to obtain a coherent detection of the symbols sent by all users across the different sub-arrays. In Section 5, we use the distributed linear data fusion (DLDF) receiver described in [10], which attempts to minimize the mean-squared error of users' signal estimates at each sub-array.

3.4 Different Update Schedule Schemes for XL-MIMO

In the context of rKA, the manner and order in which selection of the random rows occurs is often called as the *update schedule*. The convergence speed of the rKA is closely tied to the updating schedule strategy, and this has motivated the study of randomized variants in new application scenarios, such as [7], [11]. This basically translates into the choice of the probability vector $\mathbf{p}^{(s)} = [P_1^{(s)}, \dots, P_K^{(s)}]^T$ in step 12 of Algorithm 9. Below, it is introduced some possible but effective ways to select the rows $r(t)$ in the context of XL-MIMO by trying to exploit non-stationary properties. In particular, we present a novel approach, as well as alter different known ones. It is noteworthy that all three strategies described in the sequel can be thought as different *power allocation* methods. A comparison among the three update schedules is carried out in Section 5.

Power-based update schedule (pwr.)

The traditional rKA sample probability in the context of Algorithm 9 is [11]

$$P_{r(t)}^{(s)} = \frac{\|\mathbf{h}_{r(t)}^{(s)}\|_2^2 + \xi}{\|\mathbf{H}^{(s)}\|_\text{F}^2 + K\xi}. \tag{E.12}$$

[2] The version of Algorithm 9 comes from [8], which considers a self-initialization procedure to ensure and accelerate convergence for all users, i.e. both center- and edge-located users (see Step 10 of Algorithm 9).

Algorithm 9 Receive Combining Matrix Estimation for Each sub-array using rKA.

```
**Input:** Number of sub-array antennas $M^{(s)}$, number of users $K$, inverse of the SNR $\xi \geq 0$ (RZF regularization factor), sub-array channel matrix $\mathbf{H}^{(s)} \in \mathbb{C}^{M^{(s)} \times K}$, and number of iterations $T^{(s)}$.
**Initialization:** Specify $\mathbf{W}^{(s)} \in \mathbb{C}^{K \times K} = \mathbf{0}$.
**Procedure:**
**for** $k \leftarrow 1$ **to** $K$ **do**
    **if** *power of user k is not zero* **then**
        Define state vectors $\mathbf{u}^t \in \mathbb{C}^{M^{(s)}}$ and $\mathbf{z}^t \in \mathbb{C}^K$ with $\mathbf{u}^0 = \mathbf{0}$ and $\mathbf{z}^0 = \mathbf{0}$.
        Define user canonical basis $\mathbf{e}_k \in \mathbb{R}^K$, where $[\mathbf{e}_k]_k = 1$ and $[\mathbf{e}_k]_j = 0, \forall j \neq k$.
        **for** $t \leftarrow 0$ **to** $T^{(s)} - 1$ **do**
            **if** $t = 0$ **then**
                Pick row $k$ of $(\mathbf{H}^{(s)})^H$ as a way to coherently initialize the algorithm and make it fair. This is referred to as *self-initialization* [8].
            **else**
                Pick a row $r(t)$ of $(\mathbf{H}^{(s)})^H$ with $r(t) \in \{1, 2, \dots, K\}$ drawn based on $\mathbf{p}^{(s)}$ (see Section 3.4).
            **end if**
            Compute the residual:
                $$\eta^t := \frac{[\mathbf{e}_k]_{r(t)} - \langle \mathbf{h}_{r(t)}^{(s)}, \mathbf{u}^t \rangle - \xi z_{r(t)}^t}{\|\mathbf{h}_{r(t)}^{(s)}\|_2^2 + \xi}.$$
            Update $\mathbf{u}^{t+1} = \mathbf{u}^t + \eta^t \mathbf{h}_{r(t)}^{(s)}$.
            Update $z_{r(t)}^{t+1} = z_{r(t)}^t + \eta^t$.
            Repeat $z_j^{t+1} = z_j^t, \forall j \neq r(t)$.
        **end for**
        Update $\left[\mathbf{W}^{(s)}\right]_{:,k} = \mathbf{z}^{T^{(s)}-1}$.
    **end if**
**end for**
**Output:** $\mathbf{W}^{(s)}$, $\hat{\mathbf{V}}^{(s)} = \mathbf{H}^{(s)}\mathbf{W}^{(s)}$.
```

This probability can be interpreted as the relative ratio of the power of user $r(t) \in \{1, \ldots, K\}$ to the power of all users in the system. Therefore, users with better channel conditions or/and *now* with more active antennas D_k at a specific sub-array s are more often chosen. Moreover, to compute this sample probability, we need to obtain the K sample probabilities of each user in which each takes $2M^{(s)}$ complex multiplications [3, Appx. B]. In fact, due to non-stationary, not all users will be served by sub-array s, and therefore only $\bar{K}^{(s)}$ sample probabilities need to be computed, where $\bar{K}^{(s)}$ is the average number of users served by each sub-array.

Uniform update schedule (unif.)

A second strategy for the sample probability was suggested by the authors in [7]. The authors of [7] proved that, if the selection of the rows is defined to be uniform with respect to the users i.e., $P^{(s)}_{r(t)} = 1/K^{(s)}$, the rKA also achieves an expected rate of convergence, where $K^{(s)}$ denotes the number of active users at sub-array s. This method can be considered to bring fairness to the update schedule, in the sense that no user-specific equations are preferable. Different from the previous case, we assume that no extra computational complexity is required to compute $\mathbf{p}^{(s)}$.

Active-antennas-based update schedule (a.a.)

Aiming to exploit non-stationary channels, herein, we propose an update schedule scheme which is similar to the uniform one, but now the samples probabilities are based on the number of active antennas $D^{(s)}_k$ of user k at sub-array s. We define the sample probability as

$$P^{(s)}_{r(t)} = \frac{D^{(s)}_{r(t)}}{\sum_{i=1}^{K^{(s)}} D^{(s)}_i}. \tag{E.13}$$

This approach gives more attention to users that have a large number of active antennas at each sub-array. Again, no additional computational complexity is considered.

4 Computational Complexity Analysis

In this section, we characterize the computational complexity of Algorithm 9. To do so, we consider the framework for complexity analysis presented in [3, Appx. B], where only complex multiplications/divisions are taken into account.

Table E.1 summarizes the computational complexity expressions of the traditional ZF and RZF schemes [3], as benchmarks, and of Algorithm 9 when considering the three different update schedule schemes discussed in Section 3.4. Since the computational cost of rKA-based methods are related to $T^{(s)}$, a comparison is made in Section 5 after numerically characterizing some convergence notions. Some observations:

1. We consider that the vector norms $\|\mathbf{h}^{(s)}_{r(t)}\|_2$ are computed once and then, they are stored at each sub-array's processing unit. The only other operation that contributes to the computational complexity is $\langle \mathbf{h}^{(s)}_{r(t)}, \mathbf{u}^t \rangle$ at each iteration t.

2. The reception columns refers to the computation of $\hat{\mathbf{x}}^{(s)} = (\mathbf{V}^{(s)})^H \mathbf{y}^{(s)}$ at each sub-array. From the point of view of low-complexity, we can maintain the output of Algorithm 9 in the factorized form $\mathbf{W}^{(s)}$ and to recover the symbol estimates at each complex-valued sample, perform $\hat{\mathbf{x}}^{(s)} = ([\mathbf{W}^{(s)}]^H([\mathbf{H}^{(s)}]^H \mathbf{y}^{(s)}))$. Let τ_{ul} be the number of complex-valued samples reserved to the uplink phase. Thus, the above operation leads to $\tau_{\text{ul}} M^{(s)} \bar{K}^{(s)}$ at each sub-array distributed unit.
3. We assume that both the canonical in (E.5) and rKA forms of computing the ZF and RZF receive combining matrices are taking advantage of the non-stationary premise that not all users are served by all sub-arrays.
4. The overall computational complexity is given by the computation of all $\hat{\mathbf{x}}^{(s)}$'s, where it is important to note that the number of iterations may vary for each sub-array.

Table E.1: Overall computational complexity per coherence block for the XL-MIMO receive combining schemes based on complex operations

Scheme	**Receive combining matrix**		**Reception**
	Multiplications	*Divisions*	*Multiplications*
ZF	$S[(3(\bar{K}^{(s)})^2 M^{(s)})/2 + (\bar{K}^{(s)} M^{(s)})/2 + ((\bar{K}^{(s)})^3 - \bar{K}^{(s)})/3]$	$S\bar{K}^{(s)}$	$\tau_{\text{ul}} S M^{(s)} \bar{K}^{(s)}$
RZF	$S[(3(\bar{K}^{(s)})^2 M^{(s)})/2 + (3\bar{K}^{(s)} M^{(s)})/2 + ((\bar{K}^{(s)})^3 - \bar{K}^{(s)})/3]$	$S\bar{K}^{(s)}$	$\tau_{\text{ul}} S M^{(s)} \bar{K}^{(s)}$
Alg. 1 (pwr.)	$S[M^{(s)} T^{(s)} + 2M^{(s)} \bar{K}^{(s)}]$		$\tau_{\text{ul}} S M^{(s)} \bar{K}^{(s)}$
Alg. 1 (unif./a.a)	$S[M^{(s)} T^{(s)} + M^{(s)}]$		$\tau_{\text{ul}} S M^{(s)} \bar{K}^{(s)}$

4.1 Deriving Upper Bounds for the Number of Iterations

From Table E.1, one can note that the computational advantage of Algorithm 9 basically depends on the amount of iterations $T^{(s)}$ required for the algorithm to achieve a given convergence notion (an iterative stopping criterion). We now derive upper bounds for $T^{(s)}$ in the sense that, if the average number of iterations required to reach a given convergence notion exceeds these bounds, Algorithm 9 would perform worse than the canonical form of computing the ZF and RZF schemes, given in (E.5). In fact, without loss of generality, we focus only on the RZF scheme from now on[3].

[3] As discussed in [5], [8], most promising results are obtained for the RZF scheme due to the fact that the regularization factor ξ assists in the convergence of the algorithm.

Table E.2: Simulation parameters.

Parameter	Value	Parameter	Value
Cell area	0.1×0.1 km^2	Min. distance	30 m
M	100	Array type	ULA
S	4	Carrier frequency	2.6 GHz
$M^{(s)}$	25	Antenna spacing	2λ m
K	25	L	23.0610 m
p	0 dBm	Channel model	$\mathbf{R}_k = \mathbf{I}_M$
σ^2	$[-55, -40]$ dBm	c_k	$\mathcal{U}(0, L)$
Ω	4	l_k	$\mathcal{LN}(0.1L, 0.1)$
ν	3		

Comparing the rows in Table E.1 and isolating the number of iterations, we have

$$T_{\text{pwr.}}^{(s),\text{up}} = \frac{1}{3}\frac{(\bar{K}^{(s)})^3}{M^{(s)}} + \frac{2}{3}\frac{\bar{K}^{(s)}}{M^{(s)}} + \frac{3}{2}(\bar{K}^{(s)})^2 - \frac{1}{2}\bar{K}^{(s)} \tag{E.14}$$

$$\begin{aligned} T_{\text{unif.,a.a.}}^{(s),\text{up}} = &\frac{1}{3}\frac{(\bar{K}^{(s)})^3}{M^{(s)}} + \frac{2}{3}\frac{\bar{K}^{(s)}}{M^{(s)}} + \frac{3}{2}(\bar{K}^{(s)})^2 \\ &+ \frac{3}{2}\bar{K}^{(s)} - 1. \end{aligned} \tag{E.15}$$

These upper bounds are used in the convergence analysis carried out in Section 5.1.

5 Numerical Results and Discussion

To verify the efficiency of Algorithm 9 in achieving a good performance-complexity trade-off solution for XL-MIMO signal detection, we now collect some quantitative results. The simulation parameters are disposed in Table E.2. The users are uniformly distributed inside a square-cell area with a minimum distance of 30 m to the BS. The extremely large array follows a uniform linear array (ULA) arrangement with spacing between antennas of 2λ m.

5.1 Convergence Analysis

Here, we characterize SNR regions in which the proposed algorithm brings relevant computational gains. To ease the exposition, we define the following quantity called as the *computational relaxation degree* $\text{CRD}_i^{(s)}$:

$$\text{CRD}_i^{(s)} = \frac{T_i^{(s),\text{up}} - \bar{T}^{(s)}}{T_i^{(s),\text{up}}}, \text{ if } \bar{T}^{(s)} < T_i^{(s),\text{up}} \tag{E.16}$$

and 0 otherwise, where $\bar{T}^{(s)}$ is the average number of iterations per sub-array needed to achieve a sense of appropriate convergence and i indexes the different update schedules. This quantity measures the relative computational complexity gains obtained for each sub-array via Algorithm 9 compared to the canonical way of computing the RZF scheme.

Fig. E.2(a) shows the computational relaxation degree as a function of the noise variance in dBm. Note that both ways of normalizing $\mathbf{D}_k$ discussed in Section 2.1 were considered. The average number of iterations were obtained by comparing the average SINR of Algorithm 9 with the average SINR benchmark given by the canonical computation of RZF at each sub-array. Moreover, two stopping criteria were considered in relation to the performance measured via average SINR: (i) Algorithm 9 outputs an estimate of $\mathbf{V}^{(s)}$ that reduces 10% of the canonical performance of the RZF scheme, and (ii) the same but considering a losing in performance of only 1%. We now made some observations:

1. *Average system performance:* uniform update schedule outperforms all the other schemes. This is because, for users with good and bad channel conditions, the algorithm converges properly. This last conclusion may change, since our evaluated metric is based on average.
2. Active-antennas-based update schedule performs marginally better than the typical power-based one and has a considerably easier implementation.
3. Normalization 2 better accelerates the algorithm convergence because of the disparity among users' power, which reduces the overall average signal-to-interference ratio (SIR).

The most important conclusion is that we can roughly resemble the performance provided by RZF by greatly reducing the computational cost. At low SNR, this is easily achieved due to low interference among users.

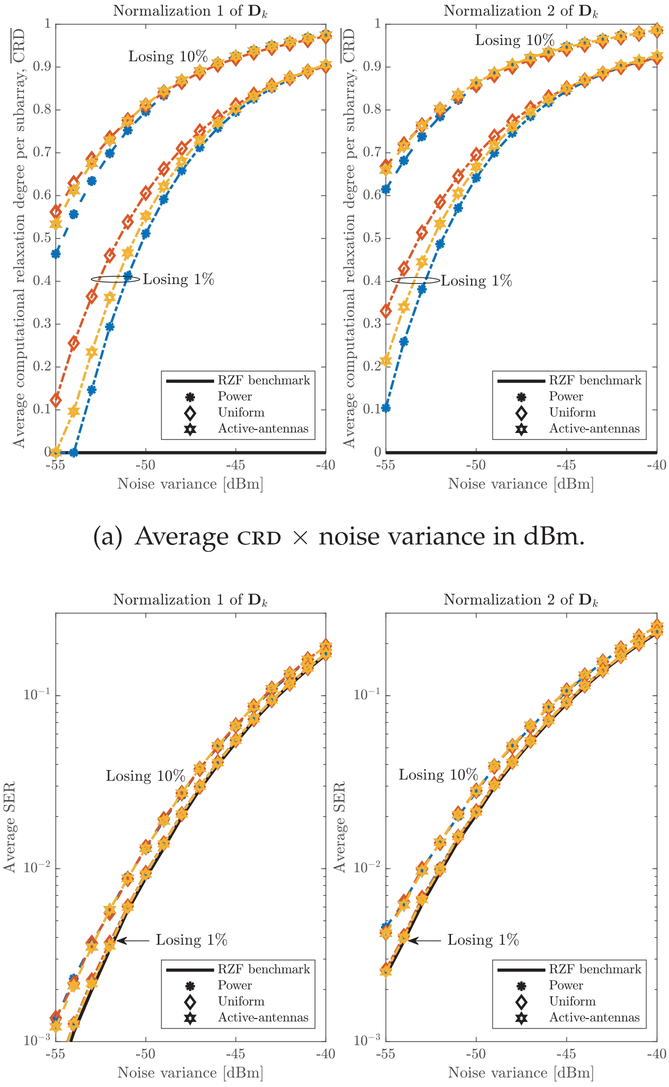
Fig. E.3 illustrates the relaxation in computational complexity brought by the algorithm when considering different system sizes. Note that the RZF complexity has a rapid growth in comparison with the rKA-based schemes as $M^{(s)}$ and K increase. Uniform and active-antennas approaches are the most attractive ones.

5.2 Performance Comparison

To give a notion of the performance gap of the two different adopted stopping criteria, Fig. E.2(b) shows the average symbol error rate (SER) as a function of the noise variance in dBm. The number of iterations used for each different noise variance point follows the results obtained in Fig. E.2(a). To fusion the signal estimates of each sub-array, we used the DLDF receiver described in [10, Algorithm 1]. It is important to observe that, although the algorithm rate of convergence when using Normalization 2 of $\mathbf{D}_k$ is faster (see Fig. E.2(a)), its performance is impaired by the different array gains for each user in comparison with the first normalization method.

6 Conclusions

In this work, we have proposed a rKA-based combiner specifically applicable to XL-MIMO systems aiming at reducing the computational burden of the signal detector



(a) Average CRD × noise variance in dBm.

(b) Average SER × noise variance in dBm.

Fig. E.2: Performance-complexity trade-off. $K/M = 0.25$ and $p = 0$ dBm. Performance gaps of 10% and 1% regarding the canonical RZF scheme and two ways of normalizing $\mathbf{D}_k$ were considered.

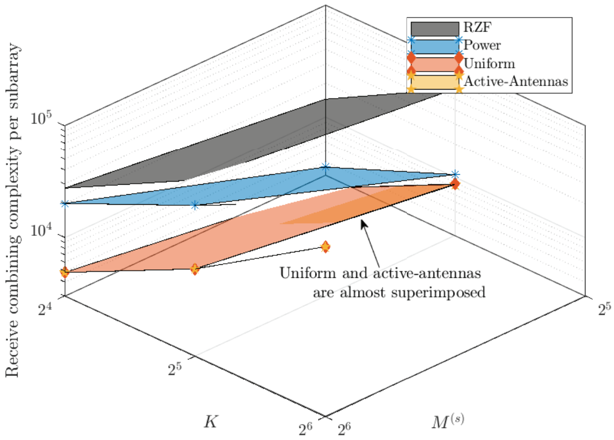


Fig. E.3: Receive combining computational complexity as a function of $M^{(s)}$ and $\bar{K}^{(s)}$. $p = 0$ dBm, $\sigma^2 = 50$ dBm, normalization 2 is considered, and the number of iterations is fixed according to the methods used in Fig. E.2(a) based on losing 10%.

with improved performance-complexity trade-off. We have provided a computational complexity analysis by deriving upper bounds for number of iterations required for convergence (10% or 1% performace losing). Besides, we have proposed a new *update scheduler* for the rKA, namely active-antenna-based update schedule, aiming at exploiting the intrinsic non-stationary properties in XL-MIMO channels. Future research will address optimizing the complexity of systems with different user requirements.

References

[1] E. Björnson, L. Sanguinetti, H. Wymeersch, J. Hoydis, and T. L. Marzetta, “Massive MIMO is a reality—What is next?: Five promising research directions for antenna arrays,” *Digital Signal Processing: A Review Journal*, vol. 94, pp. 3–20, 2019.

[2] E. De Carvalho, A. Ali, A. Amiri, M. Angjelichinoski, and R. W. Heath Jr, “Non-Stationarities in Extra-Large Scale Massive MIMO,” *arXiv preprint arXiv:1903.03085*, 2019.

[3] E. Björnson, J. Hoydis, and L. Sanguinetti, “Massive MIMO Networks: Spectral, Energy, and Hardware Efficiency,” *Foundations and Trends® in Signal Processing*, vol. 11, no. 3-4, pp. 154–655, 2017.

[4] A. Müller, A. Kammoun, E. Björnson, and M. Debbah, “Linear Precoding Based on Polynomial Expansion: Reducing Complexity in Massive MIMO,” J Wireless Com Network 2016, 63 (2016).

[5] M. N. Boroujerdi, S. Haghighatshoar, and G. Caire, “Low-Complexity Statistically Robust Precoder/Detector Computation for Massive MIMO Systems,” *IEEE Transactions on Wireless Communications*, vol. 17, no. 10, pp. 6516–6530, 2018.

[6] A. Amiri, C. N. Manchón, and E. de Carvalho, “A Message Passing Based Receiver for Extra-Large Scale MIMO,” *arXiv preprint arXiv:1912.04131*, 2019.

[7] M. N. Boroujerdi, A. Abbasfar, and M. Ghanbari, “Efficient beamforming scheme in distributed massive MIMO system,” *International Symposium on Turbo Codes and Iterative Information Processing, ISTC*, vol. 2018, pp. 1–5, 2019.

[8] V. C. Rodrigues, J. C. Marinello Filho, and T. Abrão, “Randomized Kaczmarz algorithm for massive MIMO systems with channel estimation and spatial correlation,” *International Journal of Communication Systems*, p. e4158, sep 2019.

[9] A. Ali, E. De Carvalho, and R. W. Heath, “Linear Receivers in Non-Stationary Massive MIMO Channels with Visibility Regions,” *IEEE Wireless Communications Letters*, vol. 8, no. 3, pp. 885–888, 2019.

[10] A. Amiri, M. Angjelichinoski, E. De Carvalho, and R. W. Heath, “Extremely Large Aperture Massive MIMO: Low Complexity Receiver Architectures,” *2018 IEEE Globecom Workshops, GC Wkshps 2018 - Proceedings*, 2019.

[11] T. Strohmer and R. Vershynin, “A Randomized Solver for Linear Systems with Exponential Convergence,” In: Díaz J., Jansen K., Rolim J.D.P., Zwick U. (eds) Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques. APPROX 2006, RANDOM 2006. Lecture Notes in Computer Science, vol 4110. Springer, Berlin, Heidelberg, 2006.

Paper F

Accelerated Randomized Methods for Receiver Design in Extra-Large Scale MIMO Arrays

Victor Croisfelt, Abolfazl Amiri, Taufik Abrão, Elisabeth de Carvalho, Petar Popovski

The paper has been published in the
IEEE Transactions on Vehicular Technology (TVT), vol. 70, no. 7, pp.6788–6799,
2021.

The layout has been revised.

Abstract

Massive multiple-input-multiple-output (M-MIMO) features a capability for spatial multiplexing of large number of users. This number becomes even more extreme in extra-large (XL-MIMO), a variant of M-MIMO where the antenna array is of very large size. Yet, the problem of signal processing complexity in M-MIMO is further exacerbated by the XL size of the array. The basic processing problem boils down to a sparse system of linear equations that can be addressed by the randomized Kaczmarz (RK) algorithm. This algorithm has recently been applied to devise low-complexity M-MIMO receivers; however, it is limited by the fact that certain configurations of the linear equations may significantly deteriorate the performance of the RK algorithm. In this paper, we embrace the interest in accelerated RK algorithms and introduce three new RK-based low-complexity receiver designs. In our experiments, our methods are not only able to overcome the previous scheme, but they are more robust against inter-user interference (IUI) and sparse channel matrices arising in the XL-MIMO regime. In addition, we show that the RK-based schemes use a mechanism similar to that used by successive interference cancellation (SIC) receivers to approximate the regularized zero-forcing (RZF) scheme.

Keywords— massive MIMO; extra-large scale massive MIMO; randomized Kaczmarz algorithm; receiver design.

1 Introduction

Early deployments of fifth generation (5G) networks are already exploiting massive multiple-input multiple-output (M-MIMO) technology to cope with the rapid growth in the number of users and data traffic [1]. The benefits from the M-MIMO topology come from the spatial multiplexing of the users on the same time-frequency resources. However, the common choice for compact antenna arrays limits the spatial dimension of such systems, reducing the performance gains achievable in practice. One way to enhance the promised benefits of M-MIMO is to scale up the number of antenna elements at the base station (BS). Systems that embrace antenna arrays of extremely large dimensions can better separate a large number of users, significantly increasing overall performance. This uncovers a new regime of M-MIMO referred to as the extra-large scale MIMO (XL-MIMO) [2].

Despite the potential benefits, a disadvantage of extremely large antenna arrays is the excessively high computational complexity concerning signal processing at the receiver. The reason is that inter-user interference (IUI) management is necessary to deal with a large number of users, motivating the use of more intricate receiver designs. The canonical regularized zero-forcing (RZF) is one of these schemes that can offer near-optimum performance in many scenarios [3]. Unfortunately, applying the RZF scheme implies calculating the inverse of large matrices, which needs very high computational capacity at the processing units. This motivates the design of schemes that match the RZF performance, while offering complexity that scales better with the number of antennas and users.

Another practical challenge for receivers when increasing the dimension of the antenna arrays is the emergence of new channel effects. With a larger array, different users experience the same channel paths with variable energy or totally different channel paths. This effect results in a variable mean energy value along the array that is called *spatial non-stationarities* [2]. In contrast, the term *spatial stationarity* refers to the case where the energy variations along the array is negligible. Non-stationarities give rise to sparse channel matrices due to the possibility that the user's energy is concentrated only in a small part of the large antenna array. This uneven energy distribution limits the performance of conventional linear receivers, *e.g.*, zero-forcing (ZF) [4]. Thus, there is a need for low-complexity receiver designs that are aware of such non-stationarities.

1.1 Related work

Many recent works address the design of low-complexity receivers in the context of multi-antenna systems. One of the most common techniques consists of approximating the matrix inverse in the RZF scheme. There are three main approximation techniques: approximate matrix inversion algorithms [5, 6], matrix gradient search methods [7], and iterative solvers of systems of linear equations (SLEs) [8–10]. These methods provide ways to manage the performance and complexity trade-off. However, they face some challenges that can decrease their applicability. The first two have limited control over the performance-complexity management and can involve steps that can still be considered complicated and costly from implementation point of view. For example, the truncated polynomial expansion (TPE) technique used in [5, 6] has iterations comprised of matrix products and further processing is needed to fine tune parameters. Iterative solvers of SLEs, on the other hand, depend dramatically on their convergence rate. In this paper, we focus on the third category in order to increase its applicability using acceleration techniques that are premised on simplicity.

Among the iterative solvers of SLEs, the Kaczmarz algorithm [11] is a popular approach for solving very large SLEs, fitting well with our application scenario. In [12], a randomized Kaczmarz (RK) algorithm was introduced and shown to have an excellent convergence behavior. The authors of [10] introduced a low-complexity receiver design to approximate the RZF scheme based on the RK algorithm of [12] for M-MIMO. There are two main features in favor of the RK-based RZF scheme of [10]. First, the scheme is *simple*, meaning that there is no need to adjust any parameters other than the number of iterations or to know second-order channel statistics. Second, the scheme is *flexible*, which means that it can easily control performance and complexity with great granularity by adjusting the number of iterations.

However, two known problems with the RK algorithm were not treated in [10]. The RK algorithm randomly selects one of the SLE equations to be solved in a given iteration. This equation sampling is based on a probability criterion where probabilities are proportional to the energy of the equations, giving rise to the following weaknesses: (a) low-energy equations are rarely selected, and (b) performance of the RK algorithm is deteriorated when the energy of the equations are very similar [13–15]. We refer to these weaknesses as the *problem with rare equations* and the *curse of uniform normalization*, respectively. Then, the low-complexity receiver of [10] performs poorly when some users are located at the cell-edge or a user power control scheme has been employed. Because of this, we call the receiver scheme in [10] as nRK, short for naive

RK. Besides the above intrinsic issues, it has been found in [13] that the *RK algorithm can fail under certain sparsity conditions*, hindering the operation of the nRK in sparse channels characteristics of the XL-MIMO regime.

Recent interest in solving sparse SLEs for neural network training and other machine learning problems has motivated the research on accelerated RK algorithms, such as the greedy RK (GRK) of [13] and the randomized sampling Kaczmarz (RSK) of [16]. The accelerated RK algorithms can address the three presented weaknesses of the RK algorithm to some extent. In this paper, we embrace this observation and introduce three different accelerated RK-based receiver designs to compete against the nRK scheme of [10].

Distributed receiver designs are also being studied to further alleviate the complexity in the XL-MIMO systems [17, 18]. We share the belief that the distributed approach is the way to further reduce complexity when it comes to XL-MIMO systems, due to the excess of complexity and information management brought by the large number of antennas. For the sake of tractability, however, here we focus on a centralized receiver design in order to cover simultaneously the discussion of both M-MIMO and XL-MIMO regimes. Since centralized designs do not suffer from the inevitable performance loss given the decentralization process [17], they can be advantageous when the number of antenna modules is limited and the hardware does not suffer from unsustainable processing capacity and excessive information communication. Furthermore the distributed framework derived in our previous work [19] can be used to generalize the receivers presented herein for a XL-MIMO system comprised of sub-arrays [2].

1.2 Contributions

In this paper, we introduce three low-complexity receiver designs based on the accelerated RK-algorithms for M-MIMO and XL-MIMO systems. These acceleration techniques are using the following heuristics: (i) the sampling without replacement (SwoR) technique, (ii) the GRK algorithm of [13], and (iii) the RSK algorithm of [16]. To the best of our knowledge, this is the first work that uses these accelerated RK-based algorithms to design receivers for multi-antenna systems. Our schemes work by approximating the performance of the RZF, while providing control over its complexity. This control is realized by only adjusting the number of iterations of the algorithms. The proposed schemes are executed at a central processing unit (CPU). Below, we summarize the contributions of this work:

- We present three flexible receivers that are able to select points of operation from a two-dimensional space defined by performance and complexity. The upper bound of performance is provided by the RZF scheme and the performance range is discretized by the number of iterations.
- We show that one can interpret the RK-based receivers in general to perform a kind of successive interference cancellation (SIC) procedure, giving us a better understanding of how the RK algorithms work when approaching the RZF scheme from the standpoint of classical literature.
- We provide a detailed complexity analysis, showing that our schemes have

more scalable computational complexities with respect to the number of antennas and users.

The remainder of this paper is organized as follows. Section 2 defines a single-cell uplink system model suitable for the M-MIMO and XL-MIMO regimes. Section 3 introduces the mathematical framework needed to approximate the RZF scheme based on the RK algorithms. The proposed accelerated RK-based RZF schemes are described in Section 4, while Section 5 provides a better interpretation of them. Complexity analysis and numerical results are given in Section 6 followed by the main conclusions summarized in Section 7.
Notations: We use upper and lower case boldface to denote matrices and vectors, while non-boldface are used for constants. Discrete and continuous sets are given by calligraphic $\mathcal{X}$ and blackboard bold $\mathbb{X}$. Cardinality of a set is given by $|\mathcal{X}|$. The n-th element of $\mathbf{x}$ is denoted as x_n. The m,n-th element of the matrix $\mathbf{X}$ is $[\mathbf{X}]_{m,n}$, while $[\mathbf{X}]_{:,n}$ represents the n-th column vector of $\mathbf{X}$. Vertical and horizontal matrix concatenations are $[\mathbf{X};\mathbf{Y}]$ and $[\mathbf{X},\mathbf{Y}]$, respectively. We indicate transpose and Hermitian transpose by $(\cdot)^T$ and $(\cdot)^H$. Identity matrix of size n is denoted as $\mathbf{I}_n$, while $\mathbf{0}_{m\times n}$ is an $m \times n$ matrix of zeros. Trace and diagonal matrix operators are denoted respectively by tr$\cdot$ and diag$\cdot$. The $l2$- and Frobenius norms are given by $||\mathbf{x}||_2$ and $||\mathbf{X}||_F$, respectively. Gaussian distribution is represented by $\mathcal{N}(\cdot,\cdot)$, whereas circularly symmetric complex-Gaussian distribution is $\mathcal{CN}(\cdot,\cdot)$. The indicator function with argument x over the set $\mathcal{A}$ is denoted as $\chi_{\mathcal{A}}(x)$, where $\chi_{\mathcal{A}}(x) = 1$ if $x \in \mathcal{A}$, and zero otherwise. Floor operation is $\lfloor\cdot\rfloor$.

2 System Model

We consider the uplink payload data transmission of an M-MIMO system wherein a BS with M antennas simultaneously serves a total of $K < M$ single-antenna users. For convenience, the group of users is indexed by the set of integers $\mathcal{K} = \{1, 2, \dots, K\}$, while $\mathcal{M} = \{1, 2, \dots, M\}$ is the set of antenna indexes. Moreover, we assume the block-fading channel model where the channel vector $\mathbf{h}_k \in \mathbb{C}^{M\times 1}$ of user $k \in \mathcal{K}$ is constant and frequency-flat within a coherence block [3]. When all users transmit simultaneously, the BS receives the following narrowband baseband signal $\mathbf{y} \in \mathbb{C}^{M\times 1}$:

$$\mathbf{y} = \sqrt{\rho}\mathbf{H}\mathbf{x} + \mathbf{n}, \tag{F.1}$$

where ρ is the user transmit power, $\mathbf{H} \in \mathbb{C}^{M\times K} = [\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_K]$ is the channel matrix perfectly known by the BS, $\mathbf{x} \in \mathbb{C}^{K\times 1}$ is the transmitted signal vector, and $\mathbf{n} \in \mathbb{C}^{M\times 1} \sim \mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{I}_K)$ is the receiver noise vector with noise power σ^2. The vector $\mathbf{x}$ is composed of the modulated transmission symbols sent by each user independently, where x_k is drawn from a normalized constellation sequence $\mathcal{X}$. Our goal is to design an efficient and reliable receiver that coherently combines the M observations of the received signal $\mathbf{y}$ and produce a *soft estimate* $\hat{\mathbf{x}}$ for $\mathbf{x}$. Throughout this work, we consider a traditional baseband processing architecture, where a CPU is responsible for all processing activities related to the signal reception in the antenna array.

Table F.1: Summary of the Low-Complexity Accelerated RK-Based RZF Receiver Designs

Scheme	Acceleration Method	Advantages	Disadvantages
nRK-RZF [10]	none	least costly iteration ↓ complexity due to sparsity (XL-MIMO)†	↓ performance for cell-edge users and user power control ↓ weak against IUI and sparsity
RK-RZF (Algorithm 1)	SwoR	**best benefit-cost ratio** ↑ performance for cell-edge users and user power control ↑ robust to sparsity than nRK-RZF	± robust against IUI iteration cost grows linearly with K
GRK-RZF (Algorithm 2)	complete residual info.	**best under extreme conditions** ↑ performance for cell-edge users and user power control ↑ robust against IUI and sparsity smallest number of iterations to converge	most costly iterations
RSK-RZF (Algorithm 3)	partial residual info.	same from GRK-RZF to a much smaller extent intermediate iteration cost (between RK and GRK)	↑ number of iterations to converge w.r.t. GRK

† All other receivers inherent the complexity reduction due to sparsity.

2.1 General Channel Model

We adopt the correlated and non-stationary Rayleigh fading channel model proposed in [4], suitable for transmissions at sub-6 GHz frequencies. Following this model, the channel vector of user $k \in \mathcal{K}$ is defined by $\mathbf{h}_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Theta}_k)$, where $\mathbf{\Theta}_k \in \mathbb{C}^{M \times M}$ is the general channel covariance matrix. This matrix can be decomposed as [4]:

$$\mathbf{\Theta}_k = \mathbf{D}_k^{\frac{1}{2}} \mathbf{R}_k \mathbf{D}_k^{\frac{1}{2}}. \tag{F.2}$$

where $\mathbf{R}_k \in \mathbb{C}^{M \times M}$ is the spatial correlation matrix and $\mathbf{D}_k \in \{0,1\}^{M \times M}$ is an indicator diagonal matrix. Further, the vector with the large-scale coefficients of user k is defined as $\boldsymbol{\beta}_k = \text{diag}\mathbf{R}_k = [\beta_k^1, \dots, \beta_k^M]^T$. The diagonal matrix $\mathbf{D}_k$ models the portions of the antenna array "seen" by each user through the concept of visibility regions (VRs) [2], where $[\mathbf{D}_k]_{m,m} = 1$ indicates that antenna $m \in \mathcal{M}$ sees user $k \in \mathcal{K}$, $[\mathbf{D}_k]_{m,m} = 0$ indicates otherwise. For convenience, we assume that $\text{tr}\mathbf{D}_k = D$, $\forall k \in \mathcal{K}$, where $D \leq M$ is the number of *visible* antennas. The term visible indicates that only D antennas have the major contribution in the communication of any user to the BS, but the visible antenna indices can differ between users. One of the key differences between the M-MIMO and XL-MIMO regimes is in their corresponding $\mathbf{D}$ matrix that determines the stationarity of the received energy over the BS array. For the stationary case $\mathbf{D} = \mathbf{I}_M$, since the array is compact. Spatial non-stationarities impose a sparse structure into the channel matrix $\mathbf{H}$ that can be exploited for simpler receiver designs. In fact, one of the main motivations of this work is to design efficient receivers using such information to reduce the computational complexity.

3 Preliminaries

In this section, we introduce the RZF scheme as one of the state-of-the-art receivers used in the literature [3]. We argue that the straightforward implementation of the RZF scheme may not be attractive from the hardware point of view. To solve this problem, we interpret its solution as an optimization problem that can be solved through an SLE. Then, we describe the process of acquiring a consistent SLE that meets our needs.

A conventional scheme suitable for a scenario where IUI is a problem and the signal-to-noise-ratio (SNR) of users may vary overly is the RZF scheme [3]. Employing the RZF to combine coherently the payload information in $\mathbf{y}$ yields in [3]

$$\hat{\mathbf{x}}^{\text{RZF}} = \left(\mathbf{V}^{\text{RZF}}\right)^{H} \mathbf{y} = \left(\mathbf{H}^{H}\mathbf{H} + \xi \mathbf{I}_{K}\right)^{-1} \mathbf{H}^{H}\mathbf{y}, \tag{F.3}$$

where $\mathbf{V}^{\text{RZF}} \in \mathbb{C}^{M \times K}$ is the RZF receive combining matrix, $\hat{\mathbf{x}}^{\text{RZF}} \in \mathbb{C}^{K \times 1}$ is the RZF soft estimate, $\xi = \sigma^2/\rho$ is the inverse of the pre-processing user transmit SNR, $\mathbf{G} \in \mathbb{C}^{K \times K} = \mathbf{H}^{H}\mathbf{H}$ is the channel Gramiam matrix, and $\mathbf{R}_{\mathbf{yy}} \in \mathbb{C}^{K \times K} = \mathbf{G} + \xi \mathbf{I}_{K}$ is the sample covariance matrix of the received signal $\mathbf{y}$.

The classical RZF scheme can be viewed as the solution of the following optimization problem [10]:

$$\mathbf{w}^{\star} = \arg \min_{\mathbf{w} \in \mathbb{C}^{K \times 1}} ||\mathbf{H}\mathbf{w} - \mathbf{y}||_2^2 + \xi ||\mathbf{w}||_2^2, \tag{F.4}$$

where $\mathbf{w}^{\star}$ is the optimal solution and corresponds to $\hat{\mathbf{x}}^{\text{RZF}}$. The proof simply follows by taking the derivative of the l_2-regularized least-squares cost function above and equating it to zero. A compact form of the cost function is $||\mathbf{B}\mathbf{w} - \mathbf{y}_0||_2^2$, where $\mathbf{B} = [\mathbf{H}; \sqrt{\xi}\mathbf{I}_K] \in \mathbb{C}^{(M+K) \times K}$ and $\mathbf{y}_0 = [\mathbf{y}; \mathbf{0}_{K \times 1}] \in \mathbb{C}^{(M+K) \times 1}$. Naturally, the solution of this optimization problem can be obtained by solving the thin SLE $\mathbf{B}\mathbf{w} = \mathbf{y}_0$.

The presence of noisy observations in $\mathbf{y}$ hinders the use of iterative solvers over $\mathbf{B}\mathbf{w} = \mathbf{y}_0$. On the other hand, this SLE is inconsistent; meaning that the noisy observations in $\mathbf{y}$ make $\mathbf{y}_0$ not lie in the range of $\mathbf{B}$ [20]. Thus there is no solution set. It is preferable to obtain a consistent SLE with minimum additional complexity cost. We use the transformation proposed in [10] that yields the following consistent, fat SLE:

$$\mathbf{B}^{H}\mathbf{z} = \mathbf{b} = \mathbf{H}^{H}\mathbf{y}, \tag{F.5}$$

where $\mathbf{z} \in \mathbb{C}^{(M+K) \times 1} = [\mathbf{u} \in \mathbb{C}^{M \times 1}; \sqrt{\xi}\mathbf{v} \in \mathbb{C}^{K \times 1}]$. The minimum-norm solution to the SLE above is given by $\mathbf{u} = \mathbf{H}\hat{\mathbf{x}}^{\text{RZF}}$ and $\mathbf{v} = \hat{\mathbf{x}}^{\text{RZF}}$. One can note that the k-th equation of this SLE can be associated with obtaining the k-th component of $\mathbf{v}$, which solution is $\hat{x}_k^{\text{RZF}}$ of user $k \in \mathcal{K}$. *Hence, we can use the terms equation and user interchangeably.*

Remark 1. *(MR Receiver).* The vector with constant terms $\mathbf{b}$ $=\mathbf{H}^{H}\mathbf{y}$ in (F.5) is the maximum-ratio (MR) soft estimate $\hat{\mathbf{x}}^{\text{MR}}$ and the price to pay for consistency [3]. Therefore, the respective upper and lower bounds of receiver performance and complexity are given by the MR scheme in this work.

Remark 2. *(Normal Equations).* The authors of [21] designed Kacmarz-based receivers using the normal SLE $\mathbf{B}^{H}\mathbf{B}\mathbf{w} = \mathbf{B}^{H}\mathbf{y}_0$. The solutions discussed here are easily extended to this case as well. Here, we use the SLE in (F.5) to avoid the additional complexity of acquiring the normal SLE.

Remark 3. (*Three Challenges*). (i) $\mathbf{B}^{H}$ does not have symmetry properties, impeding the application of some classical iterative methods, *e.g.*, conjugate gradient [22]. (ii) the amounts of calculations and storage are limited due to wireless nature, hindering the use of common acceleration techniques, such as preconditioning [22].[1] (iii) $\mathbf{B}^{H}$ is a sparse full rank matrix when the channel is non-stationary $D \neq M$, making the SLE difficult to be solved by some methods, *e.g.*, the RK algorithm [12] depending on the sparse structure. The methods proposed here aim to overcome these challenges.

[1] It is common the use of a relaxation parameter to improve convergence of RK methods [23]. However, this expedient requires adjusting the regularization parameter.

4 Low-Complexity Receiver Designs Based on Accelerated RK Algorithms

In this section, we exploit recently established acceleration techniques for the RK algorithm to increase the applicability of RK-based receiver designs in both M-MIMO and XL-MIMO regimes. We start with an overview of the three introduced accelerated RK-based receivers, justifying the reasons for using the chosen methods and indicating the advantages and disadvantages of each. Then, we give a detailed presentation for each of the schemes.

4.1 Overview: Proposed Receivers

We present three receivers based on variations of the RK algorithm: **(i) RK-RZF:** a receiver that improves the overall performance of nRK-RZF [10] using a SwoR-based acceleration technique, which is simpler than those used in the other two schemes; **(ii) GRK-RZF:** a greedy scheme that exploits the residual information of the SLE in (F.5) to further accelerate convergence [13]; **(iii) RSK-RZF:** a scheme introduced to deal with the complexity disadvantages of the GRK-RZF whilst still exploits part of the acceleration provided by the residual information [16]. The "-RZF" suffix explicitly denotes that the performance of the RZF scheme is being emulated by the RK-based receivers.

Table F.1 summarizes the main differences and advantages/disadvantages of each new accelerated RK-based RZF schemes. The RK-RZF scheme has the best benefit-cost ratio among the three proposed receivers. This means that the RK-RZF is able to reduce the complexity of the RZF scheme with little performance losses for typical numbers of antennas M and users K of the M-MIMO and XL-MIMO systems. On the other hand, the performance of the RK-RZF is drastically affected by high levels of IUI and/or sparsity, and when operating at high SNR regime. The GRK-RZF scheme works better under these extreme conditions, being more robust against IUI, sparsity, and increased SNR. However, the price to pay for these gains can turn the GRK-RZF receiver very costly. The RSK-RZF scheme is an effort to reduce the cost while hold part of the benefits of the GRK-RZF. The region of applicability of the RSK-RZF receiver is very limited though.

Figure F.1 illustrates the main steps that are common to all the proposed accelerated RK-based RZF receivers when considering a centralized baseband architecture coordinated by a CPU. First, the CPU uses the received signal $\mathbf{y}$ to calculate the common information which is fixed for all the iterations.[2] Then, depending on the selected scheme (from Table F.1), it starts with the user symbol estimation process. In general, the probability criterion used to select the equations from the SLE in (F.5) differs between the algorithms, while the steps used to update the solution are the same. We refer to this set of steps as the *Kaczmarz update step*, since it follows the classical Kaczmarz algorithm [11]. If a pre-defined stopping criterion is fulfilled,

[2] As a practical matter, if the coherence block is large enough, it is more efficient to calculate $\mathbf{V}^{\text{RZF}}$ in (F.3) just once and then use it until the end of the coherence block. Actually, the schemes described here can be generalized in this way by following the lines in [10, 19, 24].

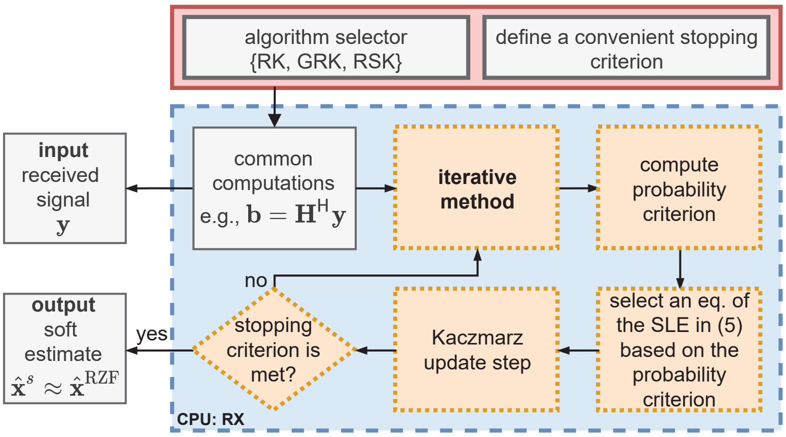

algorithm selector {RK, GRK, RSK}
define a convenient stopping criterion
input received signal $\mathbf{y}$
common computations e.g., $\mathbf{b} = \mathbf{H}^{\mathrm{H}}\mathbf{y}$
iterative method
compute probability criterion
no
output soft estimate $\hat{\mathbf{x}}^s \approx \hat{\mathbf{x}}^{\mathrm{RZF}}$
yes
stopping criterion is met?
Kaczmarz update step
select an eq. of the SLE in (5) based on the probability criterion
CPU: RX

Fig. F.1: Illustration of the basic steps realized by the proposed low-complexity receivers based on accelerated RK algorithms in a centralized baseband architecture, where the CPU is carrying out the signal reception (RX).

the iterative method converges and the CPU obtains the soft estimate $\hat{\mathbf{x}}^s$, which is an approximation of $\hat{\mathbf{x}}^{\mathrm{RZF}}$, where the superscript s indexes the proposed schemes in $\{\mathrm{RK}, \mathrm{GRK}, \mathrm{RSK}\}$ according to Table F.1. Otherwise, the algorithm will continue until a certain maximum number of iterations. The complexity analysis and the stopping criterion are discussed in Subsection 6.1.

4.2 Randomized Kaczmarz Algorithm with SwoR

Algorithm 10[3] summarizes the RK method applied to solve (F.5) when adopting the SwoR technique in Step 11. We refer to Algorithm 10 as the RK-RZF scheme. Except for the application of the SwoR technique, the description of the nRK-RZF [10] scheme is the same as that used in Algorithm 10. However, this seemingly small modification leads to important implications as we discuss below.

The RK-RZF scheme works as follows: Steps 1-7 are comprised of initialization of variables and common computations that will be used throughout the iterative process. In Step 7, the sampling probability vector $\mathbf{p}$ is calculated in (F.6), representing the probability criterion used to select the equations from the SLE in (F.5). As long as a stopping criterion is not met, the algorithm randomly selects in Step 9 one of the equations based on $\mathbf{p}$, where the superscript (t) indicates the current iteration. After choosing an equation index $i^{(t)} \in \mathcal{K}$, the method projects orthogonally the last iterative solution $\mathbf{z}^{(t)} = [\mathbf{u}^{(t)}, \mathbf{v}^{(t)}]$ onto the solution hyperplane $b_{i^{(t)}} = \mathbf{h}_{i^{(t)}}^H \mathbf{u}^{(t)} + \xi v_{i^{(t)}}$, as described in Step 10. This orthogonal projection is seen as the residual $r_{i^{(t)}}^{(t)}$ in relation to the $i^{(t)}$-th equation. In Step 11, this residual is normalized by the energy of the chosen equation $||\mathbf{h}_{i^{(t)}}||_2^2 + \xi$. In Steps 12 and 13, the solution $\mathbf{z}^{(t)}$ is updated into $\mathbf{z}^{(t+1)}$, considering the contribution $\gamma^{(t)}$ of the normalized orthogonal projection of the last iterate over equation $i^{(t)}$. The *Kaczmarz update step* in Fig. F.1 can be defined

[3] Two key observations about the description of all algorithms: (a) a count in terms of floating-point operation per second (FLOPs) is annotated after "%" and (b) the keyword "Store" stands for one time calculations.

according to the realization of Steps 10-14 and is common to all the algorithms in this paper. When the stopping criterion is met, the iterative process terminates and $\mathbf{v}^{(t)}$ is considered to be the RK-RZF soft estimate $\hat{\mathbf{x}}^{\text{RK}}$, an approximation of $\hat{\mathbf{x}}^{\text{RZF}}$.

SwoR. Let $\mathcal{P}^{(t)}$ denote the population from which the equations are sampled available in Step 9 at iteration t. At $t = 0$, we have $\mathcal{P}^{(0)} = \mathcal{K}$. Applying the SwoR technique implies that $\mathcal{P}^{(t+1)} = \mathcal{P}^{(t)} \setminus \{i^{(t)}\}$ until the end of a sweep. At an arbitrary iteration t', we define a *sweep* as the cycle of K iterations that consists of bringing $|\mathcal{P}^{(t')}|$ from $|\mathcal{K}|$ elements to 1 element. After the end of a sweep, a new sweep begins with $\mathcal{P}^{(t'+K)} = \mathcal{K}$ and so on. Because $\mathcal{P}^{(t)}$ changes at each new iteration, the sampling probability vector $\mathbf{p}$ in Step 9 needs to be constantly re-scaled in Step 11 considering the elements in $\mathcal{P}^{(t)}$.

Probability criterion based on energy information of the equations

The key feature of the RK algorithm [12] is the sub-optimal probability criterion in (F.6) that dictates how the equations of the SLE in (F.5) will be selected.[4] This criterion is based on the *energy information of the equations*. Note that the k-th entry of the sampling probability vector $\mathbf{p}$ is the ratio of the regularized channel gain of user k to the sum of the regularized channel gains of all users. And that the desired solution $v_k^{(t)}$ is only updated if the k-th equation is selected. Two bad phenomena can occur if $\mathbf{p}$ is poorly scaled: (a) the weakest users will not be selected as often resulting in a poor performance; (b) convergence performance is naturally degraded if the users experience similar channel gains, which implies that $p_k \approx 1/K$, $\forall k \in \mathcal{K}$. Both problems can be partly eliminated in a heuristic way by the SwoR technique in Step 11 of Algorithm 10. First, the SwoR technique avoids selecting the same equation in sequence, increasing the frequency of selection of the weakest users. Second, the SwoR is changing the selection probabilities constantly to lower the effect of uniform normalization curse. On the flip side, the SwoR technique comes with the price of re-scaling $\mathbf{p}$ in Step 11 after each new iteration, which causes the cost per iteration to grow linearly with K. Motivated by these limited solutions provided by the RK-RZF, in the following we seek other acceleration methods to improve the nRK-RZF [10] receiver.

4.3 Greedy Randomized Kaczmarz Algorithm

The GRK algorithm is an accelerated version of the RK algorithm proposed in [13]. The main idea is to eliminate the equations with larger residuals as quickly as possible. This heuristic design deals with the problem of rare equations and the curse of uniform normalization present in the RK algorithm. Beyond that, the GRK has improved convergence in comparison to the RK algorithm when solving sparse SLEs in general. The application of the GRK to solve (F.5) is given in Algorithm 11, namely, the GRK-RZF scheme.

We start by explaining the functionality of the GRK-RZF scheme. At the beginning of the iterative approach, the current residual vector $\mathbf{r}^{(t)}$ is calculated in Step

[4] Due to natural wireless channel variations, the SLE in (F.5) is constantly changing. Finding the optimum probability of sampling the equations is very costly and should be computed many times, and, therefore, avoided here.

Algorithm 10 RK-with-SwoR-Based Receiver (RK-RZF)

```
1: **Input:** $\mathbf{H}$, $\mathbf{y}$, $M$, $K$, $\xi$
2: **Output:** $\hat{\mathbf{x}}^{\mathrm{RK}}$, $T^{\mathrm{RK}}$
3: $\mathbf{b} = \mathbf{H}^H \mathbf{y}$                 % $8KM - 2K$ FLOPs
4: Store $\{||\mathbf{h}_k||_2^2 + \xi\}$        % $8KM - K$ FLOPs
5: Store $||\mathbf{H}||_F^2 + K\xi = \sum_{k \in \mathcal{K}} (||\mathbf{h}_k||_2^2 + \xi)$        % $K - 1$ FLOPs
6: $\mathbf{u}^{(0)} \in \mathbb{C}^{M \times 1} = \mathbf{0}_{M \times 1}$
7: $\mathbf{v}^{(0)} \in \mathbb{C}^{K \times 1} = \mathbf{0}_{K \times 1}$
8: $t = 0$
9: Probab. criterion based on energy info. of eqs. $\mathbf{p} \in \mathbb{R}^{K \times 1}$:

   $$p_k = \frac{||\mathbf{h}_k||_2^2 + \xi}{||\mathbf{H}||_F^2 + K\xi}, \ \forall k \in \mathcal{K} \qquad \text{(F.6)}$$

10: **While** stopping criterion **is** False do
11: pick $i^{(t)} \in \mathcal{K}$ by re-scaling $\mathbf{p}$ w/ SwoR % $K$ FLOPs
12: **Kaczmarz update step (Steps 12-17):**
13: $r_{i^{(t)}}^{(t)} = b_{i^{(t)}} - \mathbf{h}_{i^{(t)}}^H \mathbf{u}^{(t)} - \xi v_{i^{(t)}}^{(t)}$        % $8M + 4$ FLOPs
14: $\gamma^{(t)} = r_{i^{(t)}}^{(t)} / (||\mathbf{h}_{i^{(t)}}||_2^2 + \xi)$              % 2 FLOPs
15: $\mathbf{u}^{(t+1)} = \mathbf{u}^{(t)} + \gamma^{(t)} \mathbf{h}_{i^{(t)}}$              % $8M$ FLOPs
16: $\mathbf{v}^{(t+1)} = \mathbf{v}^{(t)} + \gamma^{(t)} [\mathbf{I}_K]_{:,i^{(t)}}$           % 2 FLOPs
17: $t = t + 1$
18: **end while**
19: $\hat{\mathbf{x}}^{\mathrm{RK}} = \mathbf{v}^{(t)}$, $T^{\mathrm{RK}} = t$
```

Algorithm 11 GRK-Based Receiver (GRK-RZF)

```
input: $\mathbf{H}$, $\mathbf{y}$, $M$, $K$, $\xi$
output: $\hat{\mathbf{x}}^{\text{GRK}}$, $T^{\text{GRK}}$
Repeat Steps 1-6 of Algorithm 10% $16KM - 2K - 1$ FLOPs
Store $(||\mathbf{H}||_F^2 + K\xi)^{-1}$     % 1 flop
While stopping criterion is Falsedo
    $\mathbf{r}^{(t)} \in \mathbb{C}^{K \times 1}$ w/     % $8KM - 8$ FLOPs
        $r_k^{(t)} = b_k - \mathbf{h}_k^H \mathbf{u}^{(t)} - \xi v_k^{(t)}$
    $\vec{\text{SAR}}^{(t)} \in \mathbb{R}^{K \times 1}$ with $\text{SAR}_k^{(t)} = |r_k^{(t)}|^2$   % $3K$ FLOPs
    $\text{RSS}^{(t)} = \sum_{k \in \mathcal{K}} \text{SAR}_k^{(t)}$    % $K - 1$ FLOPs
    Compute $\epsilon^{(t)}$ as   % $2K + 3$ FLOPs
        $\epsilon^{(t)} = \frac{1}{2} \left( \frac{1}{\text{RSS}^{(t)}} \max_{j \in \mathcal{K}} \left\{ \frac{\text{SAR}_j^{(t)}}{||\mathbf{h}_j||_2^2 + \xi} \right\} + \frac{1}{||\mathbf{H}||_F^2 + K\xi} \right)$
    Get the set of working equations:     % $K + 1$ FLOPs
        $\mathcal{U}_t = \left\{ k : \text{SAR}_k^{(t)} \geq \epsilon^{(t)} \text{RSS}^{(t)} \left( ||\mathbf{h}_k||_2^2 + \xi \right) \right\}$
    Probab. criterion based on complete residual info. $\mathbf{p}^{(t)} \in \mathbb{R}^{K \times 1}$:     % $K$ FLOPs
        $p_k^{(t)} = \left\{ \frac{\text{SAR}_k^{(t)}}{\sum_{j \in \mathcal{U}_t} \text{SAR}_j^{(t)}}, \text{ if } k \in \mathcal{U}_t \right\}.$     (F.7)
    pick $i^{(t)} \in \mathcal{U}_t$ based on $\mathbf{p}^{(t)}$
    Kaczmarz update step (Algo. 10 – Steps 10-14) % $8M + 4$ FLOPs
end while
$\hat{\mathbf{x}}^{\text{GRK}} = \mathbf{v}^{(t)}$, $T^{\text{GRK}} = t$
```

6. We refer to $\mathbf{r}^{(t)}$ as the *complete residual information* on an iteration basis. In Steps 5 and 6, we obtain the squared absolute residuals arranged in a vector $\vec{\mathrm{SAR}}^{(t)}$ and the residual sum of squares $\mathrm{RSS}^{(t)}$. Then, in Step 7, the quantity $\epsilon^{(t)}$ as a measure of the weighted average of the normalized squared absolute residuals is computed. Using this quantity, in Step 8, we can select a set $\mathcal{U}_t$ of *working equations* that corresponds to the equations with residuals larger than the weighted average. The idea is to discard equations with the lowest residuals prioritizing the equations (users) that are further away from being solved. In Step 9, the sampling probability vector $\mathbf{p}^{(t)}$ is now iteration dependent and calculated in (F.7) on the basis of the squared absolute residuals of the equations in $\mathcal{U}_t$. The subsequent steps follow the Kaczmarz update step. When the algorithm converges, we obtain the GRK-RZF soft estimate $\hat{\mathbf{x}}^{\mathrm{GRK}}$. A relaxed version of the GRK algorithm is presented in [25], where one can control the quantity $\epsilon^{(t)}$ and adjust the size of $\mathcal{U}_t$. However, we chose not to follow this method because the control of $\epsilon^{(t)}$ can generate unwanted complexity.

Probability criterion based on complete residual information

The GRK-RZF exploits the *complete residual information* as part of its probability criterion in (F.7) used to select the working equations (preferred users) from the SLE in (F.5). Evidently, this use can solve the two fundamental problems of the nRK-RZF scheme [10] of rare equations and curse of uniform normalization in a heuristic way because the residuals progress as solutions become better. The trend is that the residuals tend to zero as the number of iterations grows towards infinity. Therefore, it is expected that the number of necessary iterations for convergence of the GRK-RZF scheme is less than that of the RK-RZF. However, obtaining the complete residual information and its processing makes iteration more expensive. This indeed can lead to the case where the total complexity cost of the GRK-RZF receiver after convergence is greater than that of the RK-RZF and even of the RZF. This issue driven us to look for ways to further explore the performance gains brought by the residuals and reduce the related cost. We further elaborate the benefits brought by the probability criterion based on the residuals in Subsection 5.2.

The first way to decrease the complexity of the GRK-RZF scheme is to adopt the following recursive relationship [13]:

$$\begin{aligned} \mathbf{r}^{(t+1)} &= \mathbf{b} - \mathbf{H}^H \mathbf{u}^{(t+1)} - \xi \mathbf{v}^{(t+1)} \\ &\overset{(a)}{=} \mathbf{b} - \mathbf{H}^H (\mathbf{u}^{(t)} + \gamma^{(t)} \mathbf{h}_{i^{(t)}}) - \xi (\mathbf{v}^{(t)} + \gamma^{(t)} [\mathbf{I}_K]_{:,i^{(t)}}) \\ &= \mathbf{b} - \mathbf{H}^H \mathbf{u}^{(t)} - \xi \mathbf{v}^{(t)} - \gamma^{(t)} \mathbf{H}^H \mathbf{h}_{i^{(t)}} - \gamma^{(t)} \xi [\mathbf{I}_K]_{:,i^{(t)}} \\ &\overset{(b)}{=} \mathbf{r}^{(t)} - \gamma^{(t)} (\mathbf{H}^H \mathbf{h}_{i^{(t)}} + \xi [\mathbf{I}_K]_{:,i^{(t)}}) \\ &\overset{(c)}{=} \mathbf{r}^{(t)} - \gamma^{(t)} [\mathbf{R}_{\mathbf{yy}}]_{:,i^{(t)}}, \end{aligned} \tag{F.8}$$

where the following steps were applied: (a) the Kaczmarz iteration relationship (Algo. 11 – Step 11), (b) the definition of the residual vector $\mathbf{r}^{(t)} \in \mathbb{C}^{K \times 1}$ at iteration t (Algo. 11 – Step 4), and (c) the definition of $\mathbf{R}_{\mathbf{yy}}$ in (F.3). Henceforth, we assume that the GRK-RZF scheme adopts the above recursive updating of the complete residual information.

Algorithm 12 RSK-Based Receiver (RSK-RZF)

```
**input:** $\mathbf{H}$, $\mathbf{y}$, $M$, $K$, $\xi$, $\omega$
**Output:** $\hat{\mathbf{x}}^{\text{RSK}}$, $T^{\text{RSK}}$
Repeat Steps 1-6 of Algorithm 10% $16KM - 2K - 1$ FLOPs
Store $(||\mathbf{H}||_F^2 + K\xi)^{-1}$    % 1 flop
**While** stopping criterion **is** False **do**
Uniformly draw w/ SwoR $\mathcal{U}_t$ w/ $|\mathcal{U}_t| = \omega$
$\mathbf{r}^{(t)} \in \mathbb{C}^{K\times 1}$ with    % $\omega(8M + 4)$ FLOPs
    $$r_j^{(t)} = \left\{ b_j - \mathbf{h}_j^H \mathbf{u}^{(t)} - \xi v_j^{(t)}, \text{if } j \in \mathcal{U}_t \right\}.$$
Compute $\vec{\text{RR}}^{(t)} \in \mathbb{R}^{K\times 1}$ w/    % $4\omega$ FLOPs
    $$\text{RR}_k^{(t)} = |r_k^{(t)}|^2 / (||\mathbf{H}||_F^2 + K\xi) \, \forall k \in \mathcal{K}$$
$i^{(t)} = \arg\max_{j\in\mathcal{U}_t} \vec{\text{RR}}^{(t)}$    % $\omega$ FLOPs
Kaczmarz update step (Algo. 10 – Steps 10-14) % $8M + 4$ FLOPs
**end while**
$\hat{\mathbf{x}}^{\text{RSK}} = \mathbf{v}^{(t)}$, $T^{\text{RSK}} = t$
```

4.4 Randomized Sampling Kaczmarz Algorithm

To reduce the complexity related to the processing of residuals and still exploit part of this information, Algorithm 12 describes the RSK method proposed in [16] to solve the SLE in (F.5). We called this as the RSK-RZF scheme. Here, the iterative process starts by uniformly drawing equations to comprise the set $\mathcal{U}_t$ of working equations with a pre-defined size of ω. Subsequently, only the residuals of these ω equations are calculated, reducing the iteration cost in comparison with the GRK-RZF scheme. To select the equation at iteration t, a deterministic criterion is now adopted: the equation $k \in \mathcal{U}_t$ with the largest entry in the relative residual vector $\vec{\text{RR}}^{(t)}$ is chosen. Then, the algorithm follows the Kaczmarz update step.

Probability criterion based on partial residual information

Different from the other algorithms, randomization is used when constructing $\mathcal{U}_t$ and sorting $\vec{\text{RR}}^{(t)}$ and depends on a *partial residual information* coming from the ω equations selected at random. Because of this, the RSK-RZF scheme gives up some performance gains, since $\mathcal{U}_t$ may not have the equations (users) that have the largest residuals. Another implementation issue that arises with Algorithm 12 is how to select $|\mathcal{U}_t| = \omega$. Motivated by [16], we use $\omega = \lceil \log_2 K \rceil$.

5 RK-Based Receiver Designs for Multi-Antenna Systems

This section addresses two issues on the algorithms' operation described above when applied to the M-MIMO context: What is the meaning of a RK-based iteration for multi-antenna systems? What are the main benefits of using residual information in the equation selection probability criterion? A short answer to both of the questions is: we notice a similar operation of our schemes with the SIC and we find robustness against IUI and sparsity, respectively. We elaborate further on each of these issues in the sequel.

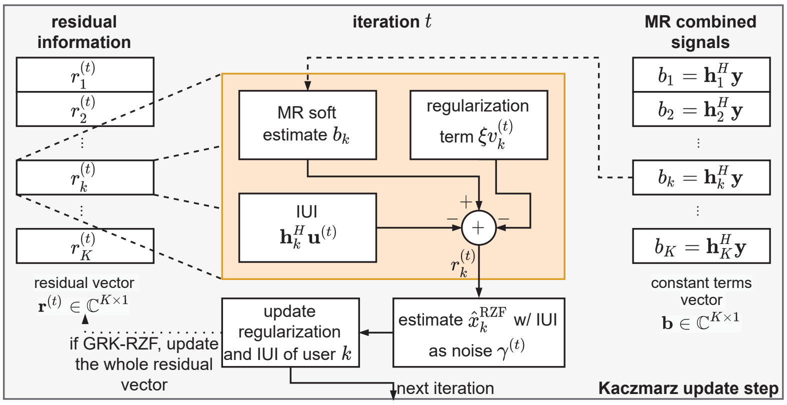

residual information
iteration t
MR combined signals
$r_1^{(t)}$
$r_2^{(t)}$
$r_k^{(t)}$
$r_K^{(t)}$
MR soft estimate b_k
regularization term $\xi v_k^{(t)}$
IUI $\mathbf{h}_k^H \mathbf{u}^{(t)}$
$+$
$-$
$-$
$r_k^{(t)}$
$b_1 = \mathbf{h}_1^H \mathbf{y}$
$b_2 = \mathbf{h}_2^H \mathbf{y}$
$b_k = \mathbf{h}_k^H \mathbf{y}$
$b_K = \mathbf{h}_K^H \mathbf{y}$
residual vector $\mathbf{r}^{(t)} \in \mathbb{C}^{K\times 1}$
constant terms vector $\mathbf{b} \in \mathbb{C}^{K\times 1}$
if GRK-RZF, update the whole residual vector
update regularization and IUI of user k
estimate $\hat{x}_k^{\mathrm{RZF}}$ w/ IUI as noise $\gamma^{(t)}$
next iteration
Kaczmarz update step

Fig. F.2: Illustration of the Kaczmarz update step when user k is selected at iteration t. Observe that the residual $r_k^{(t)}$ stores the MR combined signal of user k less: (i) IUI from previous iterations and (ii) the regularization term applied to combat noise.

Table F.2: Computational Complexity Comparison

Scheme	Computational Complexity [FLOPs]	M-MIMO [FLOPs] ($M = 64, K = 8, T = 12$)†	XL-MIMO [FLOPs] ($M = 256, K = 32, T = 64$)†
MR	$8KM - 2K$	4080	65472
RZF	$4K^2M + 12KM + 5K^3 + 10K^2 - 4K$	25696	1320832
nRK-RZF [10]	$16KM - K - 1 + (16M + 8)T^{\mathrm{nRK}}$	20567	393695
RK-RZF (Algorithm 10)	$16KM - 2K - 1 + (K + 16M + 8)T^{\mathrm{RK}}$	20653	395711
GRK-RZF (Algorithm 11)	$4K^2M + 12KM - K^2 - K + (16K + 8M + 7)T^{\mathrm{GRK}}$	30220	1310112
RSK-RZF (Algorithm 12)	$16KM - 2K + [\omega(8M + 9) + 8M + 4]T^{\mathrm{RSK}}$	33124	920576
TPE-RZF [5, 6]	$4K^2M + 12KM + 3K + 4 + (8K^2 + 4K)T^{\mathrm{TPE}}$	29696	1679460

$^\dagger T = T^{\mathrm{nRK}} = T^{\mathrm{RSK}} = T^{\mathrm{GRK}} = T^{\mathrm{RSK}} = T^{\mathrm{TPE}}$; T^s denotes the number of iterations of scheme $s \in \{\mathrm{nRK}, \mathrm{RSK}, \mathrm{GRK}, \mathrm{RSK}, \mathrm{TPE}\}$.

5.1 A Similarity with SIC receiver

We show that the structure of the RK algorithms applied to solve the SLE in (F.5) yields a mechanism similar to the used by the SIC receiver. For that, Fig. F.2 illustrates in details the Kaczmarz update step that is common to all algorithms. In this figure, we assume that user $k \in \mathcal{K}$ is selected at a given iteration t. In addition, the residual vector $\mathbf{r}^{(t)}$ and the vector of constant terms $\mathbf{b}$ with the MR soft estimate are represented by stacks of blocks. An interesting pattern can be observed from the figure. The residual $r_k^{(t)}$ is storing the MR combined signal of user k less: (i) the IUI from previous iterations and (ii) the regularization term applied to combat noise. Then, $r_k^{(t)}$ is used to get a new estimate of $\hat{x}_k^{\text{RZF}}$ considering remaining IUI as noise. We notice that this mechanism is similar to that used by the SIC receiver [26], which successively remove the contribution of the decoded data from the received signal on an iteration basis. Mathematically, we can obtain this interpretation of the residual using the recursive relationship in (F.8). From this, we can write

$$\begin{aligned} r_k^{(t)} &= \mathbf{h}_k^H \mathbf{y} - \sum_{t'=1}^{t} \gamma^{(t')} [\mathbf{R}_{\mathbf{yy}}]_{k,i^{(t')}} \\ &= \mathbf{h}_k^H \mathbf{y} - \sum_{t'=1}^{t} \gamma^{(t')} \{ \underbrace{\chi_{\mathcal{I}}(i^{(t')}) \mathbf{h}_k^H \mathbf{h}_{i^{(t')}}}_{\text{IUI}} + \\ &+ [1 - \chi_{\mathcal{I}}(i^{(t')})] \underbrace{[||\mathbf{h}_k||_2^2 + \xi]}_{\text{self-knowl. + reg.}} \} \end{aligned} \tag{F.9}$$

where $\chi_{\mathcal{I}}(i^{(t')})$ is the indicator function with argument denoting the equation selected at iteration t' and $\mathcal{I} = \mathcal{K} \setminus \{k\}$ is the set of interfering users in relation to user k. In the expression above, we used the fact that $r_k^{(0)} = b_k$ since $\mathbf{u}^{(0)}$ and $\mathbf{v}^{(0)}$ are initialized with zeros. Moreover, it can be seen that: if user k was selected at previous iterations, the previous estimates of $\hat{x}_k^{\text{RZF}}$ is also removed from $r_k^{(t)}$ together with the noise penalization (self-knowledge + regularization in (F.9)). Recall that the RK-based algorithms are approximating the RZF scheme when solving the SLE in (F.5). The application of the RK algorithms over (F.5) is then transforming the RZF scheme into a SIC-alike receiver. Effectively, we are refining the MR soft estimate $\hat{\mathbf{x}}^{\text{MR}}$ stored in $\mathbf{b}$ to approach the RZF soft estimate $\hat{\mathbf{x}}^{\text{RZF}}$, placing the SIC-alike iterations in charge of computing the weights of IUI suppression based on the RZF criterion. This adaptive mechanism implicitly helps supporting more users in multi-antenna systems with low-complexity. In addition, the RK-based receivers do no explicitly calculate metrics, such as post-processing SNR or signal-to-interference-plus-noise ratio (SINR), that are normally used to order the classical SIC receiver [26]. In fact, the RK-based algorithms give preference to the equations (users) based on the two types of information promptly provided by the SLE in eq. (F.5): the energy information of the equations in (F.6) and the complete residual information in (F.7), which is partial for the RSK-RZF. We discuss the advantages of the latter below.

5.2 Probability Criterion and Residual Information

The probability criterion in (F.7) uses the complete residual information $\mathbf{r}^{(t)}$ to select the equations of (F.5) to be solved. The residual contains inner products between channel vectors; hence, eq. (F.9) describes the IUI and naturally introduces the sparse structure arising from the spatial non-stationarities. This is a relevant contrast compared to the probability criterion in (F.6) that has probabilities proportional to $||\mathbf{h}_k||_2^2 + \xi, \forall k \in \mathcal{K}$ only. As a result, the probability criterion of (F.7) can better capture the interaction effects of IUI and sparsity existing in the XL-MIMO channels, outperforming (F.6) under the occurrence of theses effects. Also, (F.7) is dynamic in the sense that the IUI terms and consequently the probabilities are updated in the background according to the SIC-alike IUI suppression mechanism running in the foreground, changing the set $\mathcal{U}_t$ of preferred users as the solution evolves. The RSK-RZF scheme partly features these gains. However, a clear issue of using residual information is the cost of such more involved probability criterion, as best evidenced in the next sections.

6 Complexity Analysis and Numerical Results

In this section, we first discuss the computational complexity of the proposed RK-based RZF receivers in terms of floating-point operations per second (FLOPs) and how their stopping criterion can be defined in practice. The performance of the proposed receivers is numerically evaluated in the sequel by taking the bit-error-rate (BER) as a metric. Further, we assume that $\mathbf{x}$ is drawn from the equiprobable 16-QAM constellation. Besides MR, RZF, and nRK-RZF [10], in the numerical simulations we compare our proposed schemes to the TPE-RZF receiver of [5, 6]. The choice for the TPE-RZF is due to the fact that this receiver is a consolidated approach in the literature that also iteratively approximates the RZF scheme. Finally, for tractability reasons, we further consider horizontal uniform linear array (ULA) arrangements under non-line-of-sight conditions in both M-MIMO and XL-MIMO regimes, with the distance between any two neighboring antennas greater than half a wavelength when considering sub-6 GHz transmissions.

6.1 Complexity Analysis

The second column of Table F.2 summarizes the total number of FLOPs needed to compute the receiver designs relevant for this work. We always account for the worst-case when performing the complexity analysis of our proposed schemes. For Algorithm 10, this means that the cost of Step 11 is considered to be K FLOPs at most due to the re-normalization of $\mathbf{p}$. For Algorithm 11, Step 11 costs K FLOPs, since $|\mathcal{U}_t|$ is always considered to be K. Complexity of the MR and RZF schemes follows [3], while the complexity of the TPE-RZF receiver is discussed in detail in [5] and [6]. For the TPE-RZF, we adopt the eigenvalue estimation of $\mathbf{R}_{\mathbf{yy}}$ proposed in [5]. For the sake of fairness, we assume that the BS does not know second-order channel statistics. Thus, the scaling proposed in [5] to lower the scattering of the eigenvalue is not performed. This allows us to show that our methods are more robust to channel gain variations between users without the need to seek alternatives to reduce the impact of these variations.

To give a better notion of how the computational complexities are compared to each other, we evaluate two typical scenarios of both M-MIMO and XL-MIMO regimes in Table F.2. Note that we set the same number of iterations T for all the iterative algorithms in the table, where $T = 12$ and $T = 64$ iterations for M-MIMO and XL-MIMO, respectively. We chose these numbers because they allow us to show the computational gains brought by our receiver designs, while achieving good performance. We notice the hereafter trends from the table: **a**) for M-MIMO, the GRK-RZF scheme is unable to relax the RZF, exemplifying the high cost of complete residual information; **b**) in contrast, the RK-RZF receiver can relax complexity of the RZF in 19.62%. For XL-MIMO: **c**) the GRK-RZF scheme is up to reduce the complexity of the RZF in 0.81%, while the RK-RZF achieves a relaxation of 70.19%; **d**) the RSK-RZF scheme only achieves its goal of relaxing the GRK-RZF in the XL-MIMO scenario; **e**) the TPE-RZF receiver has iterations independent of M, but it has a high fixed cost due to the exact computation of $\mathbf{R}_{\mathbf{yy}}$ and the estimation of its eigenvalues. From these observations, we noticed that as M and K increase, the relaxation capacity of the GRK-RZF receiver is improved. However, the RK-RZF will always have a greater ability to relax complexity, since its iterations are cheap.

Next, we evaluate the difference in performance among the proposed receivers, identifying when the use of residual information becomes justified.

Sparsity

In the case of a sparse SLE in (F.5), we can automatically reduce the costs of the iterations of the RK-based receivers. The reason for this is to note that inner products are the most cost operations in the iterations of all the algorithms, which can be evidently reduced by only using the non-zero entries of the vectors.

Defining in practice the number of iterations

The most convenient stopping criterion of all the proposed algorithms is the maximum number of iterations. The BS can regularly adjust the maximum number of iterations after a constant period of time-frequency resources that spans multiple coherence blocks. This adjustment can be based on some performance or complexity metric that the BS wants to achieve.

6.2 Stationary Case: M-MIMO

Consider a cell that covers a square area of 0.4 km × 0.4 km served by a BS with $M = 64$ compactly installed antennas located at the cell center. The users are uniformly distributed in the cell area at locations further than 35 m from the BS. Furthermore, for this scenario we assume that all the elements of $\boldsymbol{\beta}_k$ are equal, since the distance between antennas is much smaller than the distance between users and the antenna array. Then, we model the pathloss based on the urban micro scenario as [3]: $\beta_k = -30.5 - 36.7 \log_{10} d_k$ in dB, where $\beta_k = \frac{1}{M} \text{tr} \mathbf{R}_k$ is the average large-scale coefficient, d_k is the distance in meters between user $k \in \mathcal{K}$ and the BS. In addition, we consider the more general exponential correlation model in which $[\mathbf{R}_k]_{i,j} = \iota^{|i-j|}$, $\forall i, j \in \mathcal{M}$ [27], where ι is the antenna correlation coefficient.

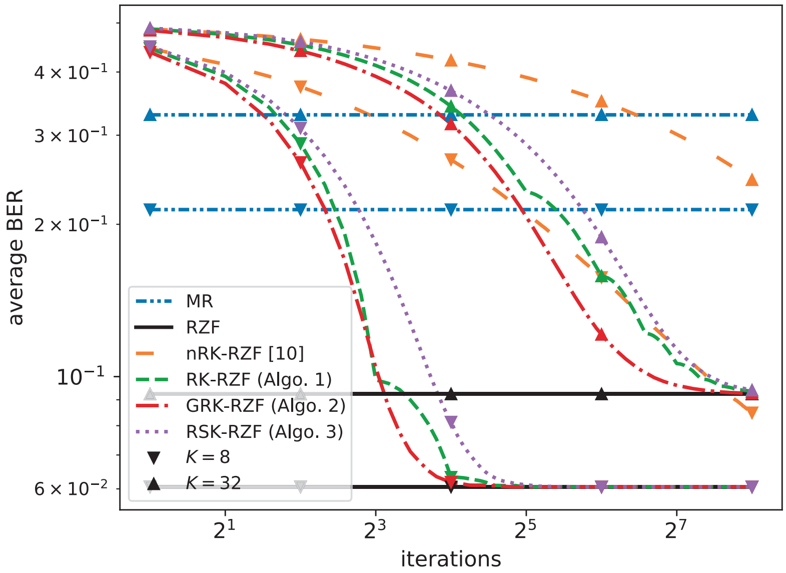
We first evaluate the convergence in terms of both performance (Fig. F.3(a)) and complexity (Fig. F.3(b)) of the proposed algorithms. A typical load of $K = 8$ users and a crowded scenario with $K = 32$ users are examined. The first observation from Fig. F.3(a) is that the performance of our three accelerated RK-based RZF schemes is much better than that obtained with the nRK-RZF proposed in [10]. We also notice that increasing the IUI with increasing K harms the convergence of RK-based schemes in general. However, the GRK-RZF suffers the least from increased IUI. This result is inline with the observed fact that randomization based on the residual information is more robust against IUI. However, we learn from Fig. F.3(b) that the GRK-RZF receiver is not suitable for typical M-MIMO scenarios in terms of complexity and only starts to get more appealing in this regime as K increases. Moreover, the bouncy behavior of the performance curves associated to RK-RZF in Fig. F.3(a) is explained by the SwoR technique and the stochastic behavior of the elements of the set $\mathcal{P}^{(t)}$. Note that the start and end point of the bounce comprehends the definition of a sweep made in Subsection 4.2 that embraces K iterations.

Fig. F.4 depicts the BER performance of different receivers as a function of the pre-processing SNR ρ/σ^2 with $K = 8$ users for uncorrelated and correlated ($\iota = 0.5$) Rayleigh fading conditions. It considers a number of iterations fixed in 12 for all the iterative schemes. The final complexity of each scheme follows the third column in Table F.2. Among the RK-based RZF schemes, the GRK-RZF attains the best performance, but needs more FLOPs than the RZF scheme. In contrast, the RK-RZF performs well as a whole while relaxing the complexity of the RZF in 19.62%. In general, the accelerated RK-based RZF schemes better approximate the performance of the RZF at low pre-processing SNRs. This is because the strength of IUI is amplified when operating at high SNRs. Finally, our schemes perform better than the TPE-RZF scheme [5, 6] with less complexity.

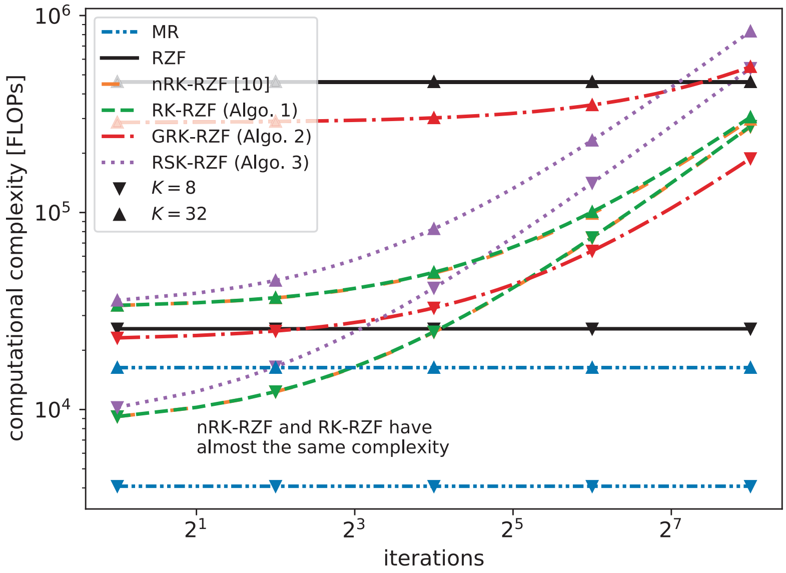
6.3 Non-Stationary Case: XL-MIMO

Let's consider a square cell with an area of 0.25 km × 0.25 km that has totally occupying one of its side by a ULA equipped with $M = 256$ antennas. The users are uniformly distributed in the cell keeping a minimum distance of 25 m from the array.[5] The distance between the user and antenna elements is now relevant. Therefore, each element β_k^m of $\boldsymbol{\beta}_k$ is modeled as $\beta_k^m = -30.5 - 36.7 \log_{10} d_k^m$, where d_k^m is the distance between user $k \in \mathcal{K}$ and antenna $m \in \mathcal{M}$. Under this setting, we focus on non-stationarities and consider antenna correlation irrelevant, then $\mathbf{R}_k = \mathbf{I}_M$. Moreover, we generate $\mathbf{D}_k$ for user $k \in \mathcal{K}$ as follows [4]: **a**) we choose an arbitrary antenna $m \in \mathcal{M}$ uniformly at random to be the center c_k of the VR; **b**) if D is odd, the VR of user k is $\mathcal{V}_k = \{c_k - \lfloor D/2 \rfloor, \ldots, c_k + \lfloor D/2 \rfloor\}$, otherwise $\mathcal{V}_k = \{c_k - \lfloor D/2 \rfloor, \ldots, c_k + \lfloor D/2 \rfloor + 1\}$; **c**) we set $[\mathbf{D}_k]_{m,m} = 1$, if $m \in \mathcal{V}_k \cap \mathcal{M}$ and $[\mathbf{D}_k]_{m,m} = 0$ otherwise, and **d**) we normalize $\mathbf{D}_k$ by M/D, hence stationary and non-

[5] The choice for these values and the geometry is motivated by the fact that the users are close enough to the array to justify the emergence of spatial non-stationarities [17, 26]. It is noteworthy that the adopted geometry of M-MIMO and XL-MIMO are comparable given that in M-MIMO, the BS is cell-centered; while in XL- MIMO geometry, the BS comprises one of the edges of the square area.



(a) Average BER versus iterations.



(b) Computational complexity in FLOPs versus iterations.

Fig. F.3: Convergence of the RK-based RZF schemes in a normally loaded ($K = 8$) and crowded ($K = 32$) M-MIMO system with $M = 64$ antennas, under Rayleigh fading channels, and a pre-processing SNR of 0 dB.

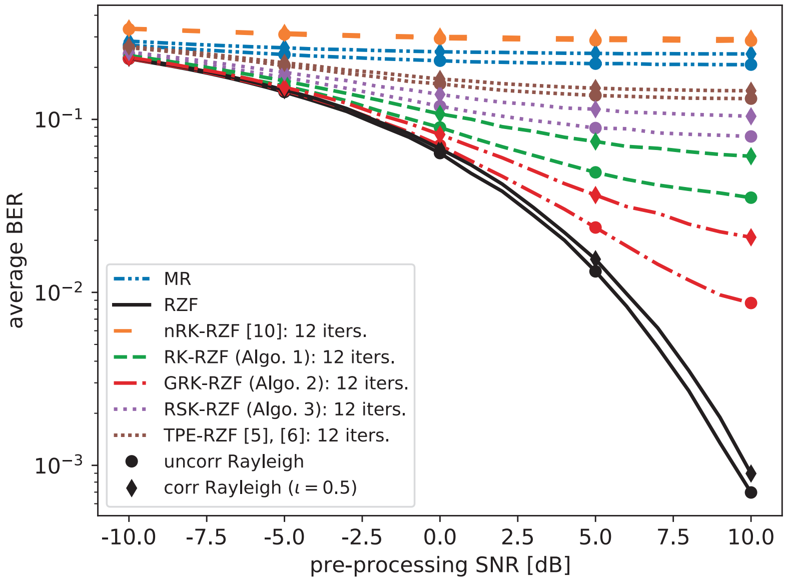
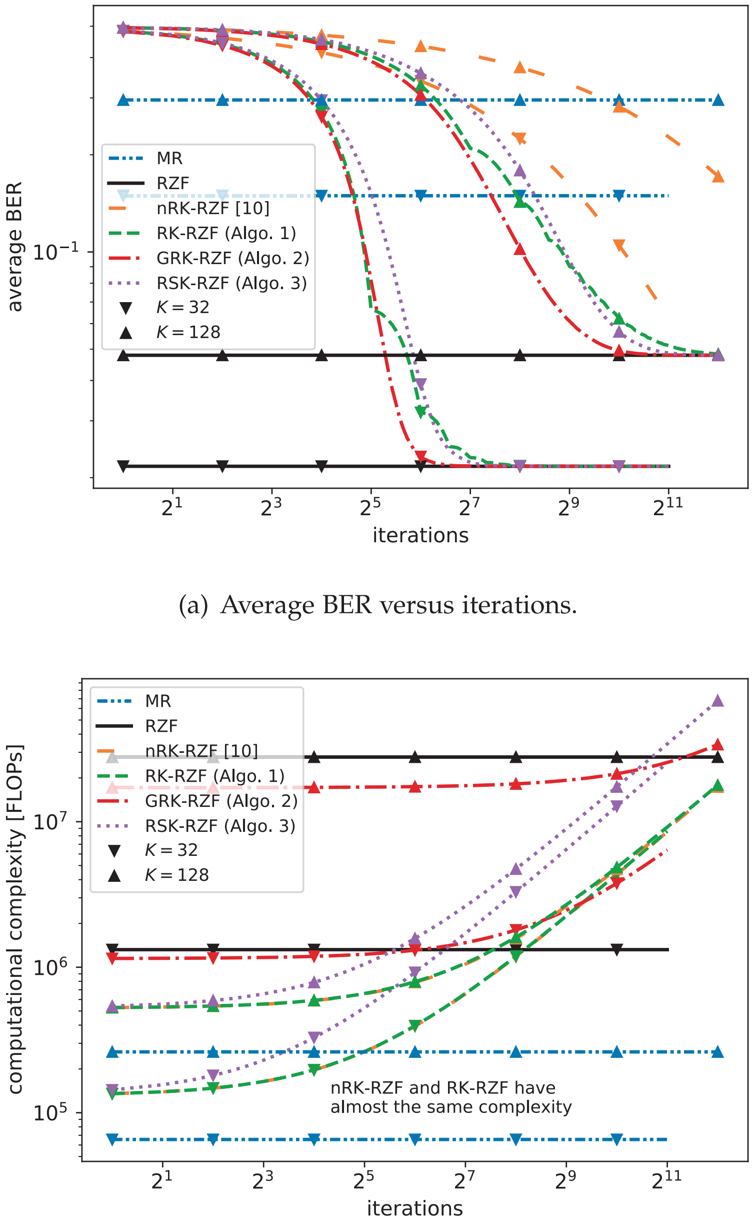


Fig. F.4: Average BER performance of various receivers *vs.* the pre-processing SNR under uncorrelated and correlated ($\iota = 0.5$) Rayleigh channels for the M-MIMO system equipped with $M = 64$ antennas and $K = 8$ users.

stationary channels have the same norm [4]. We stress that this normalization is giving an array gain for non-stationary channels similar to the stationary channels, allowing a fair comparison between the two array regimes. With this model, the users have a unique cluster of antennas representing their VRs with an average size of D. To evaluate how the receivers behave under very extreme sparse conditions, in the following, we set low values for D. In [4], for example, the authors report problems with the ZF scheme from D below 30 visible antennas. Here, the RZF regularization term makes it possible to perform the matrix inversion even in these severe conditions.

We first take a look at Fig. F.5 that exhibits the convergence of the RK-based RZF schemes and how it impacts performance (Fig. F.5(a)) and complexity (Fig. F.5(b)) with $D = 8$. Again, a typical load of $K = 32$ users and a crowded scenario with $K = 128$ users are evaluated. Comparing Fig. F.3(a) and Fig. F.5(a), one can see the bad effects of IUI over the convergence of the algorithms and how it impacts more severely the RK-RZF scheme. Furthermore, we now find from Fig. F.5(b) that the GRK-RZF scheme has more room to be able to relax the RZF scheme. One of the reasons is that the values of M and K become higher, justifying the cost related to residual information. Another is that the RK-RZF scheme needs more iterations to achieve a better performance under high levels of IUI and sparsity.

Similar to Fig. F.4, a performance comparison of the receivers is available in Fig. F.6 with $K = 32$ users and under $D = 8$ and $D = 16$ visible antennas. In addition, we fix the number of iterations of all the iterative schemes to 64. The final complexities of every scheme is reported in the fourth column of Table F.2. Definitely, we note that the performance difference between the RK-RZF and the GRK-RZF schemes increases in Fig. F.6 in comparison to Fig. F.4. This indicates that besides being more robust against IUI, the GRK-RZF scheme is more robust against the sparse structure arising from the spatial non-stationarities. Interestingly, the GRK-RZF receiver performs bet-



(a) Average BER versus iterations.

(b) Computational complexity in FLOPs versus iterations.

Fig. F.5: Convergence of the different RK-based RZF schemes in a normally loaded ($K = 32$) and crowded ($K = 128$) XL-MIMO system with $M = 256$ antennas, under Rayleigh fading channels, a pre-processing SNR of 0 dB, and a sparsity level of $D = 8$ visible antennas.

ter than the RK-RZF in high SNR regime, while relax the RZF in 0.81%; besides, the RK-RZF gives up of performance at high SNR and achieve a relaxation of 70.19%. The TPE-RZF scheme [5, 6] notably experiences a strong performance degradation from the sparse channels, corroborating a greater robustness of our proposed methods.

From the above results, we can observe that the GRK-RZF scheme becomes more convenient when the scenario is more crowded (↑ K), sparsity effects are more intense (↓ D), and/or the system is operating at high SNR. On the other hand, the RK-RZF scheme is the most appropriate choice when more relaxation is desired, the performance losses are tolerable at high SNR, and/or the system is operating at low SNRs. Although GRK-RZF operates better at high SNR, its gains in complexity compared to the RZF scheme may be only *marginal* depending on the values of M and K; when this is the case, the use of RZF may then be more advisable. The RSK-RZF receiver can achieve its goal of relaxing the GRK-RZF in some regions, at the cost of reduced performance; *e.g.*, in the range of $2^7 - 2^9$ iterations for $K = 128$ in Fig. F.5.

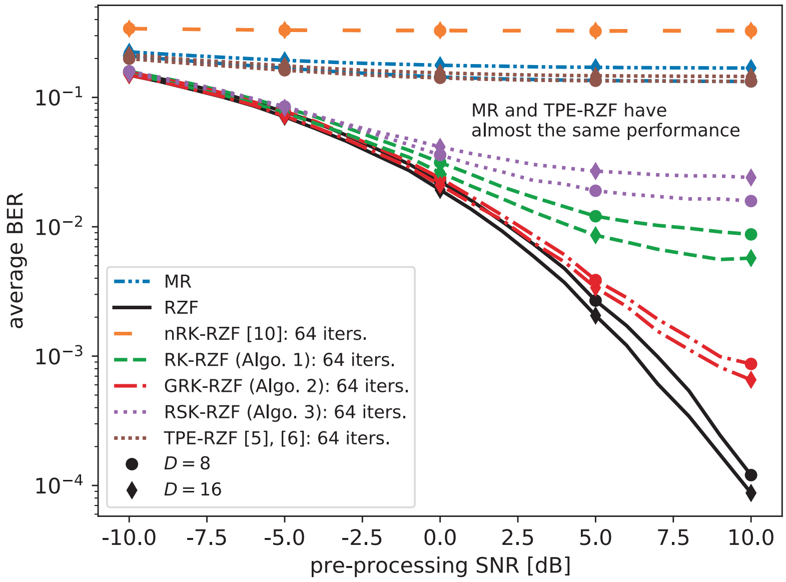


Fig. F.6: Average BER performance of various receivers *vs.* the pre-processing SNR under uncorrelated Rayleigh channels for the XL-MIMO system equipped with $M = 256$ antennas and $K = 32$ users; two distinct sparsity levels, $D = 8$, and $D = 16$ visible antennas.

7 Conclusions

We introduced three accelerated RK-based receivers, which approximate the performance of the RZF scheme, while relaxing its complexity. In our experiments, all of our proposed schemes are able to dramatically overcome the nRK-RZF introduced in [10]. The main feature of each scheme is in order. The RK-RZF (Algo. 10) is the proposed receiver with the best benefit-cost ratio, performing well in typical M-MIMO and XL-MIMO circumstances, while relaxing the complexity of the RZF in almost 20% and 70%, respectively. Moreover, the GRK-RZF scheme (Algo. 11) is more suitable for extreme cases where IUI and sparsity effects cannot be neglected. The RSK-RZF receiver (Algo. 12) can be more efficient than GRK-RZF in some scenarios but suffers from performance losses. Future work can go deeper into the theoretical analysis of the introduced receivers by using the analogy with SIC receivers revealed herein.

References

[1] E. Björnson, L. Sanguinetti, H. Wymeersch, J. Hoydis, and T. L. Marzetta, "Massive MIMO is a reality—What is next?: Five promising research directions for antenna arrays," *Digital Signal Processing: A Review Journal*, vol. 94, pp. 3–20, 2019.

[2] E. D. Carvalho, A. Ali, A. Amiri, M. Angjelichinoski, and R. W. Heath, "Non-stationarities in extra-large-scale massive MIMO," *IEEE Wireless Communications*, vol. 27, no. 4, pp. 74–80, Aug. 2020.

[3] E. Björnson, J. Hoydis, and L. Sanguinetti, "Massive MIMO networks: Spectral, energy, and hardware efficiency," *Foundations and Trends® in Signal Processing*, vol. 11, no. 3-4, pp. 154–655, 2017.

[4] A. Ali, E. D. Carvalho, and R. W. Heath, "Linear receivers in non-stationary massive MIMO channels with visibility regions," *IEEE Wireless Communications Letters*, vol. 8, no. 3, pp. 885–888, Jun. 2019.

[5] G. M. A. Sessler and F. K. Jondral, "Low complexity polynomial expansion multiuser detector for CDMA systems," *IEEE Transactions on Vehicular Technology*, vol. 54, no. 4, pp. 1379–1391, 2005.

[6] A. Kammoun, A. Müller, E. Björnson, and M. Debbah, "Linear precoding based on polynomial expansion: Large-scale multi-cell MIMO systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 861–875, 2014.

[7] B. Yin, M. Wu, J. R. Cavallaro, and C. Studer, "Conjugate gradient-based soft-output detection and precoding in massive MIMO systems," in *2014 IEEE Global Communications Conference*. IEEE, Dec. 2014, pp. 3696–3701.

[8] X. Gao, L. Dai, C. Yuen, and Y. Zhang, "Low-complexity MMSE signal detection based on Richardson method for large-scale MIMO systems," in *2014 IEEE 80th Vehicular Technology Conference (VTC2014-Fall)*. IEEE, Sep. 2014, pp. 1–5.

[9] L. Dai, X. Gao, X. Su, S. Han, C.-L. I, and Z. Wang, "Low-complexity soft-output signal detection based on Gauss–Seidel method for uplink multiuser large-scale MIMO systems," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 10, pp. 4839–4845, Oct. 2015.

[10] M. N. Boroujerdi, S. Haghighatshoar, and G. Caire, "Low-complexity statistically robust precoder/detector computation for massive MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 17, no. 10, pp. 6516–6530, Oct. 2018.

[11] S. Kaczmarz, "Angenäherte auflösung von systemen linearer gleichungen," *Bulletin International de l'Académie Polonaise des Sciences et des Lettres. Classe des Sciences Mathématiques et Naturelles. Série A, Sciences Mathématiques*, vol. 35, pp. 355–357, 1937.

[12] T. Strohmer and R. Vershynin, “A randomized Kaczmarz algorithm with exponential convergence,” *Journal of Fourier Analysis and Applications*, vol. 15, no. 2, pp. 262–278, Apr. 2009.

[13] Z. Z. Bai and W. T. Wu, “On greedy randomized Kaczmarz method for solving large sparse linear systems,” *SIAM Journal on Scientific Computing*, vol. 40, no. 1, pp. A592–A606, 2018.

[14] Y. Censor, G. T. Herman, and M. Jiang, “A note on the behavior of the randomized Kaczmarz algorithm of Strohmer and Vershynin,” *Journal of Fourier Analysis and Applications*, vol. 15, no. 4, pp. 431–436, 2009.

[15] T. Strohmer and R. Vershynin, “Comments on the randomized Kaczmarz method,” *Journal of Fourier Analysis and Applications*, vol. 15, no. 4, pp. 437–440, 2009.

[16] M.-L. Sun, C.-Q. Gu, and P.-F. Tang, “On randomized sampling Kaczmarz method with application in compressed sensing,” *Mathematical Problems in Engineering*, vol. 2020, pp. 1–11, Mar. 2020.

[17] A. Amiri, S. Rezaie, C. N. Manchon, and E. de Carvalho, “Distributed receivers for extra-large scale MIMO arrays: A message passing approach,” *arXiv*, pp. 1–31, 2020.

[18] H. Wang, A. Kosasih, C.-K. Wen, S. Jin, and W. Hardjawana, “Expectation propagation detector for extra-large scale massive MIMO,” *IEEE Transactions on Wireless Communications*, vol. 19, no. 3, p. 2036–2051, Mar. 2020. [Online]. Available: `http://dx.doi.org/10.1109/TWC.2019.2961892`

[19] V. C. Rodrigues, A. Amiri, T. Abrao, E. de Carvalho, and P. Popovski, “Low-complexity distributed XL-MIMO for multiuser detection,” in *2020 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, Jun. 2020, pp. 1–6.

[20] C. D. Meyer, *Matrix analysis and applied linear algebra*. SIAM, 2000.

[21] Wu, Hebiao and Shen, Bin and Zhao, Shufeng and Gong, Peng, “Low-complexity soft-output signal detection based on improved Kaczmarz iteration algorithm for Uplink massive MIMO system,” *Sensors*, vol. 20, no. 6, p. 1564, Mar. 2020.

[22] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical recipes: The art of scientific computing*, 3rd ed. Cambridge, UK: Cambridge University Press, 2007.

[23] Z. Z. Bai and W. T. Wu, “On partially randomized extended Kaczmarz method for solving large sparse overdetermined inconsistent linear systems,” *Linear Algebra and Its Applications*, vol. 578, pp. 225–250, 2019.

[24] V. C. Rodrigues, J. C. Marinello Filho, and T. Abrão, “Randomized Kaczmarz algorithm for Massive MIMO systems with channel estimation and spatial correlation,” *International Journal of Communication Systems*, p. e4158, Sep. 2019.

[25] Z. Z. Bai and W. T. Wu, “On relaxed greedy randomized Kaczmarz methods for solving large sparse linear systems,” *Applied Mathematics Letters*, vol. 83, pp. 21–26, 2018.

[26] T. Brown, E. De Carvalho, and P. Kyritsi, *Practical guide to the MIMO radio channel with MATLAB examples*, 1st ed. Wiley, 2012.

[27] E. Bjornson, J. Hoydis, and L. Sanguinetti, “Massive MIMO has unlimited capacity,” *IEEE Transactions on Wireless Communications*, vol. 17, no. 1, pp. 574–590, Jan. 2018.

Paper G

Antenna Selection for Improving Energy Efficiency in XL-MIMO Systems

José Carlos Marinello, Taufik Abrão, Abolfazl Amiri, Elisabeth de Carvalho, Petar Popovski

The paper has been published in the
IEEE Transactions on Vehicular Technology (TVT), vol. 66, no. 11,
pp.13305–13318, 2020.

The layout has been revised.

Abstract

We consider the recently proposed extra-large scale massive multiple-input multiple-output (XL-MIMO) systems, with some hundreds of antennas serving a smaller number of users. Since the array length is of the same order as the distance to the users, the long-term fading coefficients of a given user vary with the different antennas at the base station (BS). Thus, the signal transmitted by some antennas might reach the user with much more power than that transmitted by some others. From a green perspective, it is not effective to simultaneously activate hundreds or even thousands of antennas, since the power-hungry radio frequency (RF) chains of the active antennas increase significantly the total energy consumption. Besides, a larger number of selected antennas increases the power required by linear processing, such as precoding matrix computation, and short-term channel estimation. In this paper, we propose four antenna selection (AS) approaches to be deployed in XL-MIMO systems aiming at maximizing the total energy efficiency (EE). Besides, employing some simplifying assumptions, we derive a closed-form analytical expression for the EE of the XL-MIMO system, and propose a straightforward iterative method to determine the optimal number of selected antennas able to maximize it. The proposed AS schemes are based solely on long-term fading parameters, thus, the selected antennas set remains valid for a relatively large time/frequency intervals. Comparing the results, we find that the genetic-algorithm based AS scheme usually achieves the best EE performance, although our proposed highest normalized received power AS scheme also achieves very promising EE performance in a simple and straightforward way.

Keywords— Extra large-scale MIMO, Antenna selection, Energy efficiency, Spectral efficiency,; Visibility region (VR), Non-stationary, Near-field

1 Introduction

In the fifth-generation (5G) networks, massive multiple-input multiple-output (MIMO) is identified as a key technology for achieving large gains in spectral and energy efficiencies [1, 2]. Recently, a new type of very large antenna arrays, which can be integrated into large structures like stadiums, or shopping malls, has been conceived: the so called extra-large scale massive MIMO (XL-MIMO) [3–5]. XL-MIMO system is a very promising and recent technology, pointed out as important candidate for sixth-generation (6G) and beyond technologies [6, 7], which is still in its inception, lacking for further elaborated techniques in order to mature the technology. Indeed, due to the large dimension of the antenna array in XL-MIMO systems, different kinds of spatial non-stationarities appear accross the array [3–5]; hence, admitting constant long-term fading coefficients between a user and all the antennas of the array is not a valid assumption. This is the main difference between the XL-MIMO scenario and the typical massive MIMO system model assumed in most part of massive MIMO literature. In [8], it is shown through experimental measurements how different regions of an extremely large array see different propagation paths, and in some cases, the terminals might see just a portion of the array, called visibility region (VR). Authors also discuss how the non-stationarity properties of this new scenario change several important design aspects.

In [3] authors seek for mapping users in terms of XL-MIMO array partition, such that the downlink (DL) sum-rate using a truncated zero-forcing (ZF) precoder is maximized. Numerical results show that a properly trained network via deep learning approach solves the problem nearly as well as an optimal mapping algorithm. Hence, increasing the size of current massive MIMO arrays is promising in terms of boosting the spectral efficiency (SE) of the wireless systems.

Since the centralized processing may present very high computational complexity in XL-MIMO arrays, a useful approach is to split the signal processing between subarrays. A subarray-based system architecture for XL-MIMO systems is proposed in [4], where closed-form uplink (UL) SE approximations with linear receivers are derived; the goal is to maximize the system sum achievable SE. Two statistical channel state information (CSI) based greedy user scheduling algorithms are developed, providing improved performance for XL-MIMO systems.

In [5], a simple *non-stationary channel model* is proposed for XL-MIMO systems, and the performance of conjugate beamforming (CB) and ZF in the DL have been investigated considering such channel. The non-stationarities are modeled in a binary fashion, such that each antenna can be visible or not for a specific user, giving rise to the VRs: an area of the massive antenna array concentrating the most of the received user's energy. However, the authors did not consider long-term fading variations between the visible antennas of a given user.

In [9] authors develop procedures for XL-MIMO receivers design. There are two important challenges in designing receivers for XL-MIMO systems: increased computational cost of the multi-antenna processing, and how to deal with the variations of user energy distribution over the antenna elements due to the spatial non-stationarities across huge distributed antenna-elements in the 2D or 3D array. Indeed, non-stationarities limit the XL-MIMO system performance. Hence, the authors propose a distributed receiver based on variational message passing that can address both challenges. In the proposed receiver structures, the processing is distributed into local processing units, that can perform most of the complex processing in parallel, before sharing their outcome with a central processing unit. Such designs are specifically tailored to exploit the spatial non-stationarities and require lower computations than linear ZF or minimum mean square error (MMSE) receivers.

In [10], the ZF and regularized ZF schemes operating in XL-MIMO scenarios with a fixed number of subarrays have been emulated using the randomized Kaczmarz algorithm (rKA), deploying non-stationary properties through VRs. Numerical results have shown that, in general, the proposed rKA-based combiner applicable to XL-MIMO systems can considerably decrease computational complexity of the signal detector at the expense of small performance losses. On the other hand, in [11], an expectation propagation detector for XL-MIMO systems has been proposed. In order to reduce complexity, the subarray-based architecture employed distributes baseband data from disjoint subsets of antennas into parallel processing procedures coordinated by a central processing unit. Additionally, authors also propose strategies for further reducing the complexity and overhead of the information exchange between parallel subarrays and the central processing unit to facilitate the practical implementation of the proposed detector.

Recently, to deal with subarrays and channel scatterers in non-stationary XL-MIMO environment, [12] proposed two *channel estimation methods* based on subarray-wise and scatterer-wise near-field non-stationary channel properties. Authors model the multipath channel with the last-hop scatterers under a spherical wavefront and divide the large aperture array into multiple subarrays. The proposed channel estimation methods position the scatterers and perform a mapping between subarrays and scatterers. Hence, the scatterer-wise method simultaneously positions each scatterer and detects its VR to further enhance the positioning accuracy. Moreover, the subarray-wise method can achieve low mean square error (MSE) performance under low-complexity, whereas the scatterer-wise method can accurately arrange the scatterers and determine the non-stationary channel.

In [13], authors propose and validate realistic channel models when employing physically-large arrays, in which non-stationarities and visibility regions are present, as in the XL-MIMO system. The statistical distribution of important channel parameters are found based on measurements. Such contributions are proposed as extensions to the COST 2100 channel model. Besides, key statistical properties of the proposed extensions, e.g., autocorrelation functions, maximum likelihood estimators, and Cramer-Rao bounds, are derived and analyzed. Furthermore, the performance of a spatial modulation massive MIMO system is investigated in [14] under a non-stationary channel model. Authors show that spatial modulation can outperform typical employed spatial multiplexing transmission in certain scenarios of low correlation among sub-channels, for example under a rich scattering environment.

A novel random access (RA) protocol for crowded XL-MIMO systems is proposed in [15]. Authors have proposed a decentralized and uncoordinated decision rule, which can be evaluated at the users side, for retransmitting or not the RA pilots during the connection stage, taking advantage of the XL-MIMO propagation features. The proposed protocol achieves significant performance improvements in terms of reducing the connection delay and providing access for larger number of devices.

1.1 Motivation, Contributions and Novelties in Comparison with Existing Works

Current design approaches in telecommunication systems include a global effort in saving energy and reducing pollution [2], [16], [17]. We show in this paper that antenna selection (AS) methods in XL-MIMO systems is a very important issue since the energy expenditure of such systems could be very high if activating the radio frequency (RF) chains of all antennas simultaneously. Besides, some antennas might contribute very little with the system performance due to the non-stationarities and visibility regions, in such a way that the power required to activate their RF chain becomes a burden that severely penalizes the total energy efficiency (EE) of the system. Therefore, the very large number of antennas deployed in the XL-MIMO systems in conjunction with the spatial non-stationarities make the application of AS schemes very important.

The main *contributions* of this work are threefold:

(**i**) Reformulating the signal to interference plus noise ratio (SINR) performance expressions of [5], considering long-term fading variations across the array and incorporating the maximum transmit power constraint into the expressions for

CB and ZF, and finding more compact and comprehensive results, readily applicable for antenna selection procedures.

(ii) Based on the obtained expressions, and on a realistic power consumption model, we evaluate the total EE of the XL-MIMO system. Besides, we propose and compare four low-complexity AS procedures aiming to maximize the total EE of the system, different than [3, 4] which proposed SE-based AS schemes. Our proposed schemes are based solely on the long-term fading parameters, and the obtained solutions remain valid for larger time/frequency intervals.

(iii) Based on our proposed AS schemes, and some simplifying assumptions, we derive approximated closed-form EE expressions, and propose an iterative method for finding the optimal number of selected antennas which maximizes EE. Finally, numerical simulations have validated the proposed performance expressions and compared the different XL-MIMO AS schemes.

AS methods for typical spatially stationary massive MIMO systems [18, 19] is a well investigated topic. However, the XL-MIMO system is a different scenario. While the spatially stationary model applies for typical cellular systems, where the BS antenna array dimension is much lower than the distance to the users and a single long-term fading coefficient holds for all antennas, significant power variations appear along the XL-MIMO array, due to its large dimension and number of antennas, and proximity with users. The non-stationary XL-MIMO scenario just very recently was introduced in the literature. To the best of our knowledge, this contribution is the first evaluating the EE of the XL-MIMO scenario, showing that AS methods are especially important to improve EE due to the spatial non-stationarities that naturally arise in XL-MIMO systems, proposing long-term fading based AS procedures, and deriving the optimal number of active antennas for this new wireless communication context.

With respect to the existing XL-MIMO literature, we can point out as the *main novelties* of our paper: although our system model and CB and ZF performance expressions are similar to that of [5], authors have considered a binary visibility region model for the XL-MIMO scenario, in which no long-term fading variation occurs for the visible antennas. Besides, performance expressions are dependent of power coefficients obtained resolving a separated optimization problem for meeting power constraint, and no antenna selection is considered. Differently, we incorporated the power constraint into the performance expressions, arriving at more compact and comprehensive results, readily applicable for AS procedures, and considered long-term fading variations along the array. Besides, AS for XL-MIMO systems has been investigated only in [3, 4] at the moment of writing this paper; however, both works proposed SE-based AS schemes for XL-MIMO systems. Differently, based on only long-term fading coefficients, we propose AS schemes aiming to maximize the XL-MIMO total EE, since this is a very important issue due to the very large number of antennas at the XL-MIMO array, and the non-stationarities and visibility regions which arise in this scenario. Furthermore, the long-term fading approach has the advantages of being simpler than short-term ones, and of providing solutions which remain valid for larger time periods and all subcarriers (if employing a wideband system), reducing the computational complexity of the antenna selection approach and simplifying hardware due to switching and RF chain on-off requirements.

Notations: Boldface lower and upper case symbols represent vectors and matrices, respectively. $\mathbf{I}_N$ denotes the identity matrix of size N, while $\{\cdot\}^T$ and $\{\cdot\}^H$ denote the transpose and the Hermitian transpose operator, respectively. We use $\mathcal{CN}(m, \sigma^2)$ when referring to a circular symmetric complex Gaussian distribution with mean m and variance matrix σ^2. Besides, $\mathrm{tr}(\cdot)$ and $\mathrm{diag}(\cdot)$ are the trace and diagonal matrix operators, respectively, while $[\mathbf{A}]_{i,j}$ holds to the element in the ith row and jth column of matrix $\mathbf{A}$, and $\mathbf{a}_i$ refers to its ith column vector.

2 System Model

We consider a base station (BS) equipped with a linear XL-MIMO array with M antennas uniformly distributed along a length of L meters , Fig. G.1. In front of the extra-large array structure, K users are randomly distributed in a rectangular area, of length L in the array parallel dimension, and with a distance to the array in the range $[0.1 \cdot L, L]$[1]. Since the distances of the users to the antennas is of the same order of the array length L the average received power varies along the XL-MIMO array, and therefore we cannot consider a single long-term fading coefficient for a given user [3, 8]. Instead, we consider a long-term fading coefficient $\beta_{m,k}$ regarding the m-th antenna of the XL-MIMO array and the k-th user, similarly as in [3, 9, 10, 15], given by

$$\beta_{m,k} = q \cdot d_{m,k}^{-\kappa}, \tag{G.1}$$

in which q is a constant determining the path loss in a reference distance, $d_{m,k}$ is the distance between the m-th antenna of the XL-MIMO array and the k-th user, and κ is the path loss decay exponent. The channel matrix $\mathbf{H} \in \mathbb{C}^{M \times K}$ is thus formed by elements $h_{m,k} = \sqrt{\beta_{m,k}} \cdot \underline{h}_{m,k}$, in which $\underline{h}_{m,k} \sim \mathcal{CN}(0, 1)$, assuming a rich scattering environment as in [4, 5]. If we arrange the long-term fading coefficients of a user in a diagonal matrix:

$$\mathbf{R}_k = \mathrm{diag}([\beta_{1,k}, \beta_{2,k}, \ldots, \beta_{M,k}]) \in \mathbb{R}^{M \times M}, \tag{G.2}$$

and the elements $\underline{h}_{m,k}$ in a vector $\underline{\mathbf{h}}_k \in \mathbb{C}^{M \times 1}$, we have that each column of $\mathbf{H}$ can be defined as $\mathbf{h}_k = \mathbf{R}_k^{\frac{1}{2}} \underline{\mathbf{h}}_k$ as in [5].

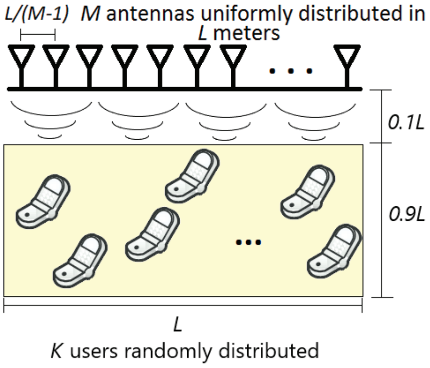

L/(M-1)
M antennas uniformly distributed in L meters
0.1L
0.9L
L
K users randomly distributed

Fig. G.1: Illustration of the adopted system model.

In the DL, considering an average received signal-to-noise ratio (SNR) ρ at the users, an average long-term fading coefficient β_{avg} (among all antennas and users'

[1] In order to guarantee a minimum distance of the users to the XL-MIMO array, as in [10, 16].

positions), and a uniform power allocation policy for the users, the total transmit power, $P_{\max}$, should satisfy [1]

$$\rho = \frac{P_{\max} \cdot \beta_{\text{avg}}}{\sigma^2}, \tag{G.3}$$

in which σ^2 is the noise power. Since the channel gain $\beta_{m,k}$ varies significantly along the array, it is more effective to select just the stronger antennas to transmit signal to the k-th user, reducing the number of active antennas , as well the power spent with power-hungry RF chains. We discuss in the next Section different approaches to obtain the set of antennas selected to serve the users, $\mathcal{A}$. For simplicity, we considered $\beta_{\text{avg}} \approx q \cdot L^{-\kappa}$ in our simulations. The signal for user k, s_k, is precoded by $\mathbf{g}_k \in \mathbb{C}^{M\times 1}$ and scaled by $p_k \geq 0$, which adjusts the signal power, before transmission. Considering a similar XL-MIMO system model than [5], the transmit vector $\mathbf{x}$ is the linear combination of the precoded and scaled signal of all the users, *i.e.*,

$$\mathbf{x} = \sum_{k=1}^{K} \sqrt{p_k} \cdot \mathbf{g}_k \cdot s_k. \tag{G.4}$$

Let $\mathbf{G} = [\mathbf{g}_1, \mathbf{g}_2, \dots, \mathbf{g}_K] \in \mathbb{C}^{M\times K}$ be the combined precoding matrix, and $\mathbf{P} = \text{diag}([p_1, p_2, \dots, p_K]) \in \mathbb{R}^{K\times K}$ be the diagonal matrix of signal powers. The combined precoding matrix $\mathbf{G}$ is normalized to satisfy the power constraint

$$\mathbb{E}[||\mathbf{x}||^2] = \text{tr}(\mathbf{P}\mathbf{G}^H\mathbf{G}) = P_{\max}. \tag{G.5}$$

The signal received by the k-th user is

$$y_k = \mathbf{h}_k^H \mathbf{x} + n_k, \quad k = 1, 2, \dots K, \tag{G.6}$$

in which $n_k \sim \mathcal{CN}(0, \sigma^2)$ is an additive white Gaussian noise (AWGN) sample. Assuming independent Gaussian signaling, *i.e.*, $s_k \sim \mathcal{CN}(0, 1)$ and $\mathbb{E}[s_i s_j^*] = 0$, $i \neq j$, the SINR γ_k of the k-th user can be defined as [5]:

$$\gamma_k = \frac{p_k |\mathbf{h}_k^H \mathbf{g}_k|^2}{\sum_{j=1, j\neq k}^{K} p_j |\mathbf{h}_k^H \mathbf{g}_j|^2 + \sigma^2}. \tag{G.7}$$

We selected the CB and ZF approaches as representative low-complexity linear precoding schemes. The CB precoder matrix is simply defined as

$$\mathbf{G}_{\text{CB}} = \alpha_{\text{CB}} \mathbf{H}, \tag{G.8}$$

and the ZF precoding matrix is

$$\mathbf{G}_{\text{ZF}} = \alpha_{\text{ZF}} \mathbf{H} (\mathbf{H}^H \mathbf{H})^{-1}, \tag{G.9}$$

where the scaling factors $\alpha_{\text{CB}} = \sqrt{P_{\max} / \text{tr}(\mathbf{P}\mathbf{H}^H\mathbf{H})}$ and $\alpha_{\text{ZF}} = \sqrt{P_{\max} / \text{tr}(\mathbf{P}(\mathbf{H}^H\mathbf{H})^{-1})}$ ensure that the power constraint (G.5) is met.

Using (G.8) in (G.7), the SINR of the kth user for CB is

$$\gamma_k^{(\text{CB})} = \frac{p_k |\mathbf{h}_k^H \mathbf{h}_k|^2}{\sum_{j=1, j\neq k}^{K} p_j |\mathbf{h}_k^H \mathbf{h}_j|^2 + \frac{\sigma^2}{P_{\max}} \text{tr}(\mathbf{P}\mathbf{H}^H\mathbf{H})}. \tag{G.10}$$

Similarly, using (G.9) in (G.7), the SINR of the kth user for ZF is

$$\gamma_k^{(\mathrm{ZF})} = \frac{p_k P_{\max}}{\sigma^2 \mathrm{tr}(\mathbf{P}(\mathbf{H}^H\mathbf{H})^{-1})}. \tag{G.11}$$

Given the system model presented in this Section in eq. (G.1)–(G.11), and the deterministic equivalent analysis of [20], it is presented in [5] the deterministic equivalent of $\gamma_k^{(\mathrm{CB})}$ in (G.10) as

$$\overline{\gamma}_k^{(\mathrm{CB})} = \frac{p_k(\mathrm{tr}(\mathbf{R}_k))^2}{\sum_{j=1,j\neq k}^{K} p_j \mathrm{tr}(\mathbf{R}_k\mathbf{R}_j) + \frac{\sigma^2}{P_{\max}} \sum_{j=1}^{K} p_j \mathrm{tr}(\mathbf{R}_j)}, \tag{G.12}$$

and the deterministic equivalent of $\gamma_k^{(\mathrm{ZF})}$ in (G.11) as

$$\overline{\gamma}_k^{(\mathrm{ZF})} = \frac{p_k P_{\max}}{\sigma^2 \sum_{i=1}^{K} p_i \left(\mathrm{tr}(\mathbf{R}_i) - \sum_{j=1,j\neq i}^{K} \frac{\mathrm{tr}(\mathbf{R}_i\mathbf{R}_j)}{\mathrm{tr}(\mathbf{R}_j)}\right)^{-1}}. \tag{G.13}$$

where $\mathbf{R}_i$ is defined as in (G.2).

Having found the SINR of the kth user, the spectral efficiency is readily obtained as $\eta_k^s = \log_2(1+\gamma_k)$. On the other hand, the energy efficiency is [16, 17]

$$\eta_e = \frac{B\sum_{k=1}^{K} \eta_k^s}{\mathcal{P}}, \tag{G.14}$$

in which B is the system bandwidth, and $\mathcal{P}$ is the total power consumption, discussed in Section 2.3.

2.1 Further Advances in the Performance Expressions

We revisit the performance expressions for non-stationary XL-MIMO discussed in [5], while propose further elaborations to arrive at lean and more comprehensive results. Note that the results of (G.12) and (G.13) depend on the signal powers in both numerator and denominators, and such coefficients should be chosen in order to satisfy the power constraint in (G.5). In the simulation code made available by the authors of [5], they apply the CVX solver of [21] to find a matrix $\mathbf{P}$ satisfying (G.5). This makes the performance expressions less intuitive, while limiting the application of AS schemes as proposed in Section 3 of this paper. Hence, in this subsection, we shed light on deriving self-contained closed-form SINR expressions recalling the channel hardening massive MIMO properties. For that, we first rewrite (G.5) in the following form:

$$\mathbb{E}[||\mathbf{x}||^2] = \mathrm{tr}(\mathbf{P}\mathbf{G}^H\mathbf{G}) = \sum_{k=1}^{K} p_k||\mathbf{g}_k||^2 = P_{\max}. \tag{G.15}$$

If a uniform power allocation scheme is applied, the following equality holds

$$p_k||\mathbf{g}_k||^2 = \frac{P_{\max}}{K}, \quad k = 1, 2, \ldots K. \tag{G.16}$$

Hence, when adopting CB, eq. (G.16) becomes

$$p_k \alpha_{\text{CB}}^2 ||\mathbf{h}_k||^2 = \frac{P_{\max}}{K}, \quad k = 1, 2, \ldots K, \tag{G.17}$$

and we have an undetermined system with K equations and $K+1$ variables. By choosing $\alpha_{\text{CB}} = 1$ for simplicity, the p_k coefficients can be obtained for CB as

$$p_k^{(\text{CB})} = \frac{P_{\max}}{K||\mathbf{h}_k||^2}, \quad k = 1, 2, \ldots K. \tag{G.18}$$

Following similar assumptions as in [5], we have that

$$||\mathbf{h}_k||^2 = \mathbf{h}_k^H \mathbf{h}_k = \underline{\mathbf{h}}_k^H \mathbf{R}_k \underline{\mathbf{h}}_k \xrightarrow{M \to \infty} \text{tr}(\mathbf{R}_k), \tag{G.19}$$

and a deterministic equivalent of (G.18) is

$$p_k^{(\text{CB})} = \frac{P_{\max}}{K\text{tr}(\mathbf{R}_k)}, \quad k = 1, 2, \ldots K. \tag{G.20}$$

Substituting (G.20) in (G.12), we arrive at

$$\overline{\gamma}_k^{(\text{CB})} = \frac{\text{tr}(\mathbf{R}_k)}{\sum_{j=1, j\neq k}^{K} \frac{\text{tr}(\mathbf{R}_k\mathbf{R}_j)}{\text{tr}(\mathbf{R}_j)} + \frac{K\sigma^2}{P_{\max}}}. \tag{G.21}$$

On the other hand, for the case of ZF, (G.5) becomes

$$\begin{aligned}
\mathbb{E}[||\mathbf{x}||^2] &= \text{tr}\left(\mathbf{P}\mathbf{G}^H\mathbf{G}\right) = P_{\max}, \\
&= \alpha_{\text{ZF}}^2 \text{tr}\left(\mathbf{P}(\mathbf{H}^H\mathbf{H})^{-1}\mathbf{H}^H\mathbf{H}(\mathbf{H}^H\mathbf{H})^{-1}\right) = P_{\max}, \\
&= \alpha_{\text{ZF}}^2 \text{tr}\left(\mathbf{P}(\mathbf{H}^H\mathbf{H})^{-1}\right) = P_{\max}, \\
&= \alpha_{\text{ZF}}^2 \text{tr}\left(\mathbf{P}\mathbf{V}\right) = P_{\max},
\end{aligned} \tag{G.22}$$

in which the matrix $\mathbf{V}$ is a diagonal matrix formed by the main diagonal elements of $(\mathbf{H}^H\mathbf{H})^{-1}$. We can thus rewrite (G.22) as

$$\alpha_{\text{ZF}}^2 \sum_{k=1}^{K} p_k [\mathbf{V}]_{k,k} = P_{\max}, \tag{G.23}$$

and if a uniform power allocation is employed

$$\alpha_{\text{ZF}}^2 \, p_k \, [\mathbf{V}]_{k,k} = \frac{P_{\max}}{K}, \quad k = 1, 2, \ldots K. \tag{G.24}$$

Again, making $\alpha_{\text{ZF}} = 1$, the p_k coefficients can be obtained for the ZF precoding as

$$p_k^{(\text{ZF})} = \frac{P_{\max}}{K\,[\mathbf{V}]_{k,k}}, \quad k = 1, 2, \ldots K. \tag{G.25}$$

Following the analysis of [5, App. A], it can be shown that

$$[\mathbf{V}]_{k,k} \xrightarrow{M \to \infty} \left(\text{tr}(\mathbf{R}_k) - \sum_{j=1, j\neq k}^{K} \frac{\text{tr}(\mathbf{R}_k\mathbf{R}_j)}{\text{tr}(\mathbf{R}_j)}\right)^{-1}, \tag{G.26}$$

and a deterministic equivalent of (G.25) is

$$p_k^{(\mathrm{ZF})} = \frac{P_{\max}}{K}\left(\mathrm{tr}(\mathbf{R}_k) - \sum_{j=1,j\neq k}^{K}\frac{\mathrm{tr}(\mathbf{R}_k\mathbf{R}_j)}{\mathrm{tr}(\mathbf{R}_j)}\right),\ k = 1,\ldots K. \tag{G.27}$$

Substituting (G.27) in (G.13), we arrive at

$$\overline{\gamma}_k^{(\mathrm{ZF})} = \frac{P_{\max}}{K\sigma^2}\left(\mathrm{tr}(\mathbf{R}_k) - \sum_{j=1,j\neq k}^{K}\frac{\mathrm{tr}(\mathbf{R}_k\mathbf{R}_j)}{\mathrm{tr}(\mathbf{R}_j)}\right). \tag{G.28}$$

Equations (G.21) and (G.28) show the XL-MIMO DL system performance employing CB and ZF, respectively, as further extensions of eq. (G.12) and (G.13) from [5]. This is a first contribution of this manuscript, which serves as basis for the following EE and AS analysis.

Remark 1: Although we have considered $\alpha_{\mathrm{CB}} = \alpha_{\mathrm{ZF}} = 1$ in our analysis, any other choice for these parameters would result in the same expressions, since would affect every numerator and denominator terms in the same way.

Remark 2: The SINR performance expressions presented in [5, Table I] can be seen as particular cases of (G.21) and (G.28) when neglecting long-term fading and applying the normalization $\mathrm{tr}(\mathbf{R}_k) = \mathrm{tr}(\mathbf{\Theta}_k) = M$ or $\mathrm{tr}(\mathbf{\Theta}_k) = D$, where $\mathbf{\Theta}_k$ and D are the matrix describing the VR of kth user and the number of visible antennas per user, respectively, as in [5].

2.2 Antenna Selection Model

Given our deterministic equivalent performance expressions for CB and ZF in eq. (G.21) and (G.28), respectively, we can rewrite these expressions considering the activation subset of antennas. Hence, denoting $\mathcal{A}$ as the set containing the indices of the active antennas, the deterministic equivalent SINR for the CB precoding results

$$\overline{\gamma}_k^{(\mathrm{CB})} = \frac{\sum_{m\in\mathcal{A}}\beta_{m,k}}{\sum_{j=1,j\neq k}^{K}\frac{\sum_{m\in\mathcal{A}}\beta_{m,k}\beta_{m,j}}{\sum_{m\in\mathcal{A}}\beta_{m,j}} + \frac{K\sigma^2}{P_{\max}}}, \tag{G.29}$$

while for the ZF:

$$\overline{\gamma}_k^{(\mathrm{ZF})} = \frac{P_{\max}}{K\sigma^2}\left(\sum_{m\in\mathcal{A}}\beta_{m,k} - \sum_{j=1,j\neq k}^{K}\frac{\sum_{m\in\mathcal{A}}\beta_{m,k}\beta_{m,j}}{\sum_{m\in\mathcal{A}}\beta_{m,j}}\right). \tag{G.30}$$

It is worth to note that, in our formulation, the *activation subset of antennas* is the same for all users, differently from [3], in which each user has its own set of active antennas aiming to maximize the system sum-rate. We justify our formulation since, when aiming to maximize the total energy efficiency, once the power-hungry RF chain of an antenna is active, it is better to take full advantage of it, transmitting signal for all users. It has no significant benefit in defining the activation subset of antennas in a per-user fashion, since the ZF approach is able to eliminate the inter-user interference, while the power increment necessary to compute the precoding vector with a slightly large number of antennas is small if compared to the power to activate the RF chain of the additional antenna, as evinced in the next subsection. Besides, it would result in more complicated performance expressions, probably in terms of short-term fading coefficients, and the dimension of the search space of the AS algorithms would scale with K, becoming considerably more complex and power consuming.

2.3 Power Consumption Model

We follow the same power consumption model of [16], which is very similar to that in [17], and is a very realistic model. However, as we focus on the DL transmission, we do not consider the UL data rates as well as the UL transmit powers. In the XL-MIMO scenario analysed herein, we consider the power expenditures of the irradiated DL data signal (with the amplifier efficiency), $P_{\text{TX}}^{\text{DL}}$, the UL training, $P_{\text{TX}}^{\text{tr}}$, the channel estimation, P_{CE}, the coding/decoding, $P_{\text{C/D}}$, the backhaul, P_{BH}, the linear processing computation, P_{PR}, the transceiver chains, P_{TC}, and a fixed quantity regarding the circuitry power consumption required for site-cooling, control signaling, and load-independent power of backhaul infrastructure and baseband processors, P_{FIX}. Thus, the overall power consumption results

$$\mathcal{P} = P_{\text{TX}}^{\text{DL}} + P_{\text{TX}}^{\text{tr}} + P_{\text{CE}} + P_{\text{C/D}} + P_{\text{BH}} + P_{\text{PR}} + P_{\text{TC}} + P_{\text{FIX}}. \tag{G.31}$$

Our objective here is to investigate the dependence of the selected subset of antennas, $\mathcal{A}$, with the total energy efficiency of the system. Note that the total energy efficiency of the system depends on $\mathcal{A}$ in different ways. First, the sum rate of the system depends on the SE of the users, which is a function of their SINRs dependent of $\mathcal{A}$. Moreover, the sum rate impacts on the power expenditures of the coding/decoding, and the backhaul. Besides, the power consumption of the transceiver chains is modeled as

$$P_{\text{TC}} = P_{\text{SYN}} + |\mathcal{A}| P_{\text{BS}} + K P_{\text{MT}}, \tag{G.32}$$

in which P_{SYN} is the power of the local oscillator, P_{BS} is the power required to each active BS antenna operate, while P_{MT} is the power required to each single-antenna mobile terminal (MT) operate. Note that M is usually very high in an XL-MIMO system[2], while P_{BS} accounting for the power-hungry RF chains is considered in [16] as 1 W per antenna. Thus, activating the RF chains of all BS antennas would result in a very large power expenditure, in such a way that it is very important to perform a suitable antenna selection procedure.

The power consumed with processing, P_{PR}, corresponds to the power required to obtain the transmit signal in (G.4), to obtain the precoding matrix, and to obtain the AS set. Note that this power is also dependent on the number of active antennas $|\mathcal{A}|$. Following the model in [16], but including the term of power related to the AS processing, we have

$$P_{\text{PR}} = B\left(1 - \frac{\tau}{\mathcal{S}}\right)\frac{\mathcal{C}_{\text{ts}}}{\mathcal{L}_{\text{BS}}} + \frac{B}{\mathcal{S}}\frac{\mathcal{C}_{\text{prec}}}{\mathcal{L}_{\text{BS}}} + \frac{1}{T_{\text{LT}}}\frac{\mathcal{C}_{\text{as}}}{\mathcal{L}_{\text{BS}}}, \tag{G.33}$$

in which τ is the length of the uplink pilot signals, $\mathcal{S}$ is the coherence block size, $\mathcal{C}_{\text{ts}}$ is the computational complexity for evaluating eq. (G.4). Besides, $\mathcal{L}_{\text{BS}}$ is the

[2] Typically hundreds or even thousands of antennas.

computational efficiency of the BS (in W/$flop$), $\mathcal{C}_{\text{prec}}$ is the complexity of obtaining the precoding vectors for all users, T_{LT} is the long-term fading coherence time, and $\mathcal{C}_{\text{as}}$ is the complexity of obtaining the antenna selection set. The obtained AS set remains valid for a long-term coherence interval, since our analysis is based only in long-term fading parameters. One can see from (G.33) that this approach results in a lower influence of the AS set computation in P_{PR}, since it is multiplied by the factor $1/T_{\text{LT}}$, which is much lower than $B/\mathcal{S}$ and $B\left(1-\frac{\tau}{\mathcal{S}}\right)$.

Following the analysis in [16], [17], we consider 1 $flop$ as an arithmetic operation between two complex numbers. Thus, the multiplication between a matrix $\mathbf{A} \in \mathbb{C}^{m\times n}$ and a matrix $\mathbf{B} \in \mathbb{C}^{n\times p}$ spends $2mnp$ flops. Therefore, we have $\mathcal{C}_{\text{ts}} = 2|\mathcal{A}|K\,flops$ from [17]. Besides, if using the CB precoder, $\mathcal{C}_{\text{prec}} = \mathcal{C}_{\text{CB}} = 3|\mathcal{A}|K\,flops$ from [17], against $\mathcal{C}_{\text{prec}} = \mathcal{C}_{\text{ZF}} = K^3/3 + 3|\mathcal{A}|K^2 + |\mathcal{A}|K\,flops$ if adopting ZF. The complexity $\mathcal{C}_{\text{as}}$ is discussed in the next Section. Besides, the terms in (G.31) not discussed in this Section can be computed in the same way as in [16].

Finally, we can rewrite (G.31) as

$$\mathcal{P} = \mathcal{P}^{\dagger} + P_{\text{CE}} + P_{\text{C/D}} + P_{\text{BH}} + P_{\text{PR}} + |\mathcal{A}|P_{\text{BS}}, \tag{G.34}$$

in which we have gathered the power components that do not depend of $\mathcal{A}$ in the term:

$$\mathcal{P}^{\dagger} = P_{\text{TX}}^{\text{DL}} + P_{\text{TX}}^{\text{tr}} + P_{\text{SYN}} + KP_{\text{MT}} + P_{\text{FIX}}. \tag{G.35}$$

The dependence of the terms in (G.34) with $\mathcal{A}$ can be justified as follows: P_{CE} depends on $\mathcal{A}$ since the short-term channel estimates are obtained only for the active antennas, $P_{\text{C/D}}$ and P_{BH} because they depend on the system sum-rate, which depends on $\mathcal{A}$, and P_{PR} because the processing complexity is dependent on the number of active antennas.

3 Antenna Selection Schemes

In this section we propose different AS schemes for XL-MIMO aiming to obtain a suitable subset of antennas $\mathcal{A}$ selected to transmit the DL signal to the mobile users subject to channel non-stationarities. First we propose a simple, deterministic, greedy scheme based on the *highest received normalized power* (HRNP) criterion. Then, three heuristic schemes are proposed using the HRNP active antennas set as initial solution: *local search* (LS), *genetic algorithm* (GA), and *particle swarm optimization* (PSO).

3.1 HRNP criterion

A first and greedy approach is to select just the M_s antennas responsible for the major part of the power received by the users. However, since closer users receive more power, this should be performed in a normalized fashion in order to achieve a fair result for all users. In this case, we first compute the metric:

$$\varphi_m = \sum_{k=1}^{K} \frac{\beta_{m,k}}{\sum_{j=1}^{M}\beta_{j,k}}, \quad m = 1, 2, \ldots, M. \tag{G.36}$$

Then, the selected subset of antennas $\mathcal{A}^{\text{HRNP}}$ will be composed by the M_s antennas with the highest values of φ_m. A pseudo-code for the HRNP-AS procedure is presented in Algorithm 13, in which $\boldsymbol{\varphi} = [\varphi_1, \varphi_2, \dots \varphi_M]$.

The complexity[3] of the HRNP AS scheme is described by

$$\mathcal{C}_{\text{as}}^{\text{HRNP}} = 3MK + M\log(M) \quad [flops], \tag{G.37}$$

corresponding to the computation of (G.36) for all antennas, and a sorting algorithm to select the M_s antennas with highest φ_m. It is noteworthy, however, that the HRNP EE performance is highly dependent on the M_s choice, since the system would provide low sum-rates with few active antennas, or it would consume a high power with many active antennas. Thus, we propose in Section 4 an approximated closed-form analytical expression for the EE of the XL-MIMO system employing ZF and HRNP-AS as a function of M_s. Then, we propose an iterative method for obtaining the M_s value which maximizes this expression. We do not consider the complexity of this method in eq. (G.37) since it is not dependent on the channel parameters, but only controlled by the system parameters, such as the number of users, transmit power, dimensions of XL-MIMO array and coverage area. Therefore, its computation can be performed over larger time periods. We discuss in Section 5.1 the complexity of the proposed method for obtaining the optimal M_s value.

Algorithm 13 Proposed HRNP AS Scheme

```
Input: $M_s$, $\beta_{m,k}$, $\forall m, k$.

Initialize $\mathcal{A}$ as an empty set;
for $m = 1, 2, \dots, M$ do
    Evaluate $\varphi_m$ as in (G.36);
end for
for $n = 1, 2, \dots, M_s$ do
    Evaluate $a = \arg\max_m \varphi_m$;
    Update $\mathcal{A}$ as $\mathcal{A} = \mathcal{A} \cup a$;
    Remove $\varphi_a$ from $\boldsymbol{\varphi}$;
end for
Output: $\mathcal{A}^{\text{HRNP}}$.
```

3.2 LS-based Antenna Selection

A simple strategy for seeking a better active antennas set is to perform a local search (LS) in the neighborhood of the HRNP solution. For this purpose, we first represent the set $\mathcal{A}$ as a binary vector $\mathbf{a}$ of length M, in which if $m \in \mathcal{A}$,

[3] We evaluate the computational complexities of the investigated schemes in terms of *floating point operations (flops)*, defined as an addition, subtraction, multiplication or division between two floating point numbers [22].

$a_m = 1$; otherwise $a_m = 0$. Then, we compute the total energy efficiency (G.14) of every candidate within a certain Hamming distance d_{Ham} from it. If a better candidate is found, the solution is updated, and the procedure is repeated on its neighborhood. This iterative procedure is repeated for a predefined number of iterations or until the convergence. A pseudo-code representation of the LS-based AS scheme is provided in Algorithm 14, in which N_{el} as defined in step 3 is the number of elements within the Hamming distance d_{Ham} from the current solution. For simplicity, we have limited our search with a unitary Hamming distance.

The complexity of the LS-AS scheme is

$$\mathcal{C}_{\text{as}}^{\text{LS}} = \mathcal{C}_{\text{as}}^{\text{HRNP}} + \overline{N}_{it} M \mathcal{C}_{\text{EE}} \quad [flops], \tag{G.38}$$

in which $\overline{N}_{it}$ is the average number of iterations until convergence, and $\mathcal{C}_{\text{EE}} = 2MK^2\ [flops]$ is the complexity of computing the total energy efficiency cost function. An interesting point to observe in the LS algorithm is that if a new solution is not found into an iteration, the search can be interrupted, since the algorithm has converged. This contributes to decrease the complexity of the algorithm, and, therefore, improve EE.

Algorithm 14 Proposed LS-based AS Scheme

```
Input: $d_{\text{Ham}}$, $N_{it}^{\max}$, $\mathcal{A}^{\text{HRNP}}$, $\beta_{m,k}$, $\forall m, k$.

Initialize $\mathbf{a}^0$ as the binary vector representation of $\mathcal{A}^{\text{HRNP}}$;
Initialize $\eta_e^{\text{best}}$ as the total energy efficiency of $\mathbf{a}^0$;
Evaluate $N_{el} = \binom{M}{d_{\text{Ham}}}$;
for $n = 1, 2, \cdots, N_{it}^{\max}$ do
    Generate the search space matrix $\mathbf{S}$ of size $M \times N_{el}$ with all vectors
    within the distance $d_{\text{Ham}}$ from $\mathbf{a}^{n-1}$;
    for $\ell = 1, 2, \ldots, N_{el}$ do
        Evaluate $\eta_e$ as the total energy efficiency of $\mathbf{s}_\ell$;
        if $\eta_e > \eta_e^{\text{best}}$ then
            Update $\eta_e^{\text{best}} = \eta_e$, and $\mathbf{a}^n = \mathbf{s}_\ell$;
        else
            Break;
        end if
    end for
end for
Output: $\mathcal{A}^{\text{LS}}$ as the set representation of $\mathbf{a}^n$.
```

3.3 GA-based Antenna Selection

The genetic algorithm is a widely-known bio-inspired heuristic optimization algorithm, which has been used to solve optimization problems in different areas. In the context of massive MIMO antenna selection, GA has been employed in the conventional stationary case in [18]. Herein, we employ a similar algorithm from [18], but adjusted to the non-stationary XL-MIMO configurations. The GA-AS uses the HRNP output as initial solution, also, other random candidates forming an initial population of size p_{GA}, which is evaluated in terms of the cost function in (G.14). A given number ϕ of the best candidates in this population is selected as parents, which will generate descendants in a new population. For this purpose, two parents are selected at random for each descendant, and the crossover operator is applied with a random crossover point. Then, the mutation operator is also applied, which inverts the entries of each candidate with certain probability p_{mut}. After a predefined number of iterations or until the convergence of the algorithm, it returns the best solution found so far. A pseudo-code representation for the GA-based AS scheme is provided in Algorithm 15.

The complexity of our proposed GA-AS procedure is

$$\mathcal{C}_{\text{as}}^{\text{GA}} = \mathcal{C}_{\text{as}}^{\text{HRNP}} + \overline{N}_{it}[p_{\text{GA}}\mathcal{C}_{\text{EE}} + p_{\text{GA}} \log(p_{\text{GA}})] \quad [flops], \tag{G.39}$$

due to the cost function evaluation of each candidate in the population, and a sorting algorithm for selecting the best candidates.

3.4 PSO-based Antenna Selection

The particle swarm optimization algorithm is another bio-inspired optimization algorithm, similarly as GA. However, it is commonly recognized as a simpler algorithm, in terms of fewer mechanisms to escape from local maxima, and reduced computational complexity per iteration. Therefore, we also suggest the use of a PSO-based AS scheme for the non-stationary XL-MIMO case, similarly as proposed in [19] for conventional stationary massive MIMO scenario.

The PSO-AS algorithm uses the HRNP output as initial solution, as well as other random candidates to form an initial swarm of p_{PSO} particles. At each iteration, each particle updates its position in terms of its previous velocity (inertial effect, with inertia weight ν), its individual best solution found (cognitive information, with cognitive factor μ_c), and the best solution found by all particles (social information, with social factor μ_s). After a predefined number of iterations or the convergence of the algorithm, it returns the best solution found. A pseudo-code representation for the PSO-based AS scheme is provided in Algorithm 16, in which $\mathbf{\Gamma} \in \mathbb{R}^{M \times p_{\text{PSO}}}$ is a random matrix generated each time it is called with each element uniformly distributed in $[0, 1]$

Algorithm 15 Proposed GA-based AS Scheme

```
Input: $p_{\text{GA}}$, $\phi$, $p_{\text{mut}}$, $\mathcal{A}^{\text{HRNP}}$, $N_{it}^{\max}$, $\beta_{m,k}$, $\forall m, k$.
Initialize the population $\boldsymbol{\Theta}^{\text{GA}}$ with the binary vector representation of $\mathcal{A}^{\text{HRNP}}$ and other $p_{\text{GA}}$-1 random binary vectors;
Evaluate the total energy efficiency of each candidate in $\boldsymbol{\Theta}^{\text{GA}}$, forming the vector $\boldsymbol{\eta}_e^{\text{GA}}$;
Sort $\boldsymbol{\eta}_e^{\text{GA}}$ in descending order, reorganizing the columns of $\boldsymbol{\Theta}^{\text{GA}}$ accordingly;
Initialize $\eta_e^{\text{best}} = \eta_{e,1}^{\text{GA}}$, and $\mathbf{a}^{\text{GA}} = \boldsymbol{\theta}_1^{\text{GA}}$;
for $n = 2, 3, \ldots, N_{it}^{\max}$ do
    for $\ell = 1, 2, \ldots, p_{\text{GA}}$ do
        Generate two different random integers $\in [1, \phi]$ to be the parents of $\boldsymbol{\theta}_\ell^{\text{GA}}$, applying the crossover operator in a random crossover point $\in [2, M]$;
        Apply the mutation operator in $\boldsymbol{\theta}_\ell^{\text{GA}}$ with probability $p_{\text{mut}}$;
        Evaluate the total energy efficiency of $\boldsymbol{\theta}_\ell^{\text{GA}}$, and assign it to $\eta_{e,\ell}^{\text{GA}}$;
    end for
    Sort $\boldsymbol{\eta}_e^{\text{GA}}$ in descending order, reorganizing the columns of $\boldsymbol{\Theta}^{\text{GA}}$ accordingly;
    if $\eta_{e,1}^{\text{GA}} > \eta_e^{\text{best}}$ then
        Update $\eta_e^{\text{best}} = \eta_{e,1}^{\text{GA}}$, and $\mathbf{a}^{\text{GA}} = \boldsymbol{\theta}_1^{\text{GA}}$;
    end if
end for
Output: $\mathcal{A}^{\text{GA}}$ as the set representation of $\mathbf{a}^{\text{GA}}$.
```

interval, and `binround`(x) is the binary round operator, which returns 1 if $x > 0.5$, and 0 otherwise.

The complexity of the proposed PSO-AS algorithm is

$$\mathcal{C}_{\text{as}}^{\text{PSO}} = \mathcal{C}_{\text{as}}^{\text{HRNP}} + \overline{N}_{it}(p_{\text{PSO}}\mathcal{C}_{\text{EE}} + p_{\text{PSO}}) \quad [flops], \tag{G.40}$$

due to the cost function evaluation (G.14) for all particles and finding the maximum EE particle, at each iteration.

4 Optimal number of selected antennas: an iterative-analytical method

In this Section we derive approximated performance analytical expressions for XL-MIMO systems employing the ZF precoder and the HRNP-based AS

Algorithm 16 Proposed PSO-based AS Scheme

```
Input: $p_{\text{PSO}}$, $\nu$, $\mu_c$, $\mu_s$, $\mathcal{A}^{\text{HRNP}}$, $N_{it}^{\max}$, $\beta_{m,k}$, $\forall m, k$.
Initialize the positions $\mathbf{\Theta}^{\text{PSO}}$ with the binary vector representation of $\mathcal{A}^{\text{HRNP}}$ and other $p_{\text{PSO}}$-1 random binary vectors;
Evaluate the total energy efficiency of each candidate in $\mathbf{\Theta}^{\text{PSO}}$, forming the vector $\boldsymbol{\eta}_e^{\text{PSO}}$;
Initialize the social information $\eta_e^{\text{best}} = \eta_{e,\phi}^{\text{PSO}}$, and $\mathbf{a}^{\text{PSO}} = \boldsymbol{\theta}_{\phi}^{\text{PSO}}$, in which $\phi = \arg\max_n \eta_{e,n}^{\text{PSO}}$;
Initialize the cognitive information $\boldsymbol{\eta}_e^{c} = \boldsymbol{\eta}_e^{\text{PSO}}$, and $\mathbf{\Theta}_c^{\text{PSO}} = \mathbf{\Theta}^{\text{PSO}}$;
Initialize the velocity matrix $\mathbf{V} \in \mathbb{R}^{M \times p_{\text{PSO}}}$ with random elements uniformly distributed in $[-1, 1]$;
for $n = 2, 3, \dots, N_{it}^{\max}$ do
    Update the velocity matrix $\mathbf{V} = \nu\mathbf{V} + \mu_c\mathbf{\Gamma}\left[\mathbf{\Theta}_c^{\text{PSO}} - \mathbf{\Theta}^{\text{PSO}}\right] + \mu_s\mathbf{\Gamma}\left[\mathbf{a}^{\text{PSO}} - \mathbf{\Theta}^{\text{PSO}}\right]$;
    Update the positions $\mathbf{\Theta}^{\text{PSO}} = \texttt{binround}\left(\mathbf{\Theta}^{\text{PSO}} + \mathbf{V}\right)$;
    for $\ell = 1, 2, \dots p_{\text{PSO}}$ do
        Evaluate the total energy efficiency of $\boldsymbol{\theta}_{\ell}^{\text{PSO}}$, and assign it to $\eta_{e,\ell}^{\text{PSO}}$;
        if $\eta_{e,\ell}^{\text{PSO}} > \eta_{e,\ell}^{c}$ then
            Update $\eta_{e,\ell}^{c} = \eta_{e,\ell}^{\text{PSO}}$, and $\boldsymbol{\theta}_{c,\ell}^{\text{PSO}} = \boldsymbol{\theta}_{\ell}^{\text{PSO}}$;
            if $\eta_{e,\ell}^{\text{PSO}} > \eta_e^{\text{best}}$ then
                Update $\eta_e^{\text{best}} = \eta_{e,\ell}^{\text{PSO}}$, and $\mathbf{a}^{\text{PSO}} = \boldsymbol{\theta}_{\ell}^{\text{PSO}}$;
            end if
        end if
    end for
end for
Output: $\mathcal{A}^{\text{PSO}}$ as the set representation of $\mathbf{a}^{\text{PSO}}$.
```

method. Such expressions are compared with numerical results obtained via Monte-Carlo simulation method in Section 5, confirming the tightness of the derivations proposed herein. Then, based on these analytical expressions, we devise an analytical iterative algorithm based on Newton-Raphson (NR) method to determine the optimal number of activated antennas for XL-MIMO systems, which maximizes the approximated EE expression.

In order to compute the average ZF SINR expression, one can directly evaluate from eq. (G.30):

$$\begin{aligned}
\mathbb{E}\left[\overline{\gamma}_k^{(\mathrm{ZF})}\right] &= \mathbb{E}\left[\frac{P_{\max}}{K\sigma^2}\left(\sum_{m\in\mathcal{A}}\beta_{m,k} - \sum_{j=1,j\neq k}^{K}\frac{\sum_{m\in\mathcal{A}}\beta_{m,k}\beta_{m,j}}{\sum_{m\in\mathcal{A}}\beta_{m,j}}\right)\right] \\
&= \mathbb{E}\left[\frac{P_{\max}}{K\sigma^2}\left(\sum_{m\in\mathcal{A}}\beta_{m,k} - \sum_{m\in\mathcal{A}}\beta_{m,k}\sum_{j=1,j\neq k}^{K}\frac{\beta_{m,j}}{\sum_{n\in\mathcal{A}}\beta_{n,j}}\right)\right] \\
&= \mathbb{E}\left[\frac{P_{\max}}{K\sigma^2}\sum_{m\in\mathcal{A}}\beta_{m,k}\left(1 - \sum_{j=1,j\neq k}^{K}\frac{\beta_{m,j}}{\sum_{n\in\mathcal{A}}\beta_{n,j}}\right)\right] \\
&= \frac{P_{\max}}{K\sigma^2}\sum_{m\in\mathcal{A}}\mathbb{E}\left[\beta_{m,k}\right]\left(1 - \sum_{j=1,j\neq k}^{K}\mathbb{E}\left[\frac{\beta_{m,j}}{\sum_{n\in\mathcal{A}}\beta_{n,j}}\right]\right) \qquad \text{(G.41)}
\end{aligned}$$

in which the expectation is taken with respect to the random users' positions.

Instead of advancing with (G.41) seeking an exact solution, we approximate the average SINR by the SINR of a *user in the most expected position* (UMEP). Given the uniform distribution of the users as illustrated in Fig. G.1, this most expected position would be as depicted in Fig. G.2.

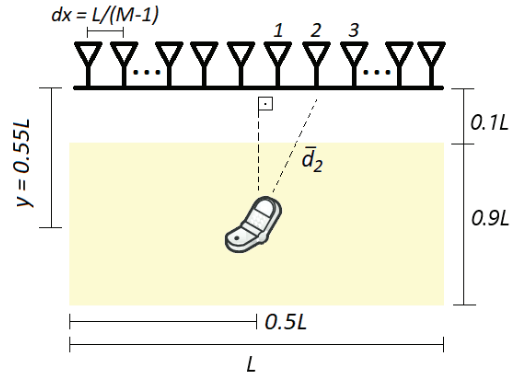


Fig. G.2: Illustration of the most expected user's position (UMEP).

Then, considering this position for the users, and noting that the HRNP AS activate in this case the M_s closest antennas, the ZF SINR expression becomes

$$\begin{aligned}
\mathbb{E}\left[\overline{\gamma}_k^{(\mathrm{ZF})}\right] &\approx \frac{P_{\max}}{K\sigma^2}\left(\sum_{m\in\mathcal{A}}\overline{\beta}_m - (K-1)\frac{\sum_{m\in\mathcal{A}}\overline{\beta}_m^2}{\sum_{m\in\mathcal{A}}\overline{\beta}_m}\right), \\
&\approx \frac{P_{\max}}{K\sigma^2}\left(2\sum_{m=1}^{M_s/2}\overline{\beta}_m - (K-1)\frac{2\sum_{m=1}^{M_s/2}\overline{\beta}_m^2}{2\sum_{m=1}^{M_s/2}\overline{\beta}_m}\right), \qquad \text{(G.42)}
\end{aligned}$$

with $\overline{\beta}_m = q\cdot(\overline{d}_m)^{-\kappa}$, and $\overline{d}_m = \sqrt{y^2+[(m-\frac{1}{2})dx]^2} = y\sqrt{1+[(m-\frac{1}{2})\frac{dx}{y}]^2} \approx y\sqrt{1+(m\frac{dx}{y})^2}$ is represented in Fig. G.2 for $m = 2$. Eq. (G.42) can thus be simplified as in the next page, in which from (G.48) to (G.49) we have used the binomial approximation $(1+x)^{\alpha} \approx 1+\alpha x$ for $|\alpha x| \ll 1$. In our scenario, this condition becomes

$$\frac{\kappa\, M_s^2\, dx^2}{8\, y^2} \ll 1, \tag{G.43}$$

which usually holds for typical XL-MIMO systems. For example, the binomial approximation results in relative errors lower than 5% for $|\alpha x| < 0.25$, which in our XL-MIMO scenario corresponds to $M_s < 225$. Besides, with this approximated ZF HRNP-AS SINR expression, we can also approximate the EE expression as in eq. (G.44). Moreover, by expanding all the power terms in the denominator of (G.44), as discussed in Section 2.3, and grouping them according to their dependence with M_s, we arrive at eq. (G.45), in which $\mathcal{P}_{\text{BC}} = \mathcal{P}_{\text{COD}} + \mathcal{P}_{\text{DEC}} + \mathcal{P}_{\text{BT}}$, and $\mathcal{T}_0$, $\mathcal{T}_1$ are defined in eq. (G.46) and (G.47), respectively.

$$\eta_e \approx \frac{BK \log_2\left(1+\mathbb{E}\left[\overline{\gamma}_k^{(\text{ZF})}\right]\right)}{P_{\text{TX}}^{\text{DL}} + P_{\text{TX}}^{\text{tr}} + P_{\text{CE}} + P_{\text{C/D}} + P_{\text{BH}} + P_{\text{PR}} + P_{\text{TC}} + P_{\text{FIX}}}, \tag{G.44}$$

$$\approx \frac{BK \log_2\left(1+\mathbb{E}\left[\overline{\gamma}_k^{(\text{ZF})}\right]\right)}{\mathcal{P}_{\text{BC}} BK \log_2\left(1+\mathbb{E}\left[\overline{\gamma}_k^{(\text{ZF})}\right]\right) + \mathcal{T}_0 + \mathcal{T}_1 M_s}. \tag{G.45}$$

$$\mathcal{T}_0 = \mathcal{P}^{\dagger} + \frac{3MK + M\log(M)}{T_{\text{LT}}\, \mathcal{L}_{\text{BS}}} + \frac{B\,K^3}{3\,\mathcal{S}\,\mathcal{L}_{\text{BS}}}. \tag{G.46}$$

$$\mathcal{T}_1 = \mathcal{P}_{\text{BS}} + \frac{5\,B\,K^2}{\mathcal{S}\,\mathcal{L}_{\text{BS}}} + \left(1-\frac{\tau}{\mathcal{S}}\right)\frac{2\,B\,K}{\mathcal{L}_{\text{BS}}} + \frac{B\,K}{\mathcal{S}\,\mathcal{L}_{\text{BS}}}. \tag{G.47}$$

4.1 Optimal Number of Activated Antennas

Considering our previous analytical results, we propose in this Section a method for obtaining the optimal M_s value when employing ZF with HRNP AS, by taking the derivative of eq. (G.45), with the SINR given in eq. (G.51), with respect to M_s, and equaling it to 0 when $M_s = M_s^*$. Following this procedure, and after some simplifications, we arrive at $f(M_s^*) = 0$, with $f(M_s)$ defined as

$$f(M_s) = \frac{\partial \mathbb{E}\left[\overline{\gamma}_k^{(\text{ZF})}\right]}{\partial M_s} - \frac{\mathcal{T}_1 \ln(2)\left(1+\mathbb{E}\left[\overline{\gamma}_k^{(\text{ZF})}\right]\right)\log_2\left(1+\mathbb{E}\left[\overline{\gamma}_k^{(\text{ZF})}\right]\right)}{\mathcal{T}_0 + \mathcal{T}_1 M_s}, \tag{G.55}$$

where $\frac{\partial \mathbb{E}\left[\overline{\gamma}_k^{(\text{ZF})}\right]}{\partial M_s}$ is given in (G.53).

Since $\mathbb{E}\left[\overline{\gamma}_k^{(\text{ZF})}\right]$ and its derivative are dependent of M_s, we cannot arrive at a closed-form expression for M_s^*. However, we can find the root of $f(M_s)$ by applying some iterative numerical method, like Newton-Raphson (NR)

$$\begin{aligned}\mathbb{E}\left[\overline{\gamma}_k^{(\mathrm{ZF})}\right] &\approx \frac{P_{\max}q}{K\sigma^2}\left(2\sum_{m=1}^{M_s/2}(\overline{d}_m)^{-\kappa} - (K-1)\frac{2\sum_{m=1}^{M_s/2}(\overline{d}_m)^{-2\kappa}}{2\sum_{m=1}^{M_s/2}(\overline{d}_m)^{-\kappa}}\right),\\ &\approx \frac{P_{\max}q}{K\sigma^2}\left\{2\sum_{m=1}^{M_s/2} y^{-\kappa}\left[1+\left(\frac{m\,dx}{y}\right)^2\right]^{-\frac{\kappa}{2}} - (K-1)\frac{2\sum_{m=1}^{M_s/2} y^{-2\kappa}\left[1+\left(\frac{m\,dx}{y}\right)^2\right]^{-\kappa}}{2\sum_{m=1}^{M_s/2} y^{-\kappa}\left[1+\left(\frac{m\,dx}{y}\right)^2\right]^{-\frac{\kappa}{2}}}\right\},\\ &\approx \frac{P_{\max}q\,y^{-\kappa}}{K\sigma^2}\left\{2\sum_{m=1}^{M_s/2}\left[1+\left(\frac{m\,dx}{y}\right)^2\right]^{-\frac{\kappa}{2}} - (K-1)\frac{2\sum_{m=1}^{M_s/2}\left[1+\left(\frac{m\,dx}{y}\right)^2\right]^{-\kappa}}{2\sum_{m=1}^{M_s/2}\left[1+\left(\frac{m\,dx}{y}\right)^2\right]^{-\frac{\kappa}{2}}}\right\}\ (ZF_{ME}),\end{aligned}\tag{G.48}$$

$$\approx \frac{P_{\max}q\,y^{-\kappa}}{K\sigma^2}\left\{2\sum_{m=1}^{M_s/2}\left[1-\frac{\kappa}{2}\left(\frac{m\,dx}{y}\right)^2\right] - (K-1)\frac{2\sum_{m=1}^{M_s/2}\left[1-\kappa\left(\frac{m\,dx}{y}\right)^2\right]}{2\sum_{m=1}^{M_s/2}\left[1-\frac{\kappa}{2}\left(\frac{m\,dx}{y}\right)^2\right]}\right\},\tag{G.49}$$

$$\begin{aligned}&\approx \frac{P_{\max}q\,y^{-\kappa}}{K\sigma^2}\Bigg[\left(\mathcal{T}_{1,1}M_s - \mathcal{T}_{1,2}M_s^2 - \mathcal{T}_{1,3}M_s^3\right)\\ &\qquad - (K-1)\frac{(\mathcal{T}_{2,1}M_s - \mathcal{T}_{2,2}M_s^2 - \mathcal{T}_{2,3}M_s^3)}{(\mathcal{T}_{1,1}M_s - \mathcal{T}_{1,2}M_s^2 - \mathcal{T}_{1,3}M_s^3)}\Bigg]\ (ZF_{BA}),\end{aligned}\tag{G.50}$$

$$\text{with}\quad \mathcal{T}_{1,1} = 1 - \frac{K\,dx^2}{12\,y^2},\quad \mathcal{T}_{1,2} = \frac{K\,dx^2}{8\,y^2},\quad \mathcal{T}_{1,3} = \frac{K\,dx^2}{24\,y^2},\tag{G.51}$$

$$\mathcal{T}_{2,1} = 1 - \frac{K\,dx^2}{6\,y^2},\quad \mathcal{T}_{2,2} = \frac{K\,dx^2}{4\,y^2},\quad \mathcal{T}_{2,3} = \frac{K\,dx^2}{12\,y^2}.$$

method, which obtains a sequence of M_s values $M_{s,0}, M_{s,1}, M_{s,2}, \ldots M_{s,n}$ converging to M_s^* if the starting point $M_{s,0}$ is not too far from it. The values in the sequence obey

$$M_{s,n} = M_{s,n-1} - \frac{f(M_{s,n-1})}{\left.\frac{\partial f(M_s)}{\partial M_s}\right|_{M_{s,n-1}}},\tag{G.56}$$

in which the derivative of $f(M_s)$ is given in (G.52).

5 Numerical Results and Discussion

Our adopted simulation parameters are indicated in Table G.1. While we have chosen very similar power consumption parameters than that of [16], [17], the XL-MIMO system parameters are chosen similarly as [3–5], as well as in accordance with common XL-MIMO scenario applications. Considering $M = 512$ antennas at the XL-MIMO BS, Fig. G.3 depicts the SINR, sum SE and the energy efficiency as a function of number of users K (from 1 to $M/2$), for both CB and ZF precoders. The sum SE is presented in units of bits per

$$\frac{\partial f(M_s)}{\partial M_s} = \frac{\partial^2 \mathbb{E}\left[\overline{\gamma}_k^{(\mathrm{ZF})}\right]}{\partial M_s^2} - \frac{\mathcal{T}_1^2 \ln(2)\left(1+\mathbb{E}\left[\overline{\gamma}_k^{(\mathrm{ZF})}\right]\right)\log_2\left(1+\mathbb{E}\left[\overline{\gamma}_k^{(\mathrm{ZF})}\right]\right)}{(\mathcal{T}_0+\mathcal{T}_1 M_s)^2}$$
$$- \frac{\mathcal{T}_1\left(\mathcal{T}_0+\mathcal{T}_1 M_s\right)\frac{\partial \mathbb{E}\left[\overline{\gamma}_k^{(\mathrm{ZF})}\right]}{\partial M_s}\left(1+\ln(2)\log_2\left(1+\mathbb{E}\left[\overline{\gamma}_k^{(\mathrm{ZF})}\right]\right)\right)}{(\mathcal{T}_0+\mathcal{T}_1 M_s)^2}, \tag{G.52}$$

$$\text{with} \quad \frac{\partial \mathbb{E}\left[\overline{\gamma}_k^{(\mathrm{ZF})}\right]}{\partial M_s} = \frac{P_{\max} q\, y^{-\kappa}}{K\sigma^2}\left[\mathcal{F}_1' - (K-1)\frac{\mathcal{F}_1\mathcal{F}_2' - \mathcal{F}_2\mathcal{F}_1'}{\mathcal{F}_1^2}\right], \tag{G.53}$$

$$\text{and} \quad \frac{\partial^2 \mathbb{E}\left[\overline{\gamma}_k^{(\mathrm{ZF})}\right]}{\partial M_s^2} = \frac{P_{\max} q\, y^{-\kappa}}{K\sigma^2}$$
$$\left[\mathcal{F}_1'' - (K-1)\frac{\mathcal{F}_1^2\left(\mathcal{F}_1\mathcal{F}_2'' - \mathcal{F}_2\mathcal{F}_1''\right) - 2\left(\mathcal{F}_1\mathcal{F}_2' - \mathcal{F}_2\mathcal{F}_1'\right)\mathcal{F}_1\mathcal{F}_1'}{\mathcal{F}_1^4}\right], \tag{G.54}$$

in which $\mathcal{F}_1 = \mathcal{T}_{1,1}M_s - \mathcal{T}_{1,2}M_s^2 - \mathcal{T}_{1,3}M_s^3, \quad \mathcal{F}_2 = \mathcal{T}_{2,1}M_s - \mathcal{T}_{2,2}M_s^2 - \mathcal{T}_{2,3}M_s^3,$
$\mathcal{F}_1' = \mathcal{T}_{1,1} - 2\mathcal{T}_{1,2}M_s - 3\mathcal{T}_{1,3}M_s^2, \quad \mathcal{F}_2' = \mathcal{T}_{2,1} - 2\mathcal{T}_{2,2}M_s - 3\mathcal{T}_{2,3}M_s^2,$
$\mathcal{F}_1'' = -2\mathcal{T}_{1,2} - 6\mathcal{T}_{1,3}M_s, \quad \mathcal{F}_2'' = -2\mathcal{T}_{2,2} - 6\mathcal{T}_{2,3}M_s.$

channel use (bpcu). One can note that ZF precoding always achieve a higher total energy efficiency than CB in the scenario investigated. The presented results were averaged among 1000 random realizations of the users' positions. It is also shown in the Figure the equivalence between the results of performance expressions from [5], eq. (G.12) and (G.13), and the expressions with our proposed simplifications, eq. (G.21) and (G.28).

Now, considering $M = 500$ antennas at the XL-MIMO BS, and the same power consumption parameters, Fig. G.4 shows the SINR, sum SE and the EE as a function of $M_s \in \{100; M\}$, with $K = 100$ users, for both CB and ZF precoders when employing the HRNP AS scheme. Notice that ZF precoding achieves a higher total energy efficiency than CB in the scenario investigated. Besides, by activating a number of $M_s = 146$ BS antennas, one can attain the maximum total energy efficiency for ZF precoder with $K = 100$ users ("M_s^* by NR" point in Fig. G.4.c), as found by our proposed NR method of Section 4.1. Fig. G.4 also compares the performance obtained by averaging eq. (G.30) with several random realizations for the users' positions (denoted as ZF), with the approximated deterministic result from eq. (G.48), denoted as ZF$_{ME}$, and with the binomial approximation in eq. (G.51), denoted as ZF$_{BA}$. It also shows the results in terms of sum SE and EE of the XL-MIMO system. One can conclude that both proposed approximations are tight, and that the M_s values that maximize them are nearly the same.

Next, in order to obtain the performance results of GA, LS, and PSO-based

Table G.1: Simulation Parameters.

Parameter	Value
Carrier frequency: f	2.6 GHz
Number of BS antennas M	$[500; 512]$
XL-MIMO array length: L	30 m
Distance of users to BS:	$[0.1 \cdot L, L]$
Path loss decay exponent: κ	3
Path loss at the reference distance: q	$10^{-3.53}$
Transmission bandwidth: B	20 MHz
Channel coherence bandwidth: B_C	100 kHz
Channel coherence time: T_C	2 ms
Long-term fading coherence time: T_{LT}	2 s
Total noise power: σ^2	-96 dBm
UL pilot transmit power: ρ_{p}	20 mW
DL radiated power: $P_{\max} = \frac{\rho\sigma^2}{qL^{-\kappa}}$	0.23 mW
Coherence block: $\mathcal{S}$	200 symbols
Length of the uplink pilot signals: τ	K
Computational efficiency at BSs: $\mathcal{L}_{\text{BS}}$	12.8 $\left[\frac{\text{Gflops}}{\text{W}}\right]$
Fraction of DL transmission: ζ^{d}	1
Fraction of UL transmission: ζ^{u}	0
PA efficiency at the BS: η^{d}	0.39
PA efficiency at the MTs: η^{uT}	0.50
Fixed power consumption: P_{FIX}	18 W
Power for local oscillators at BSs: P_{SYN}	2 W
Power for circuit components BSs: P_{BS}	1 W
Power for circuit components MTs: P_{MT}	0.10 W
Power density for coding data: $\mathcal{P}_{\text{COD}}$	0.10 $\left[\frac{\text{W}}{\text{Gbit/s}}\right]$
Power density for decoding data: $\mathcal{P}_{\text{DEC}}$	0.80 $\left[\frac{\text{W}}{\text{Gbit/s}}\right]$
Power density for backhaul traffic: $\mathcal{P}_{\text{BT}}$	0.25 $\left[\frac{\text{W}}{\text{Gbit/s}}\right]$

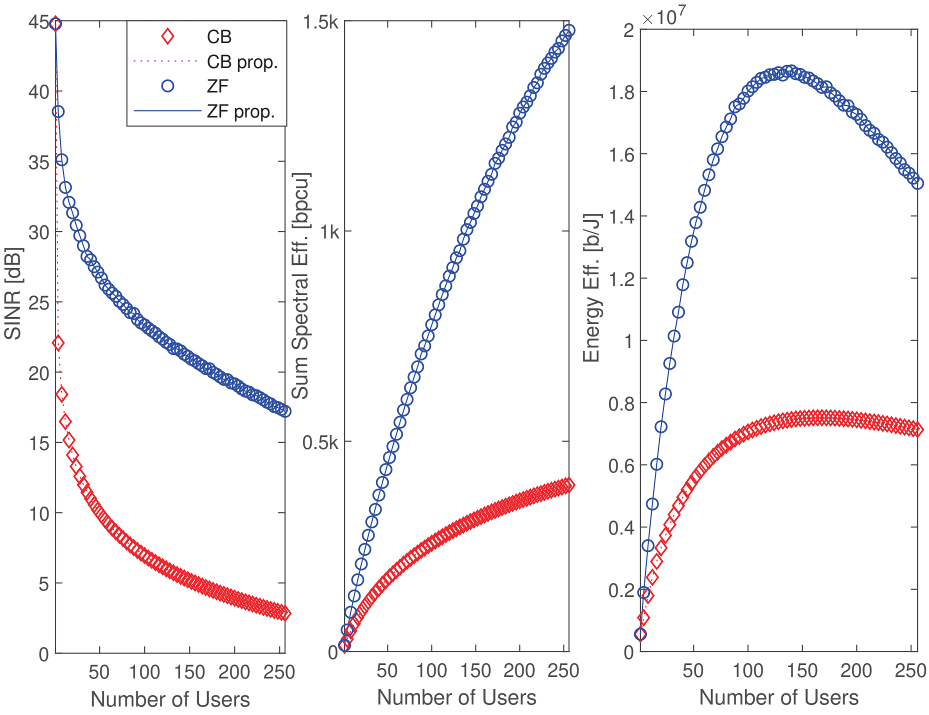
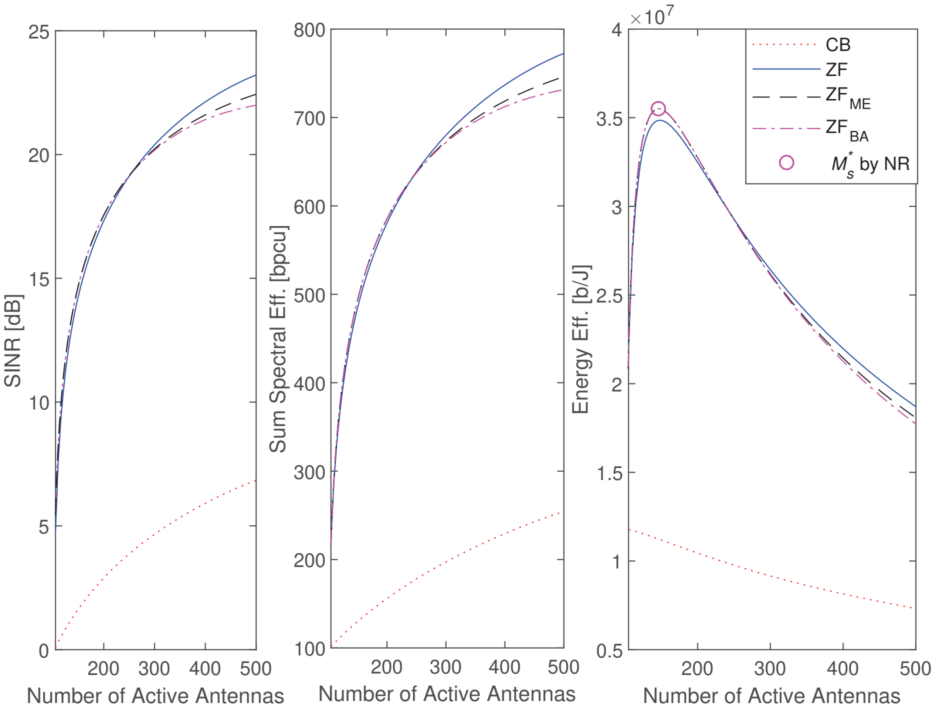


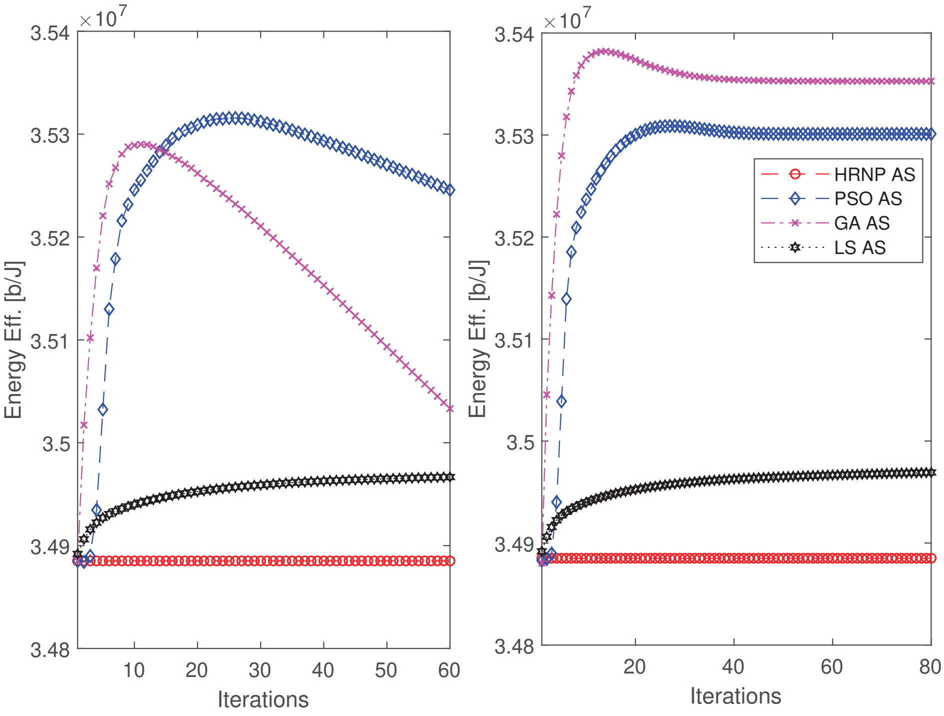
Fig. G.3: (a) SINR, (b) sum-SE, and (c) EE *vs.* K for $M = 512$ antennas, selecting all available antennas. Proposed eq. (G.21) and (G.28), are represented by dotted and solid line curves, respectively, while the performances from [5], eq. (G.12) and (G.13), are indicated by the curves with '◊' and 'o' markers.



(a) SINR (b) sum SE (c) EE

Fig. G.4: HRNP-AS scheme under ZF and CB precoders: (a) SINR, (b) sum SE, and (c) EE as a function of M_s for $M = 500$ antennas and $K = 100$.

AS schemes, we have set the maximum number of iterations $N_{it}^{\max} = 60$ for such schemes, and analysed their convergence for $K = 100$ users, as depicted in Fig. G.5.a. One can see from the Figure that the LS-AS convergence presents a non-decreasing behavior, since when a new solution is not found in certain iteration, the algorithm interrupts its search, and does not spend more processing power. On the other hand, for GA and PSO-based AS for XL-MIMO systems, if the algorithms do not find new solutions and keep searching during additional iterations, the EE of that solution decreases due to the progressive processing power consumed in the subsequent iterations. Therefore, it is not efficient to predefine the number of iterations for these two schemes in the XL-MIMO antenna selection problem, since in this optimization problem it would be very difficult do adjust the number of iterations in such a way to obtain a suitable EE solution for the algorithms. To circumvent while taking advantage of this feature, we implement an *early-interruption* criterion, in which if the GA or the PSO-based AS schemes do not find a new solution within 5 iterations, the search is interrupted, obtaining the convergences depicted in Figure G.5.b. Besides, for the GA-based AS scheme, we have considered a population size of $M/2$, of which 10% are selected as parents at each iteration, and a mutation probability of 2%. For the PSO-based one, we have considered a swarm of $M/5$ particles, and an inertia weight, cognitive factor and social factor of 0.5.

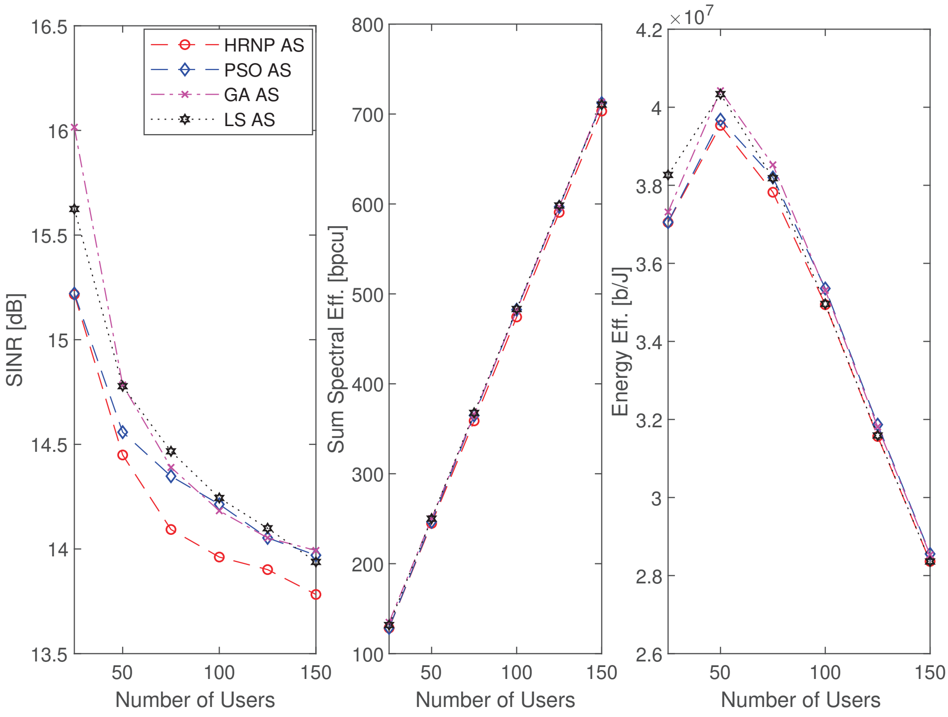


(a) without early-interruption

(b) with early-interruption

Fig. G.5: Convergence of the AS schemes: (a) without, and (b) with early-interruption stopping search criterion. $K = 100$ and $M = 500$ antennas.

Fig. G.6 depicts the SINR, sum SE, and EE as a function of K for the HRNP, GA, LS, and PSO-based AS schemes employing ZF precoding, with $M = 500$ antennas at the XL-MIMO array. While the achieved sum SE performance is nearly the same for all investigated schemes, the graphs reveal that SINR and EE gains can be achieved in comparison with HRNP. The Figure also shows that, in terms of SINR and EE, the GA, LS, and PSO-based AS schemes achieve a similar performance, and their gains in comparison with HRNP AS are small, since the processing required for finding a suitable antennas subset in the XL-MIMO system increases the energy consumption; thus, the EE gains become marginal. Except for small number of users, the GA AS scheme achieves one of the best EEs in most part of the investigated scenario, although for high number of users, its performance becomes very similar to HRNP AS scheme. Besides, due to its simplicity and celerity to return the results, one can point out that the HRNP criterion coupled to the NR procedure for M_s^* selection represents a very promising XL-MIMO AS scheme.

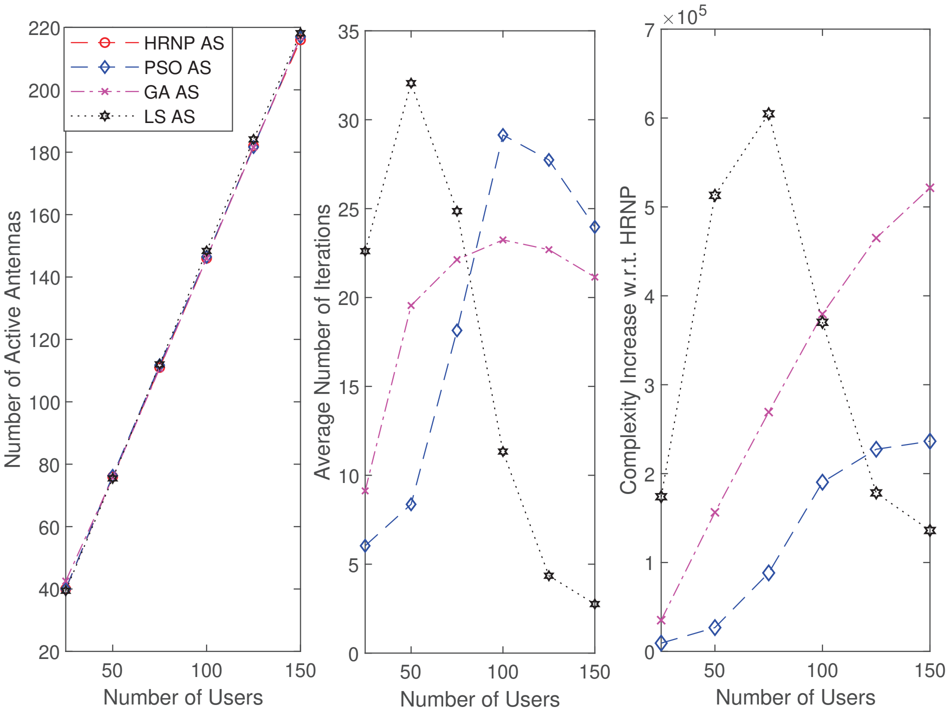


(a) SINR (b) sum SE (c) EE

Fig. G.6: AS schemes for ZF precoding: (a) SINR, (b) sum SE, and (c) EE as a function of K for $M = 500$ antennas.

5.1 Complexity of XL-MIMO AS Methods

Fig. G.7.a depicts the average number of active antennas as a function of K for the investigated AS methods. One can see that the M_s^* value obtained by our proposed NR method usually matches the number of antennas selected by LS, PSO, and GA-based AS schemes, corroborating the tightness of the approximations made and the effectiveness of the method. The major advantage of our proposed NR method for obtaining M_s^* is that it can be evaluated for any system configuration satisfying eq. (G.43). In our numerical simulations, the method has converged in at most 3 iterations from the starting point $M_{s,0} = 1.5K$. Besides, the M_s^* value is not dependent on the channel coefficients, but only on the system parameters, like number of users, transmit power, dimensions of XL-MIMO array and coverage area. Therefore, once found M_s^*, the NR method just has to be evaluated again when one of these parameters change. The fixed complexity of evaluating M_s^* under 3 NR iterations is about 380 *flops*, which is negligible in comparison with that of selecting the antennas subset, eq. (G.37), (G.38), (G.39), and (G.40), besides of remaining valid for larger time periods.



(a) Active Antennas (b) # Iterations (c) Relative Complexity

Fig. G.7: (a) Average number of active antennas, (b) average number of iterations, and (c) complexity increase of the AS schemes w.r.t. HRNP AS approach as a function of K for $M = 500$ antennas.

Fig. G.7.b depicts the average number of iterations required by each investigated AS scheme, recalling that the number of iterations are not fixed, since the LS interrupts when a new solution is not find in an iteration, and GA and PSO implement the early-interruption criterion. Besides, due to the non-decreasing behavior of the LS convergence depicted in Fig. G.5, the average number of iterations for this scheme in Fig. G.7.b does not correspond to the point in which the LS convergence curve becomes horizontal. Besides, the advantage of HRNP criterion in selecting antennas within the XL-MIMO array can also be confirmed by the extra computational complexity required for the other analysed methods. Hence, considering the average number of iterations from Fig. G.7.b, the *relative complexity increment* of the LS, GA and PSO AS schemes w.r.t. the HRNP AS method are depicted in Fig. G.7.c. The relative complexity increment metric is defined as:

$$\Delta_{\mathcal{C}} = \frac{\mathcal{C}_{\text{as}}^{\text{LS, GA,PSO}} - \mathcal{C}_{\text{as}}^{\text{HRNP}}}{\mathcal{C}_{\text{as}}^{\text{HRNP}}}$$

considering typical XL-MIMO network configurations for K users and M BS antennas. One can confirm the very large relative complexity increase of the AS methods for XL-MIMO, *i.e.*, this complexity increment is in the order of 10^5, which make the benefits they would bring less significant in terms of energy efficiency.

It is noteworthy that the computational complexity spent with the AS methods is included in the EE values, in terms of the processing power. In summary, the performance improvement of the AS scheme comes at the expense of high complexity, which results in marginal EE gains. On the other hand, the HRNP-AS procedure is able to achieve an improved EE of 34.85 Mbit/J for $K = 100$ users, in comparison with 18.71 Mbit/J of selecting all antennas, *i.e.*, not applying any AS procedure, corresponding in a 86.3% of EE increasing, as one can infer from Fig. G.4.

Elaborating further regarding the dependence of the optimal number of selected antennas M_s^* on the system parameters, such as number of users, total transmit power available, dimensions of XL-MIMO array, and coverage area, one can argue that such system parameters vary quite slowly with respect to the data symbol period. Therefore, it could be possible to evaluate the proposed AS scheme, and turning-on the optimal number of RF chains M_s^*, which are then switched to the best antenna subset according to our proposed HRNP criterion. Notice that only when the number of users changes significantly that it would be necessary to re-evaluate the (G.55)-(G.56), and then turning-on or turning-off some RF chains. Besides, the proposed method for finding the optimal number of selected antennas can provide very useful information for XL-MIMO system designers.

6 Conclusion

In this paper, we have investigated the XL-MIMO systems subject to channel non-stationarities. First, we have revisited the performance expressions from [5], and proposed to incorporate the power constraint at the SINR expressions of CB and ZF to arrive at more lean and comprehensive results.

Then, based on such obtained expressions, we have proposed four XL-MIMO AS schemes aiming at maximizing the EE based on the following criteria: HRNP, LS, GA, and PSO. Some simplifying assumptions allowed us to derive closed-form EE expressions, based on which we proposed a NR iterative method to obtain the optimal number of active antennas. Numerical results have shown that GA usually achieves one of the best EEs, although the gains were marginal in comparison with HRNP, since the processing required for achieving a suitable antennas subset increases the consumed energy, limiting the achieved EE gains. Thus, due to its simplicity and celerity in returning results, the proposed HRNP-AS scheme, with the NR method providing the optimal subarray size value M_S^*, can be seen as a very promising solution for AS XL-MIMO systems, achieving an EE gain of 86.3% in comparison with selecting all antennas strategy.

References

[1] T. Marzetta, E. Larsson, H. Yang, and H. Ngo, *Fundamentals of Massive MIMO*. New York, NY, USA: Cambridge University Press, 2016.

[2] E. Bjornson, J. Hoydis, and L. Sanguinetti, “Massive MIMO Networks: Spectral, Energy, and Hardware Efficiency,” *Foundations and Trends in Signal Processing*, vol. 11, no. 3-4, pp. 154–655, 2017.

[3] A. Amiri, C. N. Manchón, and E. de Carvalho, “Deep learning based spatial user mapping on extra large MIMO arrays,” *arXiv. 2002.00474*, 2020.

[4] X. Yang, F. Cao, M. Matthaiou, and S. Jin, “On the Uplink Transmission of Multiuser Extra-large Scale Massive MIMO Systems,” *arXiv. 1909.06760*, 2019.

[5] A. Ali, E. D. Carvalho, and R. W. Heath, “Linear Receivers in Non-Stationary Massive MIMO Channels With Visibility Regions,” *IEEE Wireless Communications Letters*, vol. 8, no. 3, pp. 885–888, June 2019.

[6] S. Chen, S. Sun, G. Xu, X. Su and Y. Cai, "Beam-Space Multiplexing: Practice, Theory, and Trends, From 4G TD-LTE, 5G, to 6G and Beyond," *IEEE Wireless Communications*, vol. 27, no. 2, pp. 162–172, April 2020.

[7] E. Björnson, L. Sanguinetti, H. Wymeersch, J. Hoydis, and T. L. Marzetta, "Massive MIMO is a reality – What is next?: Five promising research directions for antenna arrays," *Digital Signal Processing*, vol. 94, pp. 3–20, 2019.

[8] E. de Carvalho, A. Ali, A. Amiri, M. Angjelichinoski, and R. W. Heath, “Non-Stationarities in Extra-Large Scale Massive MIMO,” *arXiv. 1903.03085v2*, Oct 2019.

[9] A. Amiri, S. Rezaie, C. N. Manchón, and E. de Carvalho, “Distributed Receivers for Extra-Large Scale MIMO Arrays: A Message Passing Approach,” *arXiv. 2007.06930*, July 2020.

[10] V. C. Rodrigues, A. Amiri, T. Abrão, E. de Carvalho, and P. Popovski, "Low-Complexity Distributed XL-MIMO for Multiuser Detection," 2020 IEEE International Conference on Communications Workshops (ICC Workshops), Dublin, Ireland, 2020.

[11] H. Wang, A. Kosasih, C. Wen, S. Jin and W. Hardjawana, "Expectation Propagation Detector for Extra-Large Scale Massive MIMO," *IEEE Transactions on Wireless Communications*, vol. 19, no. 3, pp. 2036–2051, March 2020.

[12] Y. Han, S. Jin, C.-K. Wen, and X. Ma, "Channel Estimation for Extremely Large-Scale Massive MIMO Systems," *IEEE Wireless Communications Letters*, pp. 1–5, 2020.

[13] J. Flordelis, X. Li, O. Edfors and F. Tufvesson, "Massive MIMO Extensions to the COST 2100 Channel Model: Modeling and Validation," *IEEE Transactions on Wireless Communications*, vol. 19, no. 1, pp. 380–394, Jan. 2020.

[14] Y. Fu, C. Wang, X. Fang, L. Yan and S. Mclaughlin, "BER Performance of Spatial Modulation Systems Under a Non-Stationary Massive MIMO Channel Model," *IEEE Access*, vol. 8, pp. 44547–44558, Feb. 2020.

[15] O. S. Nishimura, J. C. Marinello, and T. Abrão, "A Grant-based Random Access Protocol in Extra-Large Massive MIMO System," *IEEE Communications Letters*, Early Access, July 2020.

[16] J. C. Marinello, C. Panazio, T. Abrão, and S. Tomasin, "Total Energy Efficiency of TR-MRC and FD-MRC Receivers for Massive MIMO Uplink," *IEEE Systems Journal*, vol. 13, no. 3, pp. 2285–2296, Sep. 2019.

[17] E. Björnson, L. Sanguinetti, J. Hoydis, and M. Debbah, "Optimal Design of Energy-Efficient Multi-User MIMO Systems: Is Massive MIMO the Answer?" *IEEE Transactions on Wireless Communications*, vol. 14, no. 6, pp. 3059–3075, June 2015.

[18] B. Makki, A. Ide, T. Svensson, T. Eriksson, and M. Alouini, "A Genetic Algorithm-Based Antenna Selection Approach for Large-but-Finite MIMO Networks," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 7, pp. 6591–6595, July 2017.

[19] Z. Liu, W. Du, and D. Sun, "Energy and Spectral Efficiency Tradeoff for Massive MIMO Systems With Transmit Antenna Selection," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 5, pp. 4453–4457, May 2017.

[20] J. Hoydis, S. ten Brink, and M. Debbah, "Massive MIMO in the UL/DL of Cellular Networks: How Many Antennas Do We Need?" *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 2, pp. 160–171, February 2013.

[21] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," `http://cvxr.com/cvx`, Mar. 2014.

[22] G. H. Golub and C. F. V. Loan, "Matrix Computations," Maryland, USA: Johns Hopkins University Press, 1996.

Paper H

Quasi-Distributed Antenna Selection for Spectral Efficiency Maximization in Subarray Switching XL-MIMO Systems

João Henrique Inacio de Souza, Abolfazl Amiri, Taufik Abrão,
Elisabeth de Carvalho, Petar Popovski

The paper has been published in the
IEEE Transactions on Vehicular Technology (TVT), vol. 70, no. 7, pp.6713–6725,
2021.

Abstract

In this paper, we consider the downlink (DL) of a zero-forcing (ZF) precoded extra-large scale massive MIMO (XL-MIMO) system. The base-station (BS) operates with limited number of radio-frequency (RF) transceivers due to high cost, power consumption and interconnection bandwidth associated to the fully digital implementation. The BS, which is implemented with a subarray switching architecture, selects groups of active antennas inside each subarray to transmit the DL signal. This work proposes efficient resource allocation (RA) procedures to perform joint antenna selection (AS) and power allocation (PA) to maximize the DL spectral efficiency (SE) of an XL-MIMO system operating under different loading settings. Two meta-heuristic RA procedures based on the genetic algorithm (GA) are assessed and compared in terms of performance, coordination data size and computational complexity. One algorithm is based on a quasi-distributed methodology while the other is based on the conventional centralized processing. Numerical results demonstrate that the quasi-distributed GA-based procedure results in a suitable trade-off between performance, complexity and exchanged coordination data. At the same time, it outperforms the centralized procedures with appropriate system operation settings.

Keywords**—** Extra-large scale massive MIMO (XL-MIMO), antenna selection (AS), resource allocation (RA), genetic algorithm (GA), distributed signal processing

1 Introduction

The benefits of adopting a high number of antennas at the base-station (BS) have attracted the interest on the massive MIMO transceiver design for the multi-antenna wireless communications systems beyond the fifth generation (B5G) and of the sixth generation (6G). The main advantages are the large array gain, inter-channel orthogonality and channel hardening. Also, increasing the number of antenna elements can enhance the cell coverage, improving the quality-of-service (QoS) of the border-cell users [1].

When the BS array attains extreme physical dimensions to support crowded scenario locations, such as airports and large shopping malls, the system is classified as extra-large scale massive MIMO (XL-MIMO) [2]. The XL-MIMO array provides the benefits of massive MIMO with additional beam-forming resolution due to the large array aperture [3]. The XL-MIMO array is characterized by key changes in the electromagnetic propagation conditions when compared to the conventional spatial stationary massive MIMO regime. The first property is the spherical wavefront propagation feature for the received signal due to the distance between the BS and the users being less than the Rayleigh distance [4]. Second, each cluster of scatterers sees only a portion of the array. Thus, the transmitted signal by each user reaches a small group of antennas, which comprises the visibility region (VR) of this user [2]. Additionally, the different propagation paths experienced along the array result in variations on the average received power. Results in [5, 6] demonstrate that the spatial non-stationarities produced by these two properties limit the performance of the system in terms of spectral efficiency (SE) unless an appropriated signal processing technique is applied.

Despite the benefits of high numbers of antennas, the XL-MIMO scenario imposes challenges for transceiver design. The first of them is the high cost and power consumption of fully digital implementations, which require one radio-frequency (RF) transceiver per antenna element [7, 8]. In addition, adopting a large number of antennas demands a high interconnection bandwidth to transmit the baseband data throughout the links to the BS processing unit. This turns into a serious implementation bottleneck, since the required bandwidth can not be handled by the current radio interfaces [9, 10]. Lastly, handling the complexity of signal processing techniques is a relevant issue, since the number of executed operations in linear detectors, such as zero-forcing (ZF) and minimum mean-squared error (MMSE), scales with the number of antennas [11].

In order to design practical BS architectures, one can limit the number of RF transceivers to cope with the cost constraints. The implementation with a limited the number of RF transceivers can benefit from the large array by adopting techniques such as antenna selection (AS) and hybrid precoding. Often, hybrid precoding design is associated with the solution of intricate optimization problems [12]. In addition, the commonly employed analog phase shifters are more expensive and consume more power than conventional on-off switches [8]. For these reasons, combining the AS procedures with linear precoding designs result in attainable strategies aiming at robust and effective implementations. Different approaches and tools can be adopted to perform AS, such as convex optimization [7, 13, 14], greedy heuristics [7, 15], machine learning [16] and metaheuristics [17–20].

One strategy to combat the problem of high interconnection bandwidth is to use hierarchical architectures. Adding multiple processing units to handle small groups of antennas and choosing the right signal processing methods can reduce significantly the amount of exchanged information in the regime of asymptotic number of antennas, as discussed in [9, 10]. However, the coordination of such processing units to perform different signal processing and resource allocation (RA) tasks constitutes a big challenge. In addition, many of these activities rely on the knowledge of fully reliable channel state information (CSI), which is hard to attain due to the high array dimensions. Many works on channel estimation [21], precoding and data detection [9, 10, 22–25] in massive and XL-MIMO consider distributed pre-processing at local nodes. However, studies on the distributed RA strategies, mainly involving AS, are scarce.

The signal processing complexity is an important concern in XL-MIMO due to the high number of antenna elements. However, differently from the conventional massive MIMO, the XL-MIMO can benefit from the spatial non-stationarities adopting local signal processing strategies to treat the signals inside the VRs at the BS' subarrays with reduced complexity [22, 24].

1.1 Literature Review

AS strategies for MIMO systems are extensively discussed in the literature. One AS algorithm to improve capacity in low rank matrix channels on point-to-point MIMO was first introduced in [26]. Later, the capacity distribution of systems with receive AS has been derived in [27]. These results were extended to massive MIMO regime in [28] and [29]. In these papers, the authors derived capacity bounds for systems with transmit and receive AS, respectively.

The authors in [13, 14] proposed AS procedures respectively for the channel capacity and downlink (DL) sum-capacity maximization based on the convex optimization framework. One technique based on the branch-and-bound algorithm is used in [8]. Considering linearly-precoded systems, the problems of AS for SE and sum-SINR maximization are addressed respectively in [15, 30]. Differently, the work in [31] analyzed one joint AS and power allocation (PA) procedure in a system with spatially distributed antennas. The proposed procedure runs at each antenna with side-information shared within its neighborhood. Besides, AS considering limited connections in the RF transceivers switching matrices is examined in [7].

On the other hand, there are only a few works that consider the AS problem for the XL-MIMO systems. A spatial users mapping procedure to maximize SE implemented with convolutional neural networks (CNN) is proposed in [16]. The aim is to determine each effective subarray window to precode the users signals using ZF. Results demonstrate that the CNN-based procedure achieves SE values comparable to the optimal mapping algorithm. In [17], several transmit AS procedures to maximize the energy efficiency (EE) from the long-term fading coefficients are proposed. Asymptotic SINR expressions for the received signal with AS are derived. Since the derived optimization problem is NP-hard, three of the proposed procedures are implemented by metaheuristic techniques, one being the genetic algorithm (GA). The GA is a powerful evolutionary metaheuristic that was used in different contexts to solve AS problems, as it is considered in [18–20].

1.2 Contribution

Motivated by the benefits of large numbers of antennas at the BS and the restricted number of RF transceivers, this work examines the joint AS and PA problem on the DL of a linearly-precoded XL-MIMO system. Differently from other papers adopting AS strategy, a distributed BS signal processing architecture is considered and the AS procedures are characterized in terms of the exchanged information between the processing nodes. Furthermore, we extend part of the results of [17] with the proposition of AS algorithms for XL-MIMO that use the short-term fading coefficients instead of the long-term ones. Additionally, we address the problem of joint AS and PA in XL-MIMO sub-arrays using a decentralized RA algorithm. The proposed RA algorithm uses the Sherman-Morrison-Woodbury (SMW) formula to perform optimal power allocation (OPA) and AS in a decentralized fashion.

The BS is constituted by multiple non-overlapping subarrays with dedicated remote processing units (RPUs), which perform independently channel estimation, precoding calculation and RA, mainly AS and PA. Each subarray is equipped with a fixed number of antenna elements and RF transceivers. Using the ZF precoding, the optimization goal is to maximize the SE subjected to the constraints of subarrays connections and maximum transmitted power.

The contribution of this work is fourfold. ***i***) Description of a distributed transceiver design for XL-MIMO based on a subarray switching architecture; ***ii***) proposition of a centralized procedure based on the evolutionary heuristic GA to perform joint AS and

PA to maximize the SE with subarray connection and maximum transmitted power constraints; ***iii***) proposition of a distributed version of the GA procedure for joint AS and PA which achieves performance tight to the centralized one but with low-size coordination data and less number of executed operations; ***iv***) extensive analyses of the proposed procedures in terms of number of symbols for training, coordination data size and number of floating point operations per second (flops).

The numerical results corroborate the GA-based procedures in achieving high performance, specifically in crowded XL-MIMO applications. Additionally, the decentralized GA version offers a good trade-off between performance, number of operations and coordination data size, outperforming the centralized procedures by adopting proper settings.

The rest of the paper is organized as follows. In Section 2 is described the system model, including the distributed subarrays processing at the BS. Next, in Section 3 are described the centralized and distributed GA-based optimization procedures for joint AS and PA in XL-MIMO systems, while Section 4 discusses two feasible AS procedures adopted as a result of decoupling the joint AS and PA optimization problem. Section 5 examines the complexity of the proposed algorithms. Extensive numerical results are discussed in Section 6. Final comments and conclusions are provided in Section 7.

1.3 Notation

Boldface small $\mathbf{a}$ and capital $\mathbf{A}$ letters represent respectively vectors and matrices. Capital calligraphic letters $\mathcal{A}$ represent finite sets, and $|\mathcal{A}|$ denotes the cardinality of the set $\mathcal{A}$. $\mathbf{I}_n$ denotes the identity matrix of size n. $\{\cdot\}^T$ and $\{\cdot\}^H$ denote respectively the transpose and the conjugate transpose operators. $\text{diag}(\cdot)$, $\text{tr}(\cdot)$ and $\det(\cdot)$ denote respectively the diagonal matrix, trace and determinant operators. $\lceil\cdot\rceil$ denotes the greatest integer operator. $\binom{n}{k}$ denotes the binomial coefficient. $\mathcal{CN}(\mu, \sigma^2)$ is a circularly symmetric complex Gaussian distribution with mean μ and variance σ^2. $\mathbb{E}[\cdot]$ denotes the expectation operator.

2 System Model

Consider the DL of a narrow-band multi-user XL-MIMO system with the BS equipped with M antennas and N RF transceivers serving K single-antenna users, as is depicted in Fig. H.1. During the DL, the BS uses η_{tr} symbols to perform channel estimation and η_{data} symbols to transmit the payload. We assume that the time interval used to send the total DL symbols $\eta_{\text{DL}} = \eta_{\text{tr}} + \eta_{\text{data}}$ is less than the channel coherence time.

The array in the BS is composed of B independent subarrays, each with M_b antennas and $N_b < M_b$ RF transceivers. The subarrays are equipped with a RPU to perform, in a distributed way, channel estimation, precoding calculation and RA tasks, specially AS and PA procedures. In addition, the BS has a central processing unit (CPU) to coordinate the subarrays operation. Fig. H.2 depicts all the described BS blocks.

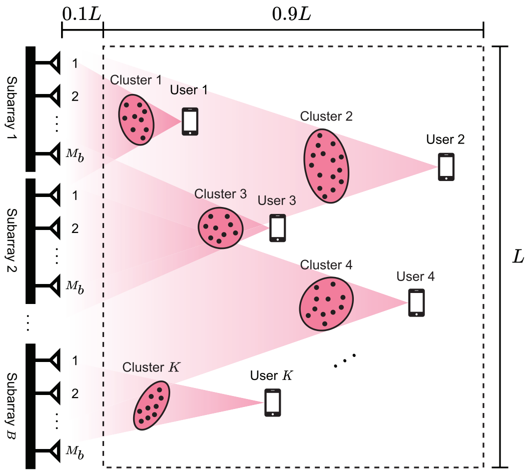


Fig. H.1: XL-MIMO system deployed inside a square cell with size L. The BS is a ULA with M antennas divided into B subarrays of M_b antennas each one. The K users are randomly distributed at a distance in the range $(0.1L, L)$ from the array.

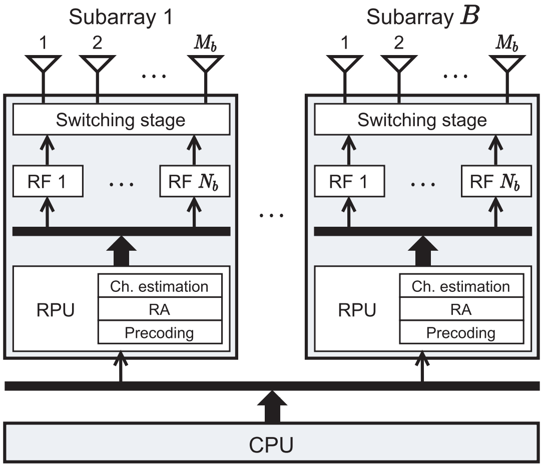


Fig. H.2: Diagram of the BS architecture for DL. The BS array is composed by B subarrays containing M_b antennas, N_b RF transceivers and one RPU. Additionally, the BS has a CPU for subarrays coordination.

Assumption 1 (Subarray switching stage): A flexible switching stage is implemented in each XL subarray. This stage allows every antenna of the subarray i to connect to any RF transceiver of it. Results in [7] demonstrate that partially connected architectures introduce lower insertion loss than fully-flexible matrices, which allows the connection of any antenna in the entire array to any RF transceiver.

We assume that each subarray has perfect knowledge of the channel coefficients associated to its antennas. See [21] for details on channel acquisition in distributed signal processing architectures. Besides, we deploy the ZF precoder to decode signals in each subarray. We adopt the technique in [21] to calculate the ZF precoder with low interconnection traffic, splitting the computations between the RPUs and the CPU.

2.1 Channel Model

In the XL-MIMO scenario, spatial non-stationarities arise due to the large array physical dimensions and number of antenna elements. Such non-stationarities are addressed in the adopted channel model as the variation of the mean received power along the array, as in [17, 22]. The path-loss coefficient associated to the BS antenna m and the user k is defined as

$$\beta_{m,k} = q_0 d_{m,k}^{-\kappa} \tag{H.1}$$

where q_0 is the path-loss attenuation at a reference distance, $d_{m,k}$ is the distance between the antenna m and the user k and κ is the path-loss exponent.

Let $\mathbf{R}_k \in \mathbb{C}^{M\times M}$, $\mathbf{R}_k = \text{diag}([\beta_{1,k} \ \cdots \ \beta_{M,k}]^T)$ be the matrix with the long-term fading coefficients of the user k. The channel vector of the user k is defined as

$$\mathbf{h}_k = \mathbf{R}_k^{\frac{1}{2}} \mathbf{h}'_k \tag{H.2}$$

where $\mathbf{h}'_k \in \mathbb{C}^{M\times 1}$, $\mathbf{h}'_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_M)$ is the short-term fading vector. From the users channel vectors, the channel matrix $\mathbf{H} \in \mathbb{C}^{M\times K}$ is defined as

$$\mathbf{H} = \begin{bmatrix} \mathbf{h}_1 & \cdots & \mathbf{h}_K \end{bmatrix} = \begin{bmatrix} \underline{\mathbf{h}}_1^T & \cdots & \underline{\mathbf{h}}_M^T \end{bmatrix}^T \tag{H.3}$$

considering $\underline{\mathbf{h}}_m \in \mathbb{C}^{1\times K}$ as the channel vector with the coefficients associated to the antenna m.

During the DL, the BS activates a group of antennas represented by the set $\mathcal{S} \subseteq \{1, \ldots, M\}$ such that $|\mathcal{S}| \leq N$. A partition of the set $\mathcal{S}$, *i.e.* $\{\mathcal{S}_b\}$, $\forall b = 1, \ldots, B$, contains the index of the selected antennas in the subarray b. This set is defined such that $|\mathcal{S}_b| \leq N_b$ $\forall b$, meeting the adopted subarray structure. The equivalent channel matrix of the active antennas is defined as a row-wise submatrix of $\mathbf{H}$, $\mathbf{H}_{\mathcal{S}} \in \mathbb{C}^{|\mathcal{S}|\times K}$. Similarly, the matrix $\mathbf{H}_{\mathcal{S}_b} \in \mathbb{C}^{|\mathcal{S}_b|\times K}$ contains only the channel vectors related to the active antennas in the subarray b.

Let $D_m \in \{0, 1\}$, $\forall m = 1, \ldots, M$ be an indicator equal to 1 if the antenna m is active during the DL and 0 otherwise. These indicators form the diagonal matrix $\mathbf{D} = \text{diag}([D_1 \ \cdots \ D_M]^T)$. During the precoding and SE computations, it is required to calculate the matrix product $\mathbf{H}_{\mathcal{S}}^H \mathbf{H}_{\mathcal{S}}$ of the active antennas channel matrix. Intended to enable this computation by the distributed signal processing architecture, the Gramian matrix is defined as in the following.

Remark 1 (Gramian matrix): Let $\mathbf{G}_m = \underline{\mathbf{h}}_m^H \underline{\mathbf{h}}_m$, $\forall m = 1, \ldots, M$ be the Gramian matrix associated with the BS antenna m. The set $\mathcal{M}_b$ is defined for $b = 1, \ldots, B$ as the group of antennas in the subarray b. The Gramian matrix associated to the b-th subarray includes only the active antennas inside it, and it can be written as

$$\mathbf{G}_{\mathcal{S}_b} = \mathbf{H}_{\mathcal{S}_b}^H \mathbf{H}_{\mathcal{S}_b} = \sum_{m\in\mathcal{M}_b} D_m \mathbf{G}_m \tag{H.4}$$

Similarly, the array Gramian matrix considering only the active antennas is defined as

$$\mathbf{G}_{\mathcal{S}} = \mathbf{H}_{\mathcal{S}}^H \mathbf{H}_{\mathcal{S}} = \sum_{m=1}^{M} D_m \mathbf{G}_m \tag{H.5}$$

An upper bound for the system performance considering the active antennas in the set $\mathcal{S}$, namely the DL sum-capacity, is calculated by [14]:

$$\begin{aligned} C_{\text{DPC}} &= \max_{\mathbf{P}} \log_2 \det\left(\mathbf{I}_K + \frac{1}{\sigma_z^2}\mathbf{P}\mathbf{H}_{\mathcal{S}}^H\mathbf{H}_{\mathcal{S}}\right) \\ &= \max_{\mathbf{P}} \log_2 \det\left(\mathbf{I}_K + \frac{1}{\sigma_z^2}\mathbf{P}\mathbf{G}_{\mathcal{S}}\right) \end{aligned} \tag{H.6}$$

where σ_z^2 is the additive noise power, while $\mathbf{P} = \text{diag}\left([p_1 \ \cdots \ p_K]\right)$ denotes the matrix with the allocated power for each user. The powers $p_k, \ \forall k = 1, \ldots, K$ are defined in order to meet the total power constraint $\sum_{k=1}^{K} p_k = P_{\max}$. The DL sum-capacity is achieved by the *dirty paper coding* (DPC) precoder, which has prohibitive high-complexity for practical implementations.

2.2 Downlink Signal

The data signal transmitted by the BS is defined as $\mathbf{x} \in \mathbb{C}^{|\mathcal{S}|\times 1}$,

$$\mathbf{x} = \mathbf{F}\mathbf{P}^{\frac{1}{2}}\mathbf{s} \tag{H.7}$$

where $\mathbf{F}\in \mathbb{C}^{|\mathcal{S}|\times K}$ denotes the ZF precoding matrix, calculated by

$$\begin{aligned} \mathbf{F} &= \mathbf{H}_{\mathcal{S}}\left(\mathbf{H}_{\mathcal{S}}^H\mathbf{H}_{\mathcal{S}}\right)^{-1} \\ &= \mathbf{H}_{\mathcal{S}}\mathbf{G}_{\mathcal{S}}^{-1} \end{aligned} \tag{H.8}$$

$\mathbf{s} = [s_1 \ \cdots \ s_K]^T$ denotes the vector of modulated data symbols such that $\mathbb{E}\left[\|s_k\|_2^2\right] = 1, \ \forall k = 1, \ldots, K$ and $\mathbb{E}\left[s_k^* s_{k'}\right] = 0, \ \forall k \neq k'$. The allocated powers in (H.7) are calculated in order to meet the following power constraint

$$\text{tr}\left[\mathbf{P}\left(\mathbf{H}_{\mathcal{S}}^H\mathbf{H}_{\mathcal{S}}\right)^{-1}\right] = \text{tr}\left(\mathbf{P}\mathbf{G}_{\mathcal{S}}^{-1}\right) = P_{\max} \tag{H.9}$$

Therefore, the entries of $\mathbf{P}$ depend on the active antennas set $\mathcal{S}$ and the PA policy.

The signal received by the users in the DL is defined as $\mathbf{y} \in \mathbb{C}^{K\times 1}$,

$$\begin{aligned} \mathbf{y} &= \mathbf{H}_{\mathcal{S}}^H\mathbf{F}\mathbf{P}^{\frac{1}{2}}\mathbf{s} + \mathbf{z} \\ &= \mathbf{P}^{\frac{1}{2}}\mathbf{s} + \mathbf{z} \end{aligned} \tag{H.10}$$

where $\mathbf{z} \in \mathbb{C}^{K\times 1}$, $\mathbf{z} \sim \mathcal{CN}\left(\mathbf{0}, \sigma_z^2\mathbf{I}_K\right)$ denotes the additive noise vector.

Given the ZF precoding design, the system SE is calculated by

$$\text{SE} = \sum_{k=1}^{K} \log_2\left(1 + \frac{p_k}{\sigma_z^2}\right) \tag{H.11}$$

which is equivalent to the SE of K independent Gaussian channels with received signal-to-noise ratio (SNR) equal to $p_k/\sigma_z^2 \ \forall k$.

2.3 Optimal Power Allocation (OPA) Policy

The OPA policy is the one that solves the problem of maximizing the system SE at (H.11), subjected to the maximum power constraint in (H.9):

$$\underset{\mathbf{P}}{\text{maximize}} \quad \text{SE} = \sum_{k=1}^{K} \log_2 \left(1 + \frac{p_k}{\sigma_z^2}\right) \tag{H.12a}$$

$$\text{subject to} \quad \text{tr}\left[\mathbf{P}(\mathbf{H}_{\mathcal{S}}^H \mathbf{H}_{\mathcal{S}})^{-1}\right] \leq P_{\max} \tag{H.12b}$$

$$p_k \geq 0,\ \forall k = 1, \ldots, K \tag{H.12c}$$

The optimization problem in (H.12) is equivalent to the well-known PA problem on independent Gaussian channels. It has an analytical closed-form solution derived by the Lagrange multipliers method (water filling solution). The optimal power distribution is calculated by [32]:

$$p_k = \left(\mu \left[(\mathbf{H}_{\mathcal{S}}^H \mathbf{H}_{\mathcal{S}})^{-1}\right]_{k,k}^{-1} - \sigma_z^2\right)^{+} \tag{H.13}$$

where $(x)^{+} = \max(x, 0)$ and μ is a constant calculated by

$$\mu = \frac{1}{K}\left\{P_{\max} + \sigma_z^2 \text{tr}\left[(\mathbf{H}_{\mathcal{S}}^H \mathbf{H}_{\mathcal{S}})^{-1}\right]\right\} \tag{H.14}$$

If $p_k = 0$ for some user k, the PA problem including this user is not feasible. For this reason, the k-th user is deactivated and the power distribution is recalculated considering only the group of the remaining active users. This process must be repeated until a group of users which results in a feasible solution is found.

3 Algorithm for Joint Antenna Selection and Power Allocation

The problem of jointly selecting the antenna-elements of the BS and allocating appropriate power amounts to maximizing the ZF SE given the constraints of maximum RF transceivers, subarray connections, and maximum power is formulated as

$$\underset{\mathbf{D},\mathbf{P}}{\text{maximize}} \quad \text{SE} = \sum_{k=1}^{K} \log_2 \left(1 + \frac{p_k}{\sigma_z^2}\right) \tag{H.15a}$$

$$\text{subject to} \quad \sum_{m \in \mathcal{M}_b} D_m \leq N_b,\ \forall b \in \{1, \ldots, B\} \tag{H.15b}$$

$$\text{tr}\left[\mathbf{P}(\mathbf{H}^H \mathbf{D} \mathbf{H})^{-1}\right] \leq P_{\max} \tag{H.15c}$$

$$D_m \in \{0, 1\},\ \forall m \in \{1, \ldots, M\} \tag{H.15d}$$

$$p_k \geq 0,\ \forall k \in \{1, \ldots, K\} \tag{H.15e}$$

The objective function in (H.15a) is the system SE. The constraints (H.15b) are the subarray connections constraints, which allow the activation of a maximum of N_b RF transceivers in each subarray. Also, the constraint (H.15c) ensures that the maximum transmitted power is equal to or less than $P_{\max}$. Moreover, the constraints (H.15d) and (H.15e) define respectively the binary antenna association variables and non-negative allocated powers.

Since $\mathbf{D}$ is binary constrained, the problem (H.15) constitutes a non-convex combinatorial optimization problem. One approach to solve (H.15) comprises two steps: firstly, determining the optimal active antennas set via exhaustive search assuming equal PA; after that, given the result $\mathbf{D}^\star$ from the exhaustive search, the allocated power matrix $\mathbf{P}^\star$ is calculated adopting the OPA policy in (H.13).

The AS via exhaustive search considering the activation of all the RF transceivers requires testing $\binom{M_b}{N_b}^B$ candidate solutions, a number that attains prohibitive dimensions in the XL-MIMO regime. For instance, in a system with $B = 8$ subarrays equipped with $M_b = 64$ antennas and $N_b = 32$ RF transceivers, there is a number of feasible solutions on the order of 10^{146}. Testing all these solution candidates in a timely manner is impracticable. An efficient alternative to the exhaustive search is to perform a guided search along the feasible set using an intelligent metaheuristic procedure. In this way, a good quality solution can be obtained in feasible time testing only a few candidates.

3.1 Genetic Algorithm

One metaheuristic procedure adopted to solve many different combinatorial problems in wireless communications is the GA. This technique implements different search phases to efficiently explore the feasible set and exploit the good candidates properties in order to find promising regions in the feasible sub-spaces. Differently from exact optimization methods, evolutionary metaheuristics do not require convex objective functions or constraints. In addition, the execution complexity can be fitted to the available computational burden by adjusting the input parameters and number of iterations. Despite the advantages, the GA, as well as other metaheuristics, does not ensure finding the optimal solution.

As the GA is a procedure inspired by principles of genetics and natural selection, it inherited several terms from biology. To simplify understanding, Table H.1 contains a glossary of some common GA terms adopted throughout this work. In the following, the implemented GA procedures, phases and variables deployed to solve the problem (H.15) are briefly described.

Optimization variables encoding: The optimization variables of the problem (H.15) are the antennas state indicators D_m and the users allocated powers p_k. The powers p_k are determined by the OPA, eq. (H.13). Therefore, only the antennas indicators should be encoded as individuals. Thus, D_ms are defined as genes and the column vectors $[\underline{\mathbf{d}}_{i,b}]_m = D_m,\ \forall m \in \mathcal{M}_b,\ b = 1, \ldots, B$ containing the optimization variables w.r.t. each subarray represent the chromosomes, where i is the individual index. Every individual is defined by a vector $\mathbf{d}_i \in \{0,1\}^{M \times 1}$,

$$\mathbf{d}_i = \begin{bmatrix} \underline{\mathbf{d}}_{i,1}^T & \cdots & \underline{\mathbf{d}}_{i,B}^T \end{bmatrix}^T = \begin{bmatrix} D_1 & \cdots & D_M \end{bmatrix}^T \tag{H.16}$$

Table H.1: Glossary of the genetic algorithm terms

Parameter	**Description**
Individual	Candidate solution for the optimization problem
Population	Set of candidate solutions for the optimization problem
Offspring	Set of candidate solutions generated during an iteration
Gene	One optimization variable of the candidate solution
Chromosome	Set of optimization variables of the candidate solution
Generation	Genetic algorithm iteration
Fitness	Objective function of the optimization problem
Score	Value of the objective function for a candidate solution

Fitness function: The fitness function considered for the implementation is the ZF SE defined in (H.11), with the power distribution computed by the OPA policy.

The implemented GA contains the following phases: *a)* elitism, *b)* tournament selection, *c)* crossover and *d)* mutation. These phases require the definition of the parameters: population size N_p, number of individuals for elitism N_e, number of tournaments N_s, crossover probability p_c and mutation probability p_m. Each procedure is summarized in the sequel.

Elitism: The elitism aims to keep the best individuals of the current generation without change. At every generation, the N_e best individuals are chosen as the first individuals of the next generation. Elitism ensures that the SE obtained with the best AS indices of the GA iteration is always a non-decreasing value.

Tournament selection: During the tournament selection, the individuals are pairwise randomly compared according to their score values. The winners of the N_s tournaments become candidates for the crossover phase. The selection step compares the sets of AS indices produced at each GA iteration according to the SE achieved by them.

Crossover: The crossover phase aims to mix the chromosomes of the tournaments winners in order to obtain new solutions. This phase exploits the good properties of the current set of AS indices. Two tournament winners, named parent 1 and parent 2, are randomly selected to generate two new individuals. Each chromosome of child 1 has the probability p_c of being inherited from parent 1 and $1 - p_c$ from parent 2. Considering child 2, every chromosome has the probability p_c of being inherited from parent 2 and $1 - p_c$ from parent 1.

Mutation: The mutation phase aims to add random small changes at the offspring generated by crossover. This phase promotes the variability among the set of AS indices, exploring different regions of the feasible set. The chromosomes are mutated with probability p_m, when one random selected gene of the chromosome is flipped. To preserve the solutions' feasibility, the mutation phase is implemented by the scheme of Algorithm 17. The set $\mathcal{P}_c$ denotes the offspring generated during the crossover, and $\mathcal{P}_m$ is the offspring after mutation.

Convergence: There are several mechanisms to check the GA convergence. Herein,

Algorithm 17 Mutation procedure

```
Input: Crossover offspring $\mathcal{P}_c$ , $p_m, B, M_b, N_b$
Output: Mutated offspring $\mathcal{P}_m$
$\mathcal{P}_m \leftarrow \emptyset$
for $\mathbf{d}_i \in \mathcal{P}_c$ do
    for $b = 1 : B$ do
        if $rand\ uniform(0,1) \leq p_m$ then
            $k \leftarrow$ rand discrete uniform$(1, M_b)$
            if $[\underline{\mathbf{d}}_{i,b}]_m == 0$ and $\sum_{j=1}^{M_b} [\underline{\mathbf{d}}_{i,b}]_j == N_b$ then
                Go to line 5
            $[\underline{\mathbf{d}}_{i,b}]_m \leftarrow$ flip$([\underline{\mathbf{d}}_{i,b}]_m)$
    $\mathcal{P}_m \leftarrow \mathcal{P}_m \cup \mathbf{d}_i$
```

the implemented algorithm has two different criteria: the maximum number of generations $T_{\max}$ and the no improvement of the best score during the last T_{stall} generations.

Algorithm 18 summarizes the implemented procedure, named *genetic algorithm for resource allocation* (GA-RA). The set $\mathcal{P}_0$ denotes the initial population, $\mathcal{P}_t$ the population of the generation t, $\mathcal{P}_s$ the winners of the tournament selection and $\mathcal{P}_{\text{temp}}$ a temporary set for the elitism phase.

3.2 Quasi-Distributed Genetic Algorithm

The proposed GA-RA procedure requires the entire channel matrix $\mathbf{H}$ knowledge at the CPU to compute the individuals score values. Such requirement is unfeasible in the XL-MIMO scenario due to the high bandwidth to transfer all the channel coefficients associated to thousands of antennas to the CPU. For this reason, one solution that does not depend on the knowledge of full CSI at the CPU is preferable.

One solution to avoid the requirement of full knowledge of the $\mathbf{H}$ matrix consists of performing local AS at each subarray, considering fixed the AS indices in the other subarrays. The contribution of these fixed AS indices can be calculated previously by the CPU and transmitted to the RPUs with reduced bandwidth and processing power resources. Therefore, each subarray can selects its antennas using the GA. The proposed *quasi-distributed genetic algorithm for resource allocation* (DGA-RA) implements this concept and is presented in the following.

Analyzing the fitness function of the GA-RA procedure in (H.11), one can observe that it depends on the inverse of the array Gramian matrix, $\mathbf{G}_{\mathcal{S}}^{-1} = (\mathbf{H}_{\mathcal{S}}^H \mathbf{H}_{\mathcal{S}})^{-1}$. The computation of $\mathbf{G}_{\mathcal{S}}^{-1}$ can be done from the subarrays Gramian matrices by

$$\mathbf{G}_{\mathcal{S}}^{-1} = \left(\sum_{b=1}^{B} \mathbf{G}_{\mathcal{S}_b} \right)^{-1} \tag{H.17}$$

Therefore, the CPU can compute the inverse of the array Gramian matrix to calculate the GA-RA fitness function only with the subarrays Gramian matrices calculated locally at the RPUs. Each subarray Gramian matrix has K^2 entries, while the channel

Algorithm 18 GA-RA

```
Input: $N_p, N_e, N_s, p_c, p_m, T_{\text{stall}}, B, M_b, N_b, \mathbf{H}$
Output: The best selected antennas set, $\mathbf{D}^\star$
$\mathcal{P}_0 \leftarrow \emptyset$  $\mathcal{P}_0 \leftarrow \mathcal{P}_0 \cup \text{N-AS}(\mathbf{H})$ (Section 4.2)
for $i = 1 : N_p - 1$ do
    $\mathcal{P}_0 \leftarrow \mathcal{P}_0 \cup \text{rand individual}()$
for $t = 0 : T_{\max}$ do
    $\mathcal{P}_{t+1}, \mathcal{P}_s, \mathcal{P}_c \leftarrow \emptyset$  $\mathcal{P}_{\text{temp}} \leftarrow \mathcal{P}_t$
    for $i = 1 : N_e$ do Elitism
        $\mathbf{d}_e \leftarrow \underset{\mathbf{d}_j}{\text{argmax}}\ \text{score}(\mathbf{d}_j),\ \mathbf{d}_j \in \mathcal{P}_{\text{temp}}$  $\mathcal{P}_{t+1} \leftarrow \mathcal{P}_{t+1} \cup \mathbf{d}_e$  $\mathcal{P}_{\text{temp}} \leftarrow$
        $\mathcal{P}_{\text{temp}} \backslash \mathbf{d}_e$
    for $i = 1 : N_s$ do Tournament selection
        $\mathbf{d}_{s_1}, \mathbf{d}_{s_2} \leftarrow \text{rand}(\mathcal{P}_t)$  $\mathbf{d}_s \leftarrow \underset{\mathbf{d}_j}{\text{argmax}}\left[\text{score}(\mathbf{d}_{s_1}),\ \text{score}(\mathbf{d}_{s_2})\right]$  $\mathcal{P}_s \leftarrow \mathcal{P}_s \cup$
        $\mathbf{d}_s$
    for $i = 1 : N_e$ do Crossover
        $\mathbf{d}_{c_1}, \mathbf{d}_{c_2} \leftarrow \text{rand}(\mathcal{P}_s)$  $\mathbf{d}_{o_1}, \mathbf{d}_{o_2} \leftarrow \mathbf{0}_M$
        for $j = 1 : B$ do
            if rand uniform$(0,1) \leq p_c$ then
                $\underline{\mathbf{d}}_{o_1,j} \leftarrow \underline{\mathbf{d}}_{c_1,j}$  $\underline{\mathbf{d}}_{o_2,j} \leftarrow \underline{\mathbf{d}}_{c_2,j}$
            else
                $\underline{\mathbf{d}}_{o_1,j} \leftarrow \underline{\mathbf{d}}_{c_2,j}$  $\underline{\mathbf{d}}_{o_2,j} \leftarrow \underline{\mathbf{d}}_{c_1,j}$
        $\mathcal{P}_c \leftarrow \mathcal{P}_c \cup \mathbf{d}_{o_1} \cup \mathbf{d}_{o_2}$
    $\mathcal{P}_m \leftarrow \text{mutation}(\mathcal{P}_c)$ (Algorithm 17)
    $\mathcal{P}_{t+1} \leftarrow \mathcal{P}_{t+1} \cup \mathcal{P}_m$
    $\mathbf{d}^\star_{t+1} \leftarrow \underset{\mathbf{d}_i}{\text{argmax}}\ \text{score}(\mathbf{d}_i),\ \mathbf{d}_i \in \mathcal{P}_{t+1}$
    if $t > T_{\text{stall}}$ then Stall convergence criterion
        $\mathbf{d}_{\text{stall}} \leftarrow \underset{\mathbf{d}_i}{\text{argmax}}\ \text{score}(\mathbf{d}_i),\ \mathbf{d}_i \in \mathcal{P}_{t-T_{\text{stall}}}$  if $score(\mathbf{d}^\star_{t+1}) == score(\mathbf{d}_{\text{stall}})$
        then
            Break the loop
$\mathbf{D}^\star \leftarrow \text{diag}(\mathbf{d}^\star_{t+1})$  return $\mathbf{D}^\star$
```

matrix has MK. Therefore, calulating the contribution of the selected antennas at the CPU using the Gramian matrix strategy requires less bandwidth than by using the centralized strategy if $BK^2 < MK$ holds.

Based on (H.17), the DGA-RA procedure operates as follows. Initially, each subarray selects an active antennas set based on a simple criterion, such as the *norm-based antenna selection* (N-AS) described in the subsection 4.2. Then, the subarrays compute their Gramian matrices based on the selected set and transmit them to the CPU. At the CPU, the array Gramian matrix is computed by (H.17) and transmitted back to the subarrays. Afterwards, every subarray performs local antenna selection by a GA implementation, considering that the other subarrays are fixed. To evaluate the fitness function in eq. (H.11), the subarrays compute the array Gramian inverse matrix adopting the SMW formula for matrix inversion, as follows.

Remark 2 (SMW formula): The SMW formula [33] gives the inverse of the matrix $(\mathbf{A}+\mathbf{U}\mathbf{V}^H)$ from $\mathbf{A}^{-1}$, $\mathbf{U}$ and $\mathbf{V}$ by computing:

$$(\mathbf{A}+\mathbf{U}\mathbf{V}^H)^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}\left(\mathbf{I}+\mathbf{V}^H\mathbf{A}^{-1}\mathbf{U}\right)^{-1}\mathbf{V}^H\mathbf{A}^{-1} \tag{H.18}$$

Adopting this formulation, the array Gramian matrix can be calculated at the subarray b during the iteration n by letting

$$\mathbf{A}^{-1} = \left(\mathbf{G}_{\mathcal{S}}^{(n-1)}\right)^{-1}, \tag{H.19}$$

$$\mathbf{U} = \left[-\left(\mathbf{H}_{\mathcal{S}_b}^{(n-1)}\right)^H \quad \left(\mathbf{H}_{\mathcal{S}_b}^{(n)}\right)^H\right], \tag{H.20}$$

$$\mathbf{V}^H = \begin{bmatrix}\mathbf{H}_{\mathcal{S}_b}^{(n-1)} \\ \mathbf{H}_{\mathcal{S}_b}^{(n)}\end{bmatrix}, \tag{H.21}$$

where the superscript (n) denotes the variable during the n-th iteration of the DGA-RA procedure (proof in Appendix A).

After performing local AS, each subarray transmits their achieved SE values to the CPU. The CPU updates the AS indices of the subarray that has achieved the maximum SE values at the iteration n. Then, the CPU requests the subarray Gramian matrix of the updated subarray, and recalculates the inverse of the array Gramian matrix, $(\mathbf{G}_{\mathcal{S}}^{(n)})^{-1}$. The process can be executed iteratively following the scheme depicted in Fig. H.3.

The GA implemented in the DGA-RA procedure is similar to that one described in the Algorithm 18, except for some details at the optimization variables encoding

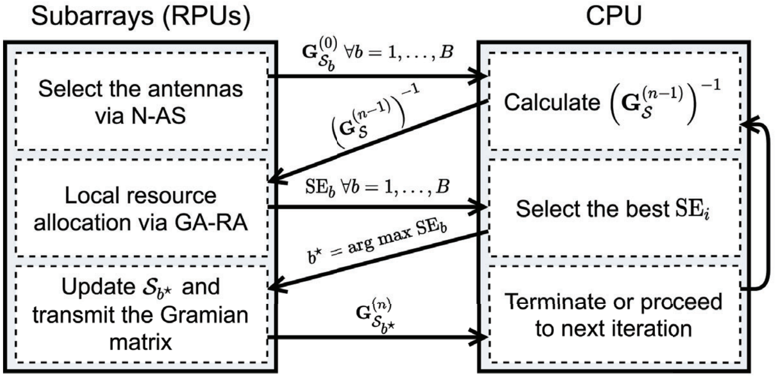

Subarrays (RPUs)
CPU
$\mathbf{G}_{\mathcal{S}_b}^{(0)}\ \forall b = 1,\dots,B$
Select the antennas via N-AS
Calculate $\left(\mathbf{G}_{\mathcal{S}}^{(n-1)}\right)^{-1}$
$\left(\mathbf{G}_{\mathcal{S}}^{(n-1)}\right)^{-1}$
Local resource allocation via GA-RA
$\mathrm{SE}_b\ \forall b = 1,\dots,B$
Select the best SE_i
$b^\star = \arg\max \mathrm{SE}_b$
Update $\mathcal{S}_{b^\star}$ and transmit the Gramian matrix
$\mathbf{G}_{\mathcal{S}_{b^\star}}^{(n)}$
Terminate or proceed to next iteration

Fig. H.3: Proposed DGA-RA procedure steps with coordination between the CPU and the RPUs. The superscript (n) denotes the n-th iteration.

and the crossover phase. About the individual encoding, the optimization variables at each subarray are reduced from M to M_b, since local AS is performed at each RPU. In addition, as the optimization variables consider only one subarray at each RPU, the individuals have two chromosomes: one represented by the first $M_b/2$ genes, and another composed by the remaining genes.

Due to this new chromosome definition, one further procedure after the crossover phase is required to preserve the solution's feasibility. The chosen method is to deactivate antennas of individuals with more than N_b antennas in a random fashion until they become feasible.

4 Antenna Selection Procedures

Two techniques to perform antenna selection are presented in the sequel, the DL *sum-capacity maximization antenna selection* (SCMAX-AS) and the N-AS method , proposed respectively in [14], [7]. The goal of solving only the antenna selection problem is to decouple the two RA problems associated to (H.15) aiming at obtaining tractable formulations.

4.1 Antenna Selection for DL Sum-Capacity Maximization

Firstly, we analize equal power allocation (EPA) strategy, *i.e.* $\mathbf{P} = \frac{P_{\max}}{K}\mathbf{I}_K$, intended to obtain a manageable optimization problem. The problem of selecting the set of active antennas in order to maximize the DL sum-capacity with the constraints of maximum number of RF transceivers and subarray connections is formulated as [14]:

$$\underset{\mathbf{D}}{\text{maximize}} \quad C_{\text{EPA}} = \log_2 \det\left(\mathbf{I}_K + \frac{P_{\max}}{K\sigma_z^2}\mathbf{H}^H\mathbf{D}\mathbf{H}\right) \tag{H.22a}$$

$$\text{subject to} \quad \sum_{m\in\mathcal{M}_b} D_m \leq N_b,\ \forall b \in \{1,\dots,B\} \tag{H.22b}$$

$$D_m \in \{0,1\},\ \forall m \in \{1,\dots,M\} \tag{H.22c}$$

Despite the concavity of the objective function in (H.22a) [13], the problem (H.22) is not convex due to the binary constraint in (H.22c). Hence, we define a convex relaxation of (H.22) by taking the variables D_m in the range $(0,1)$. This new problem, which can be solved with convex optimization tools, has the constraint (H.22c) replaced by

$$0 \leq D_m \leq 1,\ \forall m \in \{1,\dots,M\} \tag{H.23}$$

Notice that the solution of the convex relaxation results in non-binary values for the active antenna indicators D_m, which is outside the original problem domain.

One method for performing the antenna selection by solving the convex relaxation is to activate the N_b antennas with the highest D_m values at each subarray. This procedure is named in this work as SCMAX-AS, and is followed by the OPA policy in eq. (H.13). This AS procedure gives near-optimal results, except for $N \ll M$ [14]. Therefore, in a XL-MIMO system where the number of available RF transceivers is much less than the array antennas, the achieved system SE with the SCMAX-AS algorithm will be sub-optimal.

4.2 Norm-Based Antenna Selection (N-AS)

The N-AS procedure focus on selecting the subset of N_b antennas with the highest channel vector norm values [7]. We adopt this method to initiate the population of the GA-based procedures due to its low computational cost. The N-AS method solves the optimization problem formulated as

$$\begin{aligned} \underset{\mathbf{D}}{\text{maximize}} \quad & \Pi = \sum_{m=1}^{M} D_m \|\underline{\mathbf{h}}_m\|_2^2 && \text{(H.24a)} \\ \text{subject to} \quad & \sum_{m \in \mathcal{M}_b} D_m \leq N_b,\ \forall b \in \{1, \ldots, B\} && \text{(H.24b)} \\ & D_m \in \{0, 1\},\ \forall m \in \{1, \ldots, M\} && \text{(H.24c)} \end{aligned}$$

where the objective function consists of the sum of the squared norms of the channel vectors associated to the selected antennas.

The problem (H.24) can be solved quickly by selecting the N_b antennas with the highest channel vector norms at each subarray. After selection, the PA is performed by the OPA policy in (H.13).

5 Complexity Analysis

The complexity of the presented procedures is evaluated in terms of the number of symbols required for channel acquisition, the size of the coordination data exchanged between the RPUs and the CPU, and the number of flops during execution.

5.1 Training

In the following, we analyze the procedures in terms of training symbols for CSI acquisition. The length of the mutually orthogonal pilot signals used to estimate the channel vectors at the BS depends on: *a*) the number of users; *b*) the number of available RF transceivers; *c*) the number of antennas at the BS.

The number of symbols to acquire the entire channel matrix, required in all the presented procedures except in the N-AS, is $K \left\lceil \frac{M}{N} \right\rceil$. Particularly, the N-AS algorithm requires only the knowledge of the channel vector norms for selection. For this reason, the N-AS can be implemented without explicit channel estimation, supported by physical power-meters [21]. With this implementation, the N-AS requires a total of $2K$ symbols to operate. From this total, K symbols are required to estimate the norms of the channel vectors, and the remaining K symbols are used to estimate the channel vectors associated to the selected antennas.

5.2 Coordination Data Size

The coordination data is defined as the data originated at the RPUs that is required at the CPU during the RA procedures. Determining the coordination data size is

Table H.2: Coordination data exchanged between the RPUs and the CPU

Procedure	Implementation	Data type	Data size
GA-RA	Centralized	Channel matrix	MK
SCMAX-AS [14]	Centralized	Channel matrix	MK
N-AS [7]	Totally distributed	–	–
DGA-RA	Quasi-distributed	Gramian matrix	$(B + N_{\text{it}})K^2$

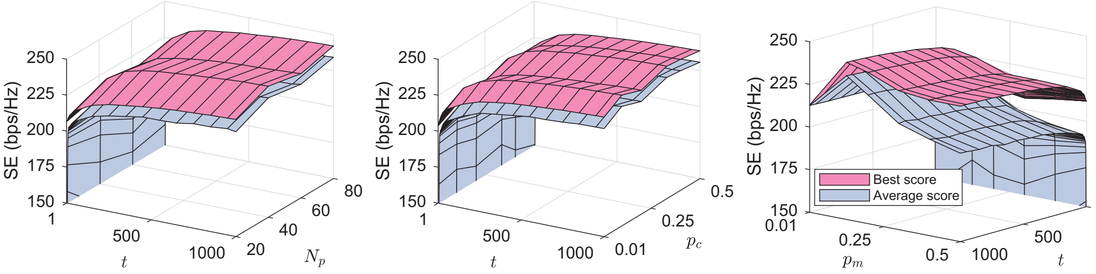
crucial since it can grow tremendously in the XL-MIMO scenario. In practical implementations, techniques as *data compression* helps alleviating the high interconnection bandwidth associated to the coordination data. However, such kind of consideration and optimization are out of the scope of this work.

Table H.2 contains the coordination data size associated to the considered RA procedures, detailing the type of required data in each one. The GA-RA and SCMAX-AS procedures require the entire channel matrix at the CPU, while the DGA-RA one relies on the subarrays Gramian matrices. On the other hand, the N-AS procedure does not require any CSI knowledge at the CPU for antenna selection purpose, being the most appealing technique in terms of the coordination data size.

5.3 Number of Flops

The third complexity metric is the number of flops executed by each procedure. The complexity analyses for the N-AS and the GA-based AS algorithms are as follows. The SCMAX-AS procedure is not considered due to the high complexity associated with computing the number of executed operations by the convex optimization solver.

N-AS: The operations executed at each subarray on the N-AS procedure consists of calculating the channel vectors' norms then sorting the obtained values to get the N_b largest ones. Assuming that the sorting operation has the complexity of the order



(a) Varying N_p (b) Varying p_c (c) Varying p_m

Fig. H.4: Convergence of the GA-RA with the number of generations t varying the GA input parameters N_p, p_c and p_m. The "best" and "average" SE surfaces are obtained over 20 realizations. In each plot, the values of the remaining input parameters are given in Table H.4.

$M_b \log(M_b)$, the per-subarray flops for N-AS is

$$\mathcal{C}_{\text{N-AS}} = M_b(2K-1) + M_b \log(M_b) \tag{H.25}$$

GA-RA: The complexity of the GA-RA method is dominated by the number of operations required for the evaluation of the GA fitness function, eq. (H.11). At the first iteration, the algorithm evaluate the fitness function for N_p individuals. During the remaining iterations, $(T-1)(N_p - N_e)$ fitness function evaluations are done, where T denotes the total number of generations.

As the OPA policy involves simple computations, the complexity of the fitness function is reduced to the inversion of the array Gramian matrix. The flops to compute the array Gramian matrix inverse is derived in Appendix B. From this result, the total flops for the GA-RA algorithm is

$$\mathcal{C}_{\text{GA-RA}} = \left[T(N_p - N_e) + N_e\right]\left(\frac{7}{3}K^3 + 2NK^2 - K^2\right) \tag{H.26}$$

DGA-RA: For the DGA-RA procedure, a similar approach to the one used for GA-RA can be followed. Despite that, the inverse of the array Gramian matrix is computed by the SMW formula, which is implemented with a different number of flops. The number of flops to obtain the inverse of the array Gramian matrix in the DGA-RA procedure is derived in Appendix C. Taking into account these differences and the fact that the DGA-RA procedure runs over N_{it} iterations, the total number of flops is given by:

$$\begin{aligned}\mathcal{C}_{\text{DGA-RA}} = N_{\text{it}} \left[T(N_p - N_e) + N_e\right] \times \\ \times \left[\frac{7}{3}N_b^3 + 2K^3 + N_b^2(4K-1) + \right. \\ \left. + K^2(4N_b - 2) + N_b^2(1-2K) + K\right]\end{aligned} \tag{H.27}$$

6 Numerical Results

The numerical evaluations of the proposed methods as well as the benchmark techniques are presented in this section. The simulation system parameters are given in Table H.3. The users are randomly located inside a square cell of size L, and the BS is equipped with a uniform linear array (ULA) positioned on one side of the cell, as depicted in Fig. H.1. Additionally, the users are random uniformly located at a distance in the range $(0.1L, L)$ from the array. Although the results in the following are obtained for the ULA, they can be easily extended to other array form factors, such as the uniform planar one.

Before comparing the proposed techniques, it is necessary to tune the GA-RA and DGA-RA GA input parameters in order to obtain a suitable performance-complexity tradeoff. The input parameter N_p, p_c and p_m values are selected using the iterated local search algorithm [34]. The number of individuals for elitism is equal to 10% of the population size, and the number of tournaments is defined in order to fill the population after the elitism phase. Additionally, the stall convergence criterion parameter

Table H.3: Simulation parameters

Parameter	**Value**
Cell size	$L = 30$ m
# Users	$K \in [1, 217]$
Maximum transmitted power	$P_{\max} = 230\ \mu$W
Path-loss at the reference distance	$q_0 = -35.3$ dB
Path-loss exponent	$\kappa = 3$
Noise power	$\sigma_z^2 = -96$ dBm
Uniform Linear Array (ULA) Setup	
# Antennas	$M \in [32, 2048]$
# RF transceivers	$N \in [64, 256]$
# Subarrays	$B = \{2, 4, 8\}$
# Antennas per subarray	$M_b = M/B$
# RF transceivers per subarray	$N_b = N/B$

Table H.4: Genetic algorithm parameters

Symbol	**Description**	**Parameter value**	
		GA-RA	**DGA-RA**
N_p	Population size	80	80
N_e	Elitism individuals	8	8
N_s	Tournaments	36	36
p_c	Crossover probability	0.33	0.35
p_m	Mutation probability	0.13	0.36
$T_{\max}$	Maximum generations	10^3	10^2
T_{stall}	Stall generations	300	30

is approximately 30% of the maximum number of generations. The selected parameters for the GA-based procedures are listed in Table H.4. Notice that the DGA-RA procedure is set to run 10 times less generations than the GA-RA, since the number of optimization variables decrease from M at the GA-RA to M_b in the DGA-RA procedure.

In Fig. H.4, the quality of convergence of the GA-RA procedure is corroborated varying the parameters N_p, p_c and p_m independently. Each surface is computed by averaging the achieved scores over 20 realizations. These results on the best and average SE scores among the generations t confirm the parameters' values adopted in Table H.4, while demonstrating a relative low tuning sensibility of the GA-RA convergence to the three input parameters.

Fig. H.5 depicts the system SE achieved by the proposed RA procedures versus the number of available RF transceivers. In addition to the proposed solutions, the SE attained by random AS scheme and using all the M antennas are plotted as the lower and upper performance bounds, respectively. The results consider $M = 512$, $B = 8$, $K = 50$ and $N_{\text{it}} \in \{5, 16\}$ for the DGA-RA procedure. Observing the Fig. H.5, one

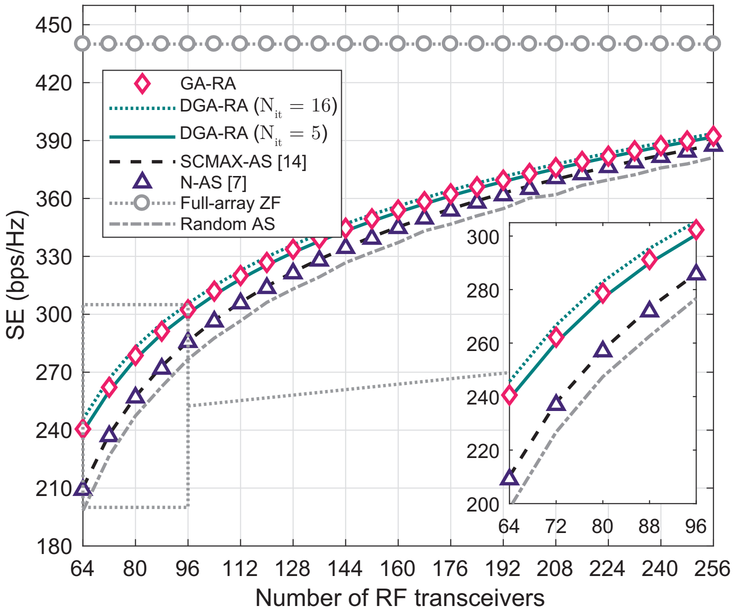


Fig. H.5: Comparison of SE *vs* the number of available RF transceivers. $M = 512$, $B = 8$, $K = 50$ and, for the DGA-RA procedure $N_{\text{it}} \in \{5, 16\}$.

realize that the GA-based procedures achieve better SE results than the other ones. In the sequence, there are respectively the SCMAX-AS and N-AS. As expected, all the performance curves are upper and lower bounded by the SE achieved using full-array ZF and random AS, respectively. The SE gap between the procedures decreases as the number of RF transceivers increases. Analyzing the GA-based procedures, the DGA-RA achieves SE values tight to the GA-RA running with only five iterations. However, setting $N_{\text{it}} = 16$ makes the DGA-RA system SE values outperform marginally the ones obtained by the GA-RA procedure. Therefore, the quasi-distributed procedure can achieve a performance comparable, or even better, to the fully centralized approach by adopting a sufficient number of iterations.

In the following, Fig. H.6 depicts the system SE achieved by the proposed RA procedures versus the number of users. These numerical results consider $M = 512$, $B = 8$, $N = 256$ and $N_{\text{it}} \in \{5, 16\}$ for the DGA-RA procedure. For better understanding, let $\mathcal{L} = K/N$ be the system effective loading factor. For all the proposed procedures, firstly the SE increases with K, assuming a decreasing behavior after a peak. This is due to the reduction of spatial degrees of freedom increasing the system loading factor, typically observed in linearly precoded systems [35]. Comparing the procedures, all of them get comparable SE values for a low loading factor. However, for high loading factor values, typically $\mathcal{L} = 0.6$, the GA-RA and DGA-RA procedures get substantial better results. Again, the DGA-RA outperforms the GA-RA in terms of SE by setting $N_{\text{it}} = 16$. Combining the results in Figs. H.5 and H.6, we conclude that the GA-based procedures perform with higher SE gains over the other available AS schemes [7, 14] in crowded XL-MIMO scenarios, *i.e.*, when the loading factor is high, $\mathcal{L} > 0.25$.

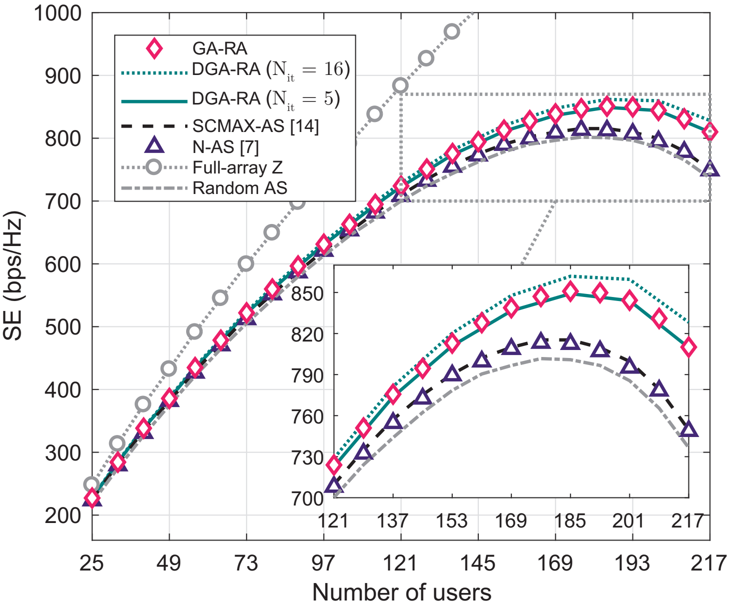


Fig. H.6: Comparison of SE *vs* the number of users. $M = 512$, $B = 8$, $N = 256$ and, for the DGA-RA procedure $N_{\text{it}} \in \{5, 16\}$.

6.1 Complexity Analysis

The numerical results in the following cover the computational complexity of the proposed procedures. In Fig. H.7 the coordination data size of the centralized procedures (GA-RA and SCMAX-AS) and the DGA-RA one versus the number of users is illustrated. The curves are evaluated by the expressions in Table H.2. The result considers $M \in \{512, 2048\}$ and,for the DGA-RA procedure, $N_{\text{it}} = 16$ and $B \in \{2, 4, 8\}$. Comparing the RA approaches when the number of users is low, the quasi-distributed one get lower coordination data sizes than the centralized procedures. For higher numbers of users, the coordination data size associated to DGA-RA acquires larger values than the obtained by the centralized procedures. This point of inversion of behavior depends on the numbers of antennas, subarrays and iterations w.r.t. the DGA-RA procedure. It is worth mentioning that the coordination data size grows quadratically with K for the DGA-RA procedure, while it grows linearly with K for the centralized RA procedure.

Fig. H.7 depicts the coordination data size of the centralized procedures and the DGA-RA one versus the number of antennas in the BS. The results consider $K = 50$ and, for the DGA-RA method, $N_{\text{it}} \in \{5, 16\}$ and $B \in \{2, 4, 8\}$. The coordination data size grows linearly with M in the centralized procedures, while for the DGA-RA procedure, it does not depend on M. In fact, this is the primary aim for choosing a distributed RA technique in XL-MIMO, in which the BS is equipped with an asymptotically high number of antennas.

The next results are related to the complexity in terms of flops. Fig. H.8 illustrates the number of flops per processing unit of the GA-based procedures versus the number of available RF transceivers. The curves are evaluated by the eqs. (H.26) and (H.27). Such results consider $K = 50$ and, for the DGA-RA procedure, $B = 8$ and

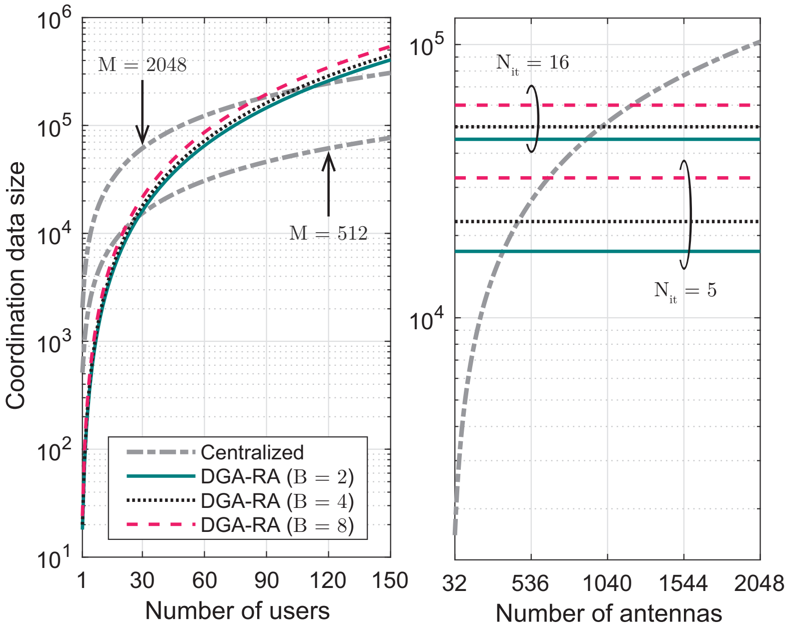


Fig. H.7: Coordination data size of the GA-based RA schemes *vs* the number of (a) users and (b) antennas. When it is not specified, $N_{\text{it}} = 16$ and $K = 50$.

$N_{\text{it}} \in \{1, 5, 16\}$. For low numbers of RF transceivers, the flops' values for the DGA-RA procedure are lower than the GA-RA algorithm. Again, after a point of inversion of behavior, the flops' values for GA-RA get lower than the ones for the quasi-distributed procedure. This point of changing of behavior decreases as N_{it} increases.

The curves with the number of flops per processing unit of the GA-based procedures versus the number of users are depicted in Fig. H.8. This result considers $N = 256$ and, for the DGA-RA procedure, $B = 8$ and $N_{\text{it}} = \{1, 5, 16\}$. For low numbers of users, the flops' values of the GA-RA procedure are lower than the ones get for the DGA-RA. However, this behavior inverts quickly, and the gap between the flops' values for both centralized and distributed procedures becomes constant. This constant behavior for large K is due to the fact that both eqs. (H.26) and (H.27) grow asymptotically with K^3.

7 Conclusions

This works proposes a subarray switching architecture for the BS antenna array, while examining the problem of joint AS and PA optimization aiming at maximizing the SE of XL-MIMO systems with limited number of RF transceivers. Two GA-based near-optimal and low-complexity procedures are proposed. One is the centralized GA-RA, designed to operate with the entire channel matrix available at the CPU. The other is the quasi-distributed DGA-RA, based on the subarrays Gramian matrices. Both evolutionary metaheuristic optimization methods are analysed in terms of achieved SE, coordination data size and flops , and compared with benchmarks, including two procedures from the literature, the SCMAX-AS and the N-AS followed by optimal PA.

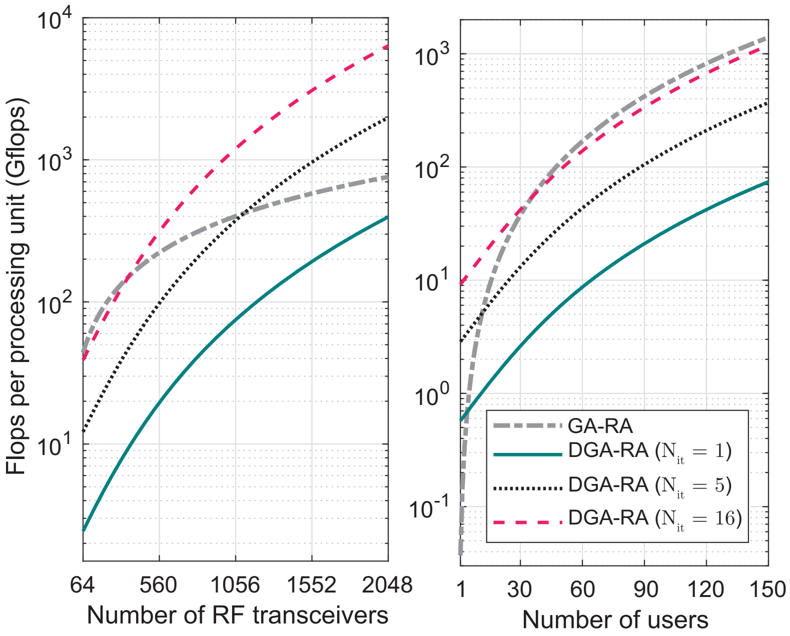


Fig. H.8: Flops per processing unit of the proposed GA-based procedures versus the number of (a) available RF transceivers and (b) users. $B = 8$ and, when it is not specified, $K = 50$ and $N = 256$.

Numerical results corroborate that the GA-based AS and PA procedures achieve high SE gains compared to the selected benchmarks, particularly in crowded XL-MIMO scenarios, *i.e.*, when the effective loading factor $\mathcal{L} > 0.25$. At the same time, the distributed DGA-RA method can outperform the other procedures with low-size coordination data and low computational complexity by taking the appropriate system operation settings.

Acknowledgment

This work was supported in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (CAPES) - Finance Code 001 (scholarship), in part by the Ministry of Science, Technology and Innovation (MCTIC), by the National Council for Scientific and Technological Development (CNPq) of Brazil under Grant 310681/2019-7, and in part by State University of Londrina – Paraná State Government (UEL) and in part by the Danish Council for Independent Research DFF-701700271.

A Local Computation of the Inverse of the Array Gramian Matrix via the Sherman - Morrison - Woodbury Formula

To compute the array Gramian matrix at the subarray b, the RPU must follow these two steps. Firstly, remove the contribution of the selected antennas at the subarray b at the iteration $n-1$. Then, add the contribution of the selected antennas at the iteration n. Therefore, it needs to compute the inverse of the array Gramian matrix by the expression

$$\left(\mathbf{G}_{\mathcal{S}}^{(n)}\right)^{-1} = \left(\mathbf{G}_{\mathcal{S}}^{(n-1)} - \mathbf{G}_{\mathcal{S}_b}^{(n-1)} + \mathbf{G}_{\mathcal{S}_b}^{(n)}\right)^{-1} \tag{H.28}$$

which evaluation would be straightforward if all the terms were available at the subarray.

However, the subarray needs to compute $(\mathbf{G}_{\mathcal{S}}^{(n)})^{-1}$ knowing only $(\mathbf{G}_{\mathcal{S}}^{(n-1)})^{-1}$ and the local channel vectors, *i.e.* $\underline{\mathbf{h}}_m\ \forall m \in \mathcal{M}_b$ for the subarray b. Writing the subarray Gramian matrices of (H.28) in terms of the local channel matrices results in

$$\begin{aligned} & -\mathbf{G}_{\mathcal{S}_b}^{(n-1)} + \mathbf{G}_{\mathcal{S}_b}^{(n)} \\ & \quad = -\left(\mathbf{H}_{\mathcal{S}_b}^{(n-1)}\right)^H \mathbf{H}_{\mathcal{S}_b}^{(n-1)} + \left(\mathbf{H}_{\mathcal{S}_b}^{(n)}\right)^H \mathbf{H}_{\mathcal{S}_b}^{(n)} \\ & \quad = \begin{bmatrix} -\left(\mathbf{H}_{\mathcal{S}_b}^{(n-1)}\right)^H & \left(\mathbf{H}_{\mathcal{S}_b}^{(n)}\right)^H \end{bmatrix} \begin{bmatrix} \mathbf{H}_{\mathcal{S}_b}^{(n-1)} \\ \mathbf{H}_{\mathcal{S}_b}^{(n)} \end{bmatrix} \end{aligned} \tag{H.29}$$

From (H.28) and (H.29), it is possible to define the SMW formula variables, $\mathbf{A}^{-1}$, $\mathbf{U}$ and $\mathbf{V}^H$, in terms of the available information at the subarray as the eqs. (H.19), (H.20) and (H.21), respectively.

B Flops to Compute the Inverse of the Array Gramian Matrix via the Cholesky Decomposition

Initially, the computation of the array Gramian matrix is done by solving the product in (H.5), which costs $2K^2N - K^2$ flops [33]. Afterwards, define the Cholesky decomposition of the array Gramian matrix as

$$\mathbf{G}_{\mathcal{S}} = \mathbf{L}\mathbf{L}^H \tag{H.30}$$

where $\mathbf{L}$ is a lower triangular matrix. The computation of $\mathbf{L}$ can be done with $K^3/3$ flops [33]. Then, each column of the inverse of the Gramian matrix can be computed solving the set of linear systems below by backforward substitution,

$$\mathbf{L}\mathbf{L}^H\mathbf{x} = \mathbf{e}_i,\ \forall i = 1,\ldots,K \tag{H.31}$$

where $\mathbf{e}_i$ denotes the canonical basis vector, *i.e.* a row vector with all entries equal to 0, except the entry i which is equal to 1. Each linear system can be solved with $2K^2$

flops [33], totaling $2K^3$ flops for all the columns of $\mathbf{G}_{\mathcal{S}}^{-1}$. Therefore, the total flops for the array Gramian matrix computation and inversion is equal to

$$\mathcal{C}_{Chol.} = \frac{7}{3}K^3 + 2NK^2 - K^2 \tag{H.32}$$

C Flops to Compute the Inverse of the Array Gramian Matrix via the Sherman-Morrison-Woodbury Formula

To count the flops to compute the matrix inversion by the SMW formula, the eq. (H.18) is decomposed in six parts. The computations involved in each part and their respective flops are organized in Table H.5. The flops in Table H.5 are counted assuming that the contribution of the selected antennas during the previous iteration is removed. Such assumption is reasonable since the expression in (H.28) can be done sequentially, by keeping only the terms $-\mathbf{G}_{\mathcal{S}_b}^{(n-1)}$ or $\mathbf{G}_{\mathcal{S}_b}^{(n)}$ at a time.

All the parts include only simple matrix multiplications and sums, except for the part $\mathbf{Q}_3$. This part can be efficiently computed by the Cholesky decomposition approach followed by the backforward substitution procedure described in Appendix B. Therefore, the total flops required to compute the inverse of the array Gramian matrix via the SMW formula is equal to

$$\begin{aligned}\mathcal{C}_{SMW} = \frac{7}{3}N_b^3 + 2K^3 + N_b^2(4K-1) \\ + K^2(4N_b-2) + N_b^2(1-2K) + K\end{aligned} \tag{H.33}$$

Table H.5: Flops involved on the Sherman-Morrison-Woodbury formula computation

Symbol	Expression	Number of flops
$\mathbf{Q}_1$	$\mathbf{V}^H\mathbf{A}^{-1}$	$2N_bK^2 - N_bK$
$\mathbf{Q}_2$	$\mathbf{I} + \mathbf{Q}_1\mathbf{U}$	$2N_b^2K - N_b^2 + N_b$
$\mathbf{Q}_3$	$\mathbf{Q}_2^{-1}$	$7/3N_b^3$
$\mathbf{Q}_4$	$\mathbf{U}\mathbf{Q}_3$	$2N_b^2K - N_bK$
$\mathbf{Q}_5$	$\mathbf{I} - \mathbf{Q}_4\mathbf{Q}_1$	$2N_bK^2 - K^2 + K$
$\mathbf{Q}_6$	$\mathbf{A}^{-1}\mathbf{Q}_5$	$2K^3 - K^2$

References

[1] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, “Massive MIMO for next generation wireless systems,” *IEEE Communications Magazine*, vol. 52, no. 2, pp. 186–195, Feb. 2014.

[2] E. D. Carvalho, A. Ali, A. Amiri, M. Angjelichinoski, and R. W. Heath, "Non-stationarities in extra-large-scale massive MIMO," *IEEE Wireless Communications*, vol. 27, no. 4, pp. 74–80, Aug. 2020.

[3] À. O. Martínez, E. De Carvalho, and J. Ø. Nielsen, "Towards very large aperture massive MIMO: A measurement based study," in *2014 IEEE Globecom Workshops*, Dec. 8–12 2014, pp. 281–286.

[4] Z. Zhou, X. Gao, J. Fang, and Z. Chen, "Spherical wave channel and analysis for large linear array in LoS conditions," in *2015 IEEE Globecom Workshops*, Dec. 6–10 2015, pp. 1–6.

[5] X. Li, S. Zhou, E. Björnson, and J. Wang, "Capacity analysis for spatially non-wide sense stationary uplink massive MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 14, no. 12, pp. 7044–7056, Dec. 2015.

[6] A. Ali, E. D. Carvalho, and R. W. Heath, "Linear receivers in non-stationary massive MIMO channels with visibility regions," *IEEE Wireless Communications Letters*, vol. 8, no. 3, pp. 885–888, Jun. 2019.

[7] A. Garcia-Rodriguez, C. Masouros, and P. Rulikowski, "Reduced switching connectivity for large scale antenna selection," *IEEE Transactions on Communications*, vol. 65, no. 5, pp. 2250–2263, May 2017.

[8] Y. Gao, H. Vinck, and T. Kaiser, "Massive MIMO antenna selection: Switching architectures, capacity bounds, and optimal antenna selection algorithms," *IEEE Transactions on Signal Processing*, vol. 66, no. 5, pp. 1346–1360, Mar. 2018.

[9] K. Li, R. R. Sharan, Y. Chen, T. Goldstein, J. R. Cavallaro, and C. Studer, "Decentralized baseband processing for massive MU-MIMO systems," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 7, no. 4, pp. 491–507, Dec. 2017.

[10] J. Rodríguez Sánchez, F. Rusek, O. Edfors, M. Sarajlić, and L. Liu, "Decentralized massive MIMO processing exploring daisy-chain architecture and recursive algorithms," *IEEE Transactions on Signal Processing*, vol. 68, pp. 687–700, Jan. 2020.

[11] A. Mueller, A. Kammoun, E. Björnson, and M. Debbah, "Linear precoding based on polynomial expansion: reducing complexity in massive MIMO," *EURASIP Journal on Wireless Communications and Networking*, no. 63, pp. 1687–1499, Feb. 2016.

[12] R. W. Heath, N. González-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 3, pp. 436–453, Apr. 2016.

[13] A. Dua, K. Medepalli, and A. J. Paulraj, "Receive antenna selection in MIMO systems using convex optimization," *IEEE Transactions on Wireless Communications*, vol. 5, no. 9, pp. 2353–2357, Sep. 2006.

[14] X. Gao, O. Edfors, F. Tufvesson, and E. G. Larsson, "Massive MIMO in real propagation environments: Do all antennas contribute equally?" *IEEE Transactions on Communications*, vol. 63, no. 11, pp. 3917–3928, Nov. 2015.

[15] P. Lin and S. Tsai, "Performance analysis and algorithm designs for transmit antenna selection in linearly precoded multiuser MIMO systems," *IEEE Transactions on Vehicular Technology*, vol. 61, no. 4, pp. 1698–1708, May 2012.

[16] A. Amiri, C. N. Manchon, and E. de Carvalho, "Deep learning based spatial user mapping on extra large MIMO arrays," *arXiv. 2002.00474*, Feb. 2020.

[17] J. C. Marinello, T. Abrão, A. Amiri, E. de Carvalho, and P. Popovski, "Antenna selection for improving energy efficiency in XL-MIMO systems," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 11, pp. 13 305–13 318, Nov. 2020.

[18] H. Lu and W. Fang, "Joint transmit/receive antenna selection in MIMO systems based on the priority-based genetic algorithm," *IEEE Antennas and Wireless Propagation Letters*, vol. 6, pp. 588–591, Dec. 2007.

[19] J. Lain, "Joint transmit/receive antenna selection for MIMO systems: A real-valued genetic approach," *IEEE Communications Letters*, vol. 15, no. 1, pp. 58–60, Jan. 2011.

[20] B. Makki, A. Ide, T. Svensson, T. Eriksson, and M. Alouini, "A genetic algorithm-based antenna selection approach for large-but-finite MIMO networks," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 7, pp. 6591–6595, Jul. 2017.

[21] J. R. Sánchez, J. Vidal Alegría, and F. Rusek, "Decentralized massive MIMO systems: Is there anything to be discussed?" in *2019 IEEE International Symposium on Information Theory*, Jul. 7–12 2019, pp. 787–791.

[22] A. Amiri, M. Angjelichinoski, E. de Carvalho, and R. W. Heath, "Extremely large aperture massive MIMO: Low complexity receiver architectures," in *2018 IEEE Globecom Workshops*, Dec. 9–13 2018, pp. 1–6.

[23] A. Amiri, C. N. Manchón, and E. de Carvalho, "A message passing based receiver for extra-large scale mimo," in *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, 2019, pp. 564–568.

[24] A. Amiri, S. Rezaie, C. N. Manchon, and E. de Carvalho, "Distributed receivers for extra-large scale MIMO arrays: A message passing approach," *arXiv. 2007.06930*, Jul. 2020.

[25] X. Yang, F. Cao, M. Matthaiou, and S. Jin, "On the uplink transmission of multiuser extra-large scale massive MIMO systems," *arXiv. 1909.06760*, Nov. 2019.

[26] D. A. Gore, R. U. Nabar, and A. Paulraj, "Selecting an optimal set of transmit antennas for a low rank matrix channel," in *2000 International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, Jun. 5–9 2000, pp. 2785–2788.

[27] A. F. Molisch, M. Z. Win, Yang-Seok Choi, and J. H. Winters, "Capacity of MIMO systems with antenna selection," *IEEE Transactions on Wireless Communications*, vol. 4, no. 4, pp. 1759–1772, Jul. 2005.

[28] S. Asaad, A. M. Rabiei, and R. R. Müller, "Massive MIMO with antenna selection: Fundamental limits and applications," *IEEE Transactions on Wireless Communications*, vol. 17, no. 12, pp. 8502–8516, Dec. 2018.

[29] C. Ouyang, Z. Ou, L. Zhang, P. Yang, and H. Yang, “Asymptotic upper capacity bound for receive antenna selection in massive MIMO systems,” in *2019 IEEE International Conference on Communications*, May. 20–24 2019, pp. 1–6.

[30] Z. Abdullah, C. C. Tsimenidis, G. Chen, M. Johnston, and J. A. Chambers, “Efficient low-complexity antenna selection algorithms in multi-user massive MIMO systems with matched filter precoding,” *IEEE Transactions on Vehicular Technology*, vol. 69, no. 3, pp. 2993–3007, Mar. 2020.

[31] H. Siljak, I. Macaluso, and N. Marchetti, “Distributing complexity: A new approach to antenna selection for distributed massive MIMO,” *IEEE Wireless Communications Letters*, vol. 7, no. 6, pp. 902–905, Dec. 2018.

[32] P. He, L. Zhao, S. Zhou, and Z. Niu, “Water-filling: A geometric approach and its application to solve generalized radio resource allocation problems,” *IEEE Transactions on Wireless Communications*, vol. 12, no. 7, pp. 3637–3647, Jun. 2013.

[33] G. H. Golub and C. F. V. Loan, *Matrix Computations*. Baltimore, MD, USA: Johns Hopkins University Press, 2013.

[34] E. Montero, M.-C. Riff, and B. Neveu, “A beginner’s guide to tuning methods,” *Applied Soft Computing*, vol. 17, pp. 39–51, Apr. 2014.

[35] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, *Fundamentals of Massive MIMO*. Cambridge, Cambridgeshire, UK: Cambridge University Press, 2016.

Paper I

Non-Stationarities in Extra-Large Scale Massive MIMO

Elisabeth De Carvalho, Anum Ali, Abolfazl Amiri, Marko Angjelichinoski, Robert W. Heath Jr.

The paper has been published in the
IEEE Wireless Communications Magazine, vol. 27, no. 4, pp.74–80, 2020.

Abstract

Massive MIMO, a key technology for increasing area spectral efficiency in cellular systems, was developed assuming moderately sized apertures. In this paper, we argue that massive MIMO systems behave differently in large-scale regimes due to spatial non-stationarity. In the large-scale regime, with arrays of around fifty wavelengths, the terminals see the whole array but non-stationarities occur because different regions of the array see different propagation paths. At even larger dimensions, which we call the extra-large scale regime, terminals see a portion of the array and inside the first type of non-stationarities might occur. We show that the non-stationarity properties of the massive MIMO channel change several important MIMO design aspects. In simulations, we demonstrate how non-stationarity is a curse when neglected but a blessing when embraced in terms of computational load and multi-user transceiver design.

1 Introduction

Massive multiple-input-multiple-output (MIMO) is a key technology in 5G wireless communication systems in sub-6 GHz bands. It is characterized by the use of many antennas at the base station serving many terminals simultaneously. In current cellular deployments, massive MIMO will likely be implementing compact planar arrays. The small footprint leads to reduced infrastructure costs. Even with a large number of antennas, though, compact design does not expose enough spatial dimensions.

Spatial dimensions are essential in uncovering the fundamental properties of massive MIMO: channel hardening, asymptotic inter-terminal channel orthogonality, and large array gains. Increasing the array dimension contributes to achieving the performance gains originally promised by massive MIMO and providing high data rates when the number of terminals is much smaller than the number of antennas. Increasing the array dimension further allows the support of high data rates to a much larger number of terminals. Distributing the arrays across a building, for example, allows for cost-efficient implementation of an extremely large array while bringing other benefits such as better coverage.

The impact of the array dimension has motivated new types of deployment where the dimension of the arrays is pushed to the extreme. Such arrays would be integrated into large structures, for example along the walls of buildings in a mega-city, in airports, large shopping malls or along the structure of a stadium [1, 2] (see Fig. I.1) and serve a large number of devices. This type of deployment is considered an extension of massive MIMO with an implementation based on discrete antenna elements. We refer to this extreme case as extra-large scale massive MIMO (XL-MIMO). We argue in this paper that XL-MIMO should be considered a distinct operating regime of massive MIMO with its unique challenges and opportunities.

When the antenna arrays reach such a large dimension, spatial non-wide sense stationary properties appear along the array. Different parts of the array may have different views of the propagation environment, observing the same channel paths with different power, or different channel paths [3]. When the dimension of the array becomes extremely large, different parts of the array may also view different terminals

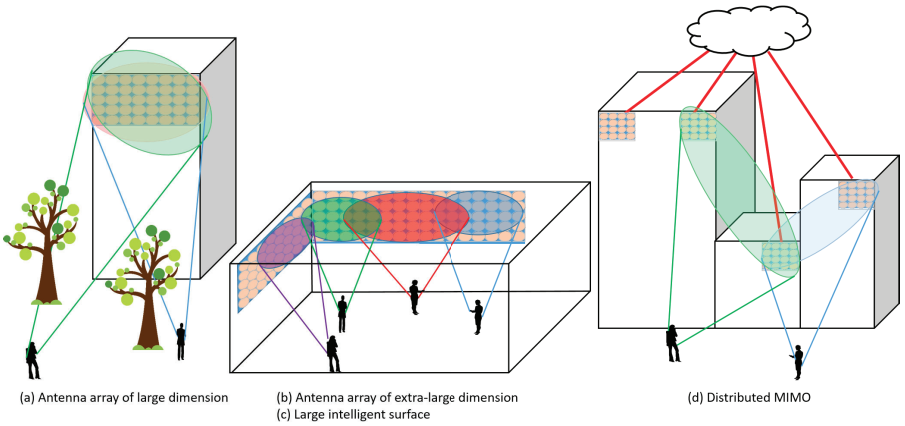

(a) Antenna array of large dimension
(b) Antenna array of extra-large dimension
(c) Large intelligent surface
(d) Distributed MIMO

Fig. I.1: The ways to create larger apertures in a massive MIMO. (a) Antenna array with large dimension. (b) Antenna array with extra-large dimension. (c) Large intelligent surface. (d) Distributed antenna system.

as the energy of each terminal is focused on a portion of the array, called visibility region (VR). As the array dimension increases, the performance for each terminal is limited by its VR, i.e., the effective array dimension viewed from the array. However, the ability to serve multiple terminals with high data rates is highly enhanced, hence bringing benefits in crowded scenarios.

Wireless communications involving large electromagnetic elements is an emerging concept. The term Large Intelligent Surface (LIS) has appeared recently and denotes generically a large electromagnetic surface [4] that is active and hence possesses communication capabilities. Another possible implementation of very large arrays is through radio stripes as described in [5] that can be easily attached to existing construction structures and are connected to a central unit to form a distributed cell-free system. Interestingly, research is also focusing on passive large electromagnetic surfaces [6]. A passive LIS acts as a reflecting surface that changes the properties of the incoming electromagnetic waves. It acts as a relay to enhance the propagation features of the reflected waves.

In this article, we focus on discrete arrays of antennas, not continuous surfaces, and the effect of non-stationary properties along the array. Our emphasis is on VRs and their impact on performance and transceiver design. The primary differentiating feature from stationary massive MIMO is that the terminals have overlapping VRs with an inter-terminal interference pattern that changes along the array. Non-stationarity is accounted for in the performance assessment of linear multi-terminal transceivers and design of hybrid analog-digital beamforming and serves as the main tool to alleviate the transceiver computational complexity.

2 Types of spatial non-stationary regimes

Fig. I.1 gives an overview of the types of deployments considered in this paper and the ways to create larger apertures for XL-MIMO.

1. An antenna array of large or extra-large dimension: typically embedded in a

building of large dimension [1].

2. Large intelligent surface: a generic term for a large electromagnetic surface [1]. A possible implementation is with a discrete array of antennas (as in case (a)) but possibly other material.
3. Distributed antenna system: cooperating antennas or arrays of antenna units placed at distant geographical locations [7].

To illustrate the spatial non-stationarity properties in massive MIMO, we rely on a cluster-based channel model. Fig. I.2 depicts a conventional massive MIMO channel model that is spatially stationary, along with two types of spatial non-stationarities defined according to the concept of VR along an antenna array. The concept of VR was introduced in the COST 2100 channel model [8]. In its original definition, a VR is a terminal geographical area. When the terminal is located in this area, it sees a given set of clusters. This is the set of clusters associated with the VR. When it moves out of the VR, the terminal sees a different set of clusters. We extend the concept of VR to denote a portion of the array from which a given set of clusters is visible. We distinguish between VRs in the terminal domain VR-T and in the array domain VR-A.

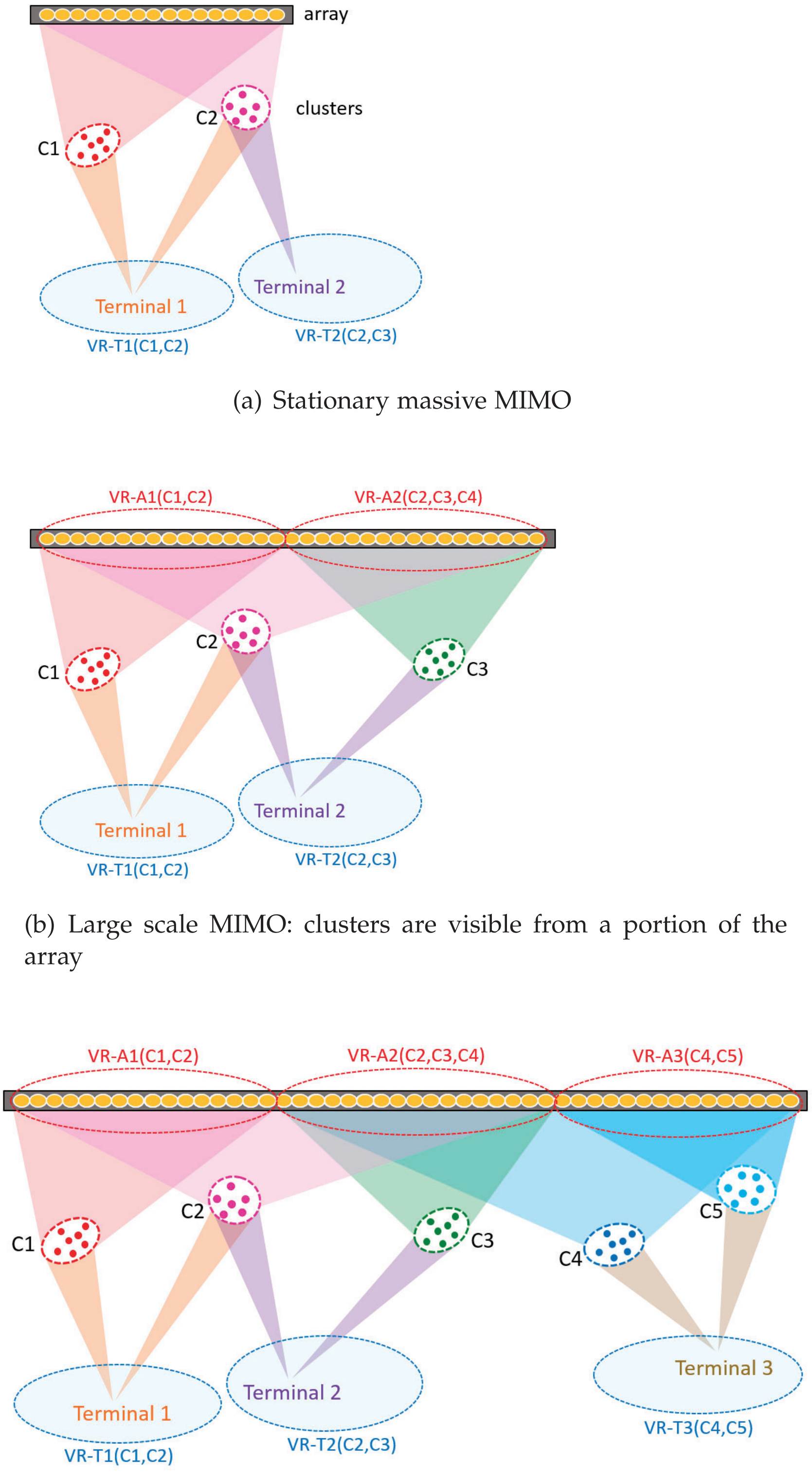
2.1 Large-scale massive MIMO

The L-MIMO regime applies when different sets of clusters are visible from different portions of the array and the whole array is visible by all terminals. In general, this implies that the terminals are at a significant distance from the array. Fig. I.2(b) illustrates a simple case where the array is divided into two disjoint VR-As. This regime was highlighted in an early measurement [3] involving a long array of 7.4 meters in a courtyard where, at a different portion of the array, different propagation paths were measured.

2.2 Extra-large scale massive MIMO

The XL-MIMO regime applies when different sets of clusters as well as different sets of terminals are visible from different portions of the array. The main difference with the L-MIMO regime is that the terminals are much closer to the array (or the array is much larger). As seen in Fig. I.2 (c), one can define another type of VR: the portion of the array that is visible from a given terminal. For example, the VRs of terminal 1 along the array includes VR-A1 and VR-A2.

Aalborg University initiated a measurement campaign specifically dedicated to XL-MIMO [2] in a large indoor venue. Fig. I.3(a) shows a striking result from the campaign that illustrates the complexity of the propagation environment. The massive array is six-meter long and comprises 64 antennas. It is placed along a wall on a line parallel to the floor and made of units of eight antennas. Eight terminals, around three meters apart, holding a two-antenna device are located at 2 and 6 meters in front of the array and send uplink signals as shown in Fig. I.3(b). Fig. I.3(a) displays the average receive power of the channel when the terminals move locally. First, we



(a) Stationary massive MIMO

(b) Large scale MIMO: clusters are visible from a portion of the array

(c) Extra-large scale MIMO: terminals are visible from a portion of the array

Fig. I.2: Three MIMO scales

observe very large variations of the power across the array, more than 10dB and different patterns for the two signals coming from the same device. Terminals 5 to 8 are located behind a stair case, which brings an attenuation of the signal visible from a portion of the array.

2.3 Distributed massive MIMO

XL-MIMO can be seen as a special case of distributed massive MIMO where the whole set of arrays is collocated. Especially in a dense distribution, the same kind of model holds where clusters, as well as terminals, are visible from a subset of the arrays.

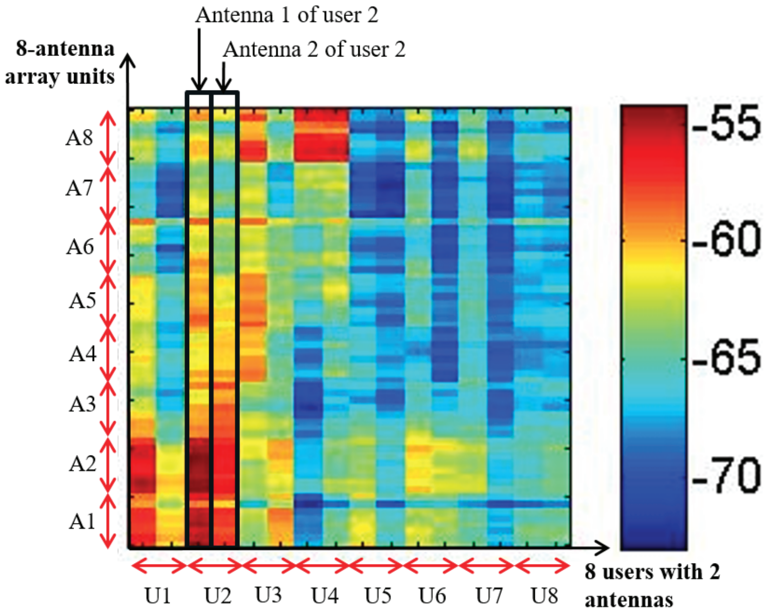
2.4 Impact on key channel assumptions

The non-stationary properties of arrays of very large dimension impose a departure from the conventional channel models, especially the widely used correlated channel model. This model assumes that the channel has a centered Gaussian distribution with a covariance matrix that reflects stationary properties in the correlation among antennas as well as the propagation. While the Gaussian assumption might still hold, the most basic modification on the channel assumptions is that the average channel gain varies along the array. A cluster-based geometric channel model reflects more appropriately the source of non-stationarity, i.e. cluster VRs. The major change compared to traditional models is in the expression of the steering vectors. First, near the array, the phase of each element should account for a spherical wave modeling as the planar wave approximation is not valid anymore. Second, the amplitude of each element varies. This is due to the path loss along the array as well as the interplay between clusters and obstacles in the environment as different portions of the spherical wavefront might experience different propagation characteristics. The main drawback of this modeling is that it depends on the position of the clusters and terminals relative to the array, which makes it scenario-dependent and increases its complexity.

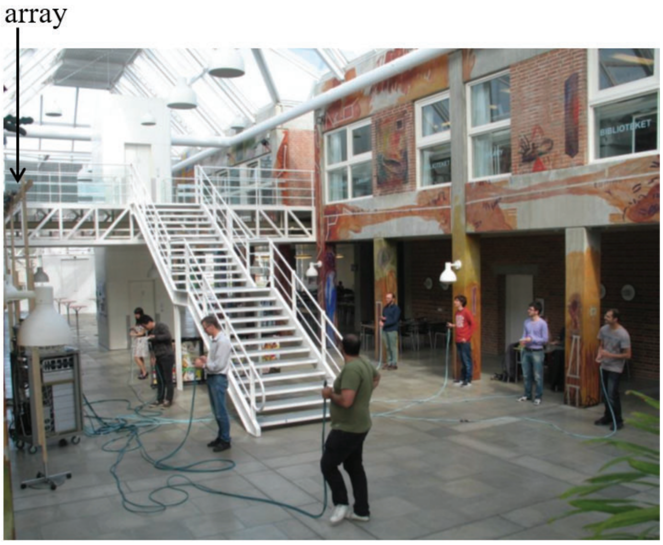
A simplification consists in decomposing the array in sub-arrays in which the channel is approximated as stationary. This model can be enhanced by adding a transition zone between the sub-arrays [9]. This type of assumption can facilitate performance analysis of XL-MIMO systems [10]. It motivates multi-antenna processing based on sub-arrays where the sub-array processing is adapted to the non-stationarity patterns.

3 Exploiting spatial non-stationarity

This section advocates that non-stationary properties, with a focus on VRs, should be accounted in performance assessment as well as transceiver design. We provide performance bounds of the zero-forcing (ZF) precoder using a simple VR model. Further, we demonstrate that VRs should be taken into account when designing hybrid analog-digital precoders/combiners. Finally, we exploit VRs, i.e., array regions where signals have low power, to design low complexity receivers.



(a) Average received power (dBm) in a 64 antenna array made of 8-antenna units (y-axis). The received power is averaged over the small movements of the terminals.



(b) The array is 6 meters long. Eight terminals (y-axis) holding a 2-antenna device are around 2 and 6 meters from the array and move in a square of 1 square meter.

Fig. I.3: The measurement set-up and results for XL-MIMO [2].

3.1 Performance bounds

Only a handful of studies have been conducted to study the impact of non-stationarity on the performance of massive MIMO systems. Channel capacity is studied in [11] for a spherical wave-front based LOS channel model, while cluster non-stationarity visibility along the array is treated in [12]. The focus of this section is on the impact of array VRs associated with the terminals and how it compares to the conventional stationary case. In [10], a simple non-stationary massive MIMO channel model was proposed so that it is conducive to the analysis of the effect of VRs. This model is employed to assess the performance of simple linear multi-terminal precoders (conjugate-beamforming (CB) and ZF precoders).

The channel of a given terminal is modeled as stationary within its VRs and is set to zero outside of the VR. Though the model was developed for correlated channels (see [10]), here, we limit our discussions to independent and identically distributed channels, for simplicity, and ZF precoding. Consider a MIMO broadcast channel with K single-antenna terminals served by a BS with M antennas. The SINR of terminal k for ZF precoding averaged over stationary channels has a well-known expression. However, VR-based channels are not easily amenable to analysis. It is possible however to find an approximation of the SINR, valid in asymptotic conditions, as a function of the VR size of each terminal and the size of the overlap regions. For simplicity, we assume that the terminals have the same VR size equal to D antennas and total transmit energy per VR is equal to M. We examine the worst and best case terminal configuration. The SINR can be written as $\frac{\rho}{K}(M - L(K, M, D))$ where the loss term $L(K, M, D)$ differs in each case. The term ρ is the transmit signal-to-noise ratio.

In the worst case, all the terminals have completely overlapping VRs - i.e., they receive the signal from the same D antennas - the inter-terminal interference is high. In the best-case, inter-terminal interference is minimized asymptotically for all K terminals. The terminals are grouped in M/D groups where each group contains $\frac{KD}{M}$ terminals. Hence, there are $\frac{KD}{M} - 1$ interfering terminals for any terminal k with an overlapping zone of D antennas.

The SINRs all scale as M/K and differ in lower order quantities. The best-case non-stationary scenario results in better performance than the stationary case. It reaches its largest value when $\frac{M}{D}$ is large, i.e., for small VRs or non-overlapping VRs. The worst-case non-stationary scenario results in worst performance than the stationary case. The smaller the VR of the terminal, the more SINR loss compared to the stationary case.

In Fig. I.4, we provide an example result to demonstrate the impact of non-stationarity on the performance of ZF precoding. We plot the SINR results against the active number of antennas per terminal, i.e., D. We can see that depending on the configuration (i.e., best-case or worst-case) the SINR can be significantly higher/lower than the SINR of the stationary channels. As expected, the differences are greater for smaller values of D.

The non-stationarity captured using VRs and subsequent analysis shows that non-stationarity has a significant impact on the performance of a massive MIMO system. As such, it is imperative to understand this impact and to exploit it in designing massive MIMO systems.

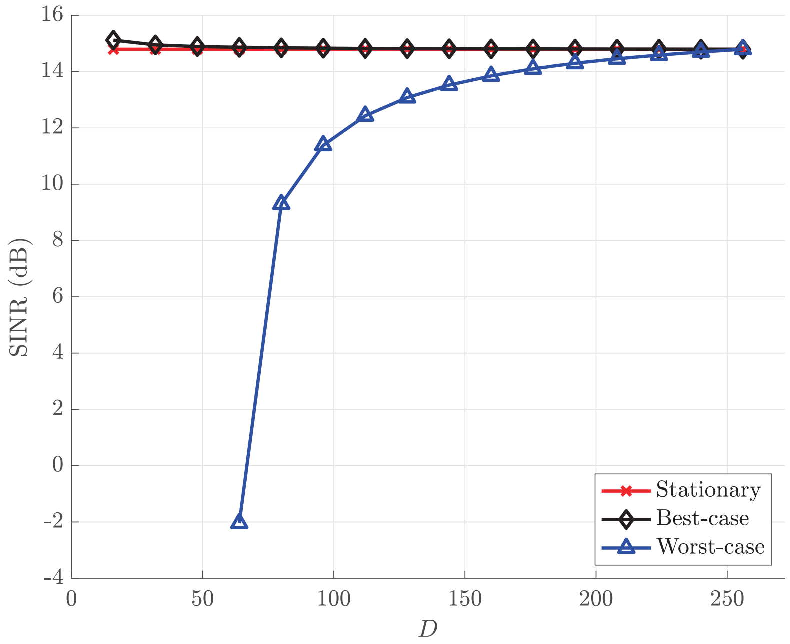


Fig. I.4: The SINR of kth user vs the active number of antennas D ($M = 256$, $K = 64$, and $\rho = 10$dB).

3.2 Hybrid beamforming

Hardware and cost constraints make it challenging to connect all the antennas in a massive MIMO system with dedicated RF-chains and high-resolution ADCs. Therefore hybrid analog-digital architectures, where a few RF-chains are connected to a large number of antennas are suitable for massive MIMO systems. The hybrid analog-digital architectures keep the cost and complexity under control by using fewer RF-chains compared to the number of antennas but allow multi-terminal multi-stream precoding that is not possible using analog-only architectures.

There are several possibilities for implementing hybrid analog-digital architectures. The more flexible (but complex to implement) architecture is fully-connected architecture, where all the RF-chains are connected to all the antennas. A simpler (but less flexible) architecture is partially connected architecture in which every antenna is connected to a subset of RF-chains. Recently, dynamic hybrid architectures are also considered that adapt to the channel, hence providing flexible yet simpler implementation [13].

The dynamic hybrid analog-digital architectures can be particularly beneficial in non-stationary channels. Motivated by the VR-based channel model discussed in the last section, it can be argued that a simple dynamic hybrid analog-digital architecture is one in which only the antennas corresponding to the VR are connected to the RF-chains. This is feasible as the antennas outside the VR do not have significant channel power. Thus a low complexity dynamic architecture can potentially provide performance close to the fully digital system but at low hardware cost.

To show the benefit of non-stationarity aware system design, we provide simulation results. Assuming $D = M/2$ size VR for each terminal (where visible antennas are chosen uniformly at random), we provide the average SINR of ZF precoder with

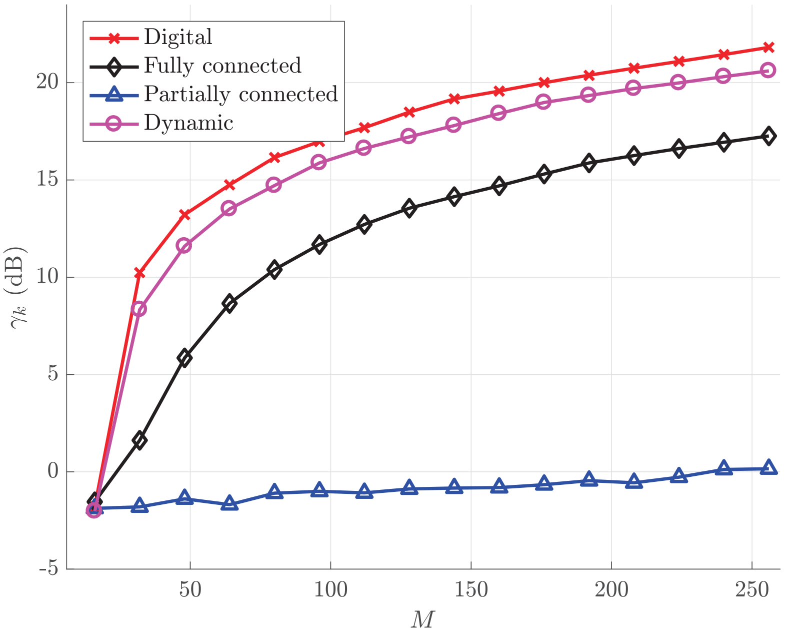


Fig. I.5: The SINR vs the number of antennas M ($D = M/2$, $K = 16$, and 16 RF-chains).

different hardware architectures. There are $K = 8$ terminals in the system. The fully digital architecture has M RF-chains, and we use the algorithm proposed in [14] to obtain the hybrid precoders. The hybrid analog-digital architectures have RF-chains equal to the number of terminals K. The partially connected hybrid architecture has an RF-chain connected to M/K successive antennas. The dynamic architecture has kth RF-chain connected only to the antennas visible to the terminal K. From the results in Fig. I.5 we can see that the dynamic architecture can provide performance better than fully-connected architecture (in non-stationary channels) and close to fully digital system.

Dynamic hybrid architectures are interesting for non-stationary massive MIMO. These architectures are an example of a system design that exploits the non-stationary nature of the massive MIMO channel. The results presented herein, however, are preliminary and a lot of research is required for practical designs. One major challenge in dynamic hybrid architectures is the efficient acquisition of channel state information (the presented results are based on genie aided CSI).

3.3 Low complexity transceivers

One obvious consequence of having an extremely large number of antennas at the base station is its high complexity architecture. Even with simple linear transceivers, the base station should perform a large number of complex operations. This problem gets even worse when it comes to crowded scenarios with many terminals in the system. Therefore, implementing low complexity techniques is one major challenge. One possible way is to adapt the transceiver design to the non-stationary energy patterns of the terminals, complemented by distributed processing methods such as sub-array based architectures. To determine low complexity transceivers, acquiring information about the VRs is critical.

The existence of VRs is the basis to implement low complexity linear transceivers such as the ZF. Indeed, the computational cost of implementing a ZF operation is dominated by the inversion of a matrix that has a band structure due to the VRs and might even be sparse. However, implementing distributed techniques is more favorable due to lower complexity and more flexibility. Distributed processing is motivated not only by the computational cost but also by the ease of installation of very large arrays that are made out of smaller sub-arrays. Each sub-array carries out local processing of the signals while a central unit is responsible for the final data fusion step.

As the terminals are connected to a subset of sub-arrays, a graph can be used to describe the connections between terminals and sub-arrays. When the terminals are connected to a small number of sub-arrays, the graph is sparse and becomes a convenient tool to facilitate low complexity transceiver designs. Compared to a fixed sub-array division, a dynamic division leads to a better performance outcome where the division fits ideally the multi-terminal VR patterns and should be updated for the changes in the VR patterns. Simple learning algorithms can help in tracking the power pattern of the terminals over the array.

Considering the uplink, a linear fusion of the sub-array output signals is carried out at the central unit. When arrays are deployed in a very large structure, such as around the roof of a stadium, the processing can be structured hierarchically with a multi-stage fusion involving a hierarchical subset of sub-arrays at each step.

Non-linear processing can be beneficial in some situations to improve performance compared to linear fusion. Due to spatial non-stationarities, multi-terminal interference patterns vary over the array so that one terminal experiences different interference conditions at each of the sub-array. Therefore, it becomes beneficial to detect a terminal from the sub-array with favorable interference conditions and then remove its contributions from the other sub-arrays, enhancing the signal to interference ratio of all the other terminals. This nonlinear method follows the principle of successive interference cancellation technique and was tested in [15]. More advanced receiver based on message passing among sub-arrays can be employed to reduce the performance gap with the optimal methods such as maximum likelihood.

In Fig. I.6, we test the notion of VR in a multi-terminal processing. The terminals are uniformly distributed in front of a linear array comprising 1024 antennas. The channel is assumed to have a Gaussian distribution while the energy variations along the array come from the path loss. The figure displays the spectral efficiency per terminal as a decreased number of antennas is considered in the VR of each terminal. For this channel model, we observe a saturation in the performance of the centralized processing. With some performance degradation, the processing can be reduced to a relatively small number of antennas per terminals (two sub-arrays of 128 antennas). In distributed processing, we have 128 antennas per sub-array. We observe a degradation when the number of antennas in the VR increases for a loaded system (128 terminals). As the VR becomes larger, the number of terminals to be processed per sub-array increases until reaching a regime where the sub-optimality of distributed processing becomes apparent.

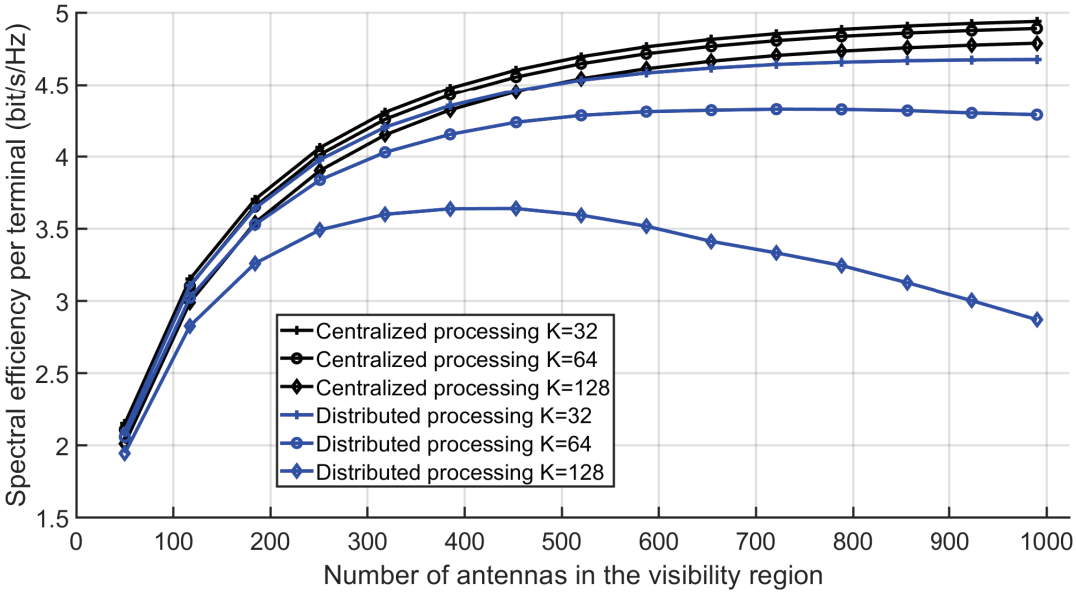


Fig. I.6: Rate per user comparison between centralized and distributed ZF processing vs number of contributing antennas in the VR of each terminal. ($M = 1024$ and SNR= 15dB).

4 Next steps

4.1 Characterizing the channel

Near-field channel measurements involving extremely large arrays where non - stationary patterns are visible are scarce (see section I.2). Yet, they are necessary, as little is known about their non-stationary features outside of a theoretical framework. Channel measurements are needed to understand in more depth the propagation behavior in real-life set-ups. For example, it appears important to uncover how the wireless channel behaves in a large indoor venue with very large antenna panels deployed along the walls: is the channel sparse or does it demonstrate rich scattering? The issue is not only about the phase front that is spherical and not planar anymore. It is also about the channel energy variations along the array as was the focus in this article. Those variations are the results of the path loss in line-of-sight but also the geometry of the building and reflecting structures (ceiling, floor, stairs, various objects) that are near the communicating panels and the end-terminals.

To characterize the channel, a multitude of measurements are needed in different deployment scenarios (e.g., different room types, outdoor scenarios) to guarantee statistical significance. Many more measurements are necessary to extract the non-stationary channel attributes but also other important features impacting channel modeling. An extremely large array views the propagation environment with super-resolution. The objects along the propagation do not look the same when illuminated by a large array. For example, large arrays can differentiate a set of reflecting entities that would be part of a cluster otherwise. Hence, the definition of clusters can be questioned, as well as the distribution of the small scale fading along the propagation path. Another example is about the modeling of large scale fading. As a terminal moves locally, large scale fading remains identical for a compact array size. With very large arrays, even very small movements might impact large scale quantities.

4.2 Embracing electromagnetics

Communication in the near-field and hence spherical wave modeling implies that the propagation features are dependent on the relative position of the end-terminals to the electromagnetic panels, the size of the panels, as well as their magnetic properties. This paper has considered very simplified models to highlight the impact of energy variations along the array panels. There is a need, though, to revisit communication theory to incorporate those electromagnetic attributes more faithfully.

The incorporation of advanced electromagnetic features impacts the development of algorithms and likely adds to their complexity. As an example, compressed sensing methods are employed widely for sparse channel estimation. Those methods rely usually on a dictionary, i.e., typically an over-complete set of vectors that span the propagation space. Spherical waves imply that more parameters are necessary to describe the dictionary.

Compared to stationary massive MIMO, the array aperture offers an additional degree of freedom: the assignment of a subset of antennas to each terminal. In section 3.3, we have restricted the processing area per terminal to the visibility region. Computational cost motivates the shrinkage of the processing area. Inside the processing area, the signal of a terminal is a signal of interest while it is treated as interference outside. Satisfying specific metrics provides another motivation: a terminal requiring more data is assigned a larger area while a fairness criterion might lead to a balanced assignment. This processing leads to a system model that lies between full network MIMO (the processing areas correspond to the visibility regions) and MIMO interference channel (the processing areas to all terminals are disjoint). This is reminiscent of the access point association in distributed settings. It is different though due to the high resolution in the assignment problem and an assignment that could be highly dynamic and follows the movements of the terminals.

5 Conclusions

XL-MIMO is an extreme but practical case of massive MIMO with larger apertures. This paper has focused on discrete antenna arrays of extremely large dimension that are deployed as part of a new large building structure. Along with active and passive large electromagnetic surfaces, they participate in a vision of ubiquitous connectivity where a connection is not achieved through access points anymore but rather through diffuse access that is located much closer to the end terminals. Such a vision is not realized yet and comprises many practical challenges. We have discussed how non-stationarities along the array found in XL-MIMO change the performance of MIMO systems and how visibility regions can be accounted for to decrease the computational load associated to MIMO transceivers in centralized or distributed implementations. When communication happens in the near-field, many other communication aspects are impacted. For example, propagation attributes become different from the conventional far-field so that channel models need to be properly adjusted calling for new measurements. Directional beamforming is more complex because the beam does not depend on the directions only but also on the position of the terminals relative to the array. Those might be well-known properties of near-field communications. However, the array dimension brings specific challenges in terms of computational load that need to be addressed.

References

[1] J. Medbo, K. Börner, K. Haneda, V. Hovinen, T. Imai, J. Järvelainen, T. Jämsä, A. Karttunen, K. Kusume, J. Kyröläinen, P. Kyösti, J. Meinilä, V. Nurmela, L. Raschkowski, A. Roivainen, and J. Ylitalo, “Channel modelling for the fifth generation mobile communications,” in *The 8th European Conference on Antennas and Propagation (EuCAP 2014)*, April 2014, pp. 219–223.

[2] Á. O. Martínez, E. De Carvalho, and J. Ø. Nielsen, “Towards very large aperture massive mimo: A measurement based study,” in *2014 IEEE Globecom Workshops (GC Wkshps)*, Dec 2014, pp. 281–286.

[3] X. Gao, F. Tufvesson, and O. Edfors, “Massive mimo channels - measurements and models,” in *2013 Asilomar Conference on Signals, Systems and Computers*, Nov 2013, pp. 280–284.

[4] S. Hu, F. Rusek, and O. Edfors, “Beyond massive mimo: The potential of data transmission with large intelligent surfaces,” *IEEE Transactions on Signal Processing*, vol. 66, no. 10, pp. 2746–2758, May 2018.

[5] G. Interdonato, E. Björnson, H. Q. Ngo, P. Frenger, and E. G. Larsson, “Ubiquitous cell-free massive mimo communications,” *https://arxiv.org/abs/1804.03421*, October 2018.

[6] C. Liaskos, S. Nie, A. Tsioliaridou, A. Pitsillides, S. Ioannidis, and I. Akyildiz, “A new wireless communication paradigm through software-controlled metasurfaces,” *IEEE Communications Magazine*, June 2018.

[7] K. T. Truong and R. W. Heath, “The viability of distributed antennas for massive mimo systems,” in *2013 Asilomar Conference on Signals, Systems and Computers*, Nov 2013, pp. 1318–1323.

[8] L. Liu, C. Oestges, J. Poutanen, K. Haneda, P. Vainikainen, F. Quitin, F. Tufvesson, and P. D. Doncker, “The cost 2100 mimo channel model,” *IEEE Wireless Communications*, vol. 19, no. 6, pp. 92–99, December 2012.

[9] A. O. Martinez, P. Eggers, and E. De Carvalho, “Geometry-based stochastic channel models for 5G: Extending key features for massive mimo,” in *Personal, Indoor, and Mobile Radio Communications (PIMRC), 2016 IEEE 27th Annual International Symposium on*. IEEE, 2016, pp. 1–6.

[10] A. Ali, E. de Carvalho, and R. W. Heath Jr, “Linear Receivers in Non-stationary Massive MIMO Channels with Visibility Regions,” *IEEE Wireless Commun. Lett.*, 2019, (Early Access).

[11] Z. Zhou, X. Gao, J. Fang, and Z. Chen, "Spherical wave channel and analysis for large linear array in los conditions," in *Proceedings of IEEE Global Telecommunications Conference*, 2015, pp. 1–6.

[12] X. Li, S. Zhou, E. Björnson, and J. Wang, "Capacity analysis for spatially non-wide sense stationary uplink Massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 14, no. 12, pp. 7044–7056, 2015.

[13] S. Park, A. Alkhateeb, and R. W. Heath, "Dynamic subarrays for hybrid precoding in wideband mmWave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 2907–2920, 2017.

[14] L. Liang, W. Xu, and X. Dong, "Low-complexity hybrid precoding in massive multiuser MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 3, no. 6, pp. 653–656, 2014.

[15] A. Amiri, M. Angjelichinoski, E. De Carvalho, and R. W. Heath Jr, "Extremely large aperture massive mimo: Low complexity receiver architectures," *arXiv preprint arXiv:1810.02092*, 2018.

SUMMARY

Massive MIMO (multiple-input multiple-output) systems are key candidates for the fifth generation (5G) of cellular networks. Having a lot of antenna elements at the base station (BS) is an important enabler to provide a very high spatial resolution. Therefore, systems beyond 5G rely on increasing the number of elements at the BS to support future applications. At very large dimensions, e.g. aperture sizes bigger than 100 wavelengths, a new type of array called extra-large scale MIMO (XL-MIMO) emerges that offers enhanced spectral and energy efficiency.However, practical implementation of such arrays requires overcoming several challenges such as computational complexity, hardware limitations and non-stationary propagation patterns.

This thesis presents several techniques to handle major existing concerns in the XL-MIMO arrays, namely: computational complexity of receiver algorithms, scalability and interconnection overheads. In order to address the complexity issue, different low complexity methods are proposed. One of the main differences between these methods and conventional linear receivers in massive MIMO systems is, that they exploit the information about user energy patterns over the array to operate more effectively. Another approach is to distribute the receiver processing tasks between several nodes and create a hierarchy between processing nodes. The thesis studies different architectures and mostly focuses on a distributed way that uses sub-arrays to obtain local estimates at local nodes. Then, a central node collects all the local data to perform a global decision. Furthermore, the thesis suggests several antenna selection methods to limit the area of the array being processed and control the amount of computations. These methods directly use the received energy patterns at the BS to find the best active antenna sets and turn off the rest of the array to save energy. Moreover, to address the hardware considerations such as scalability and inter-connection overheads, a fully decentralized method is proposed that works without a central node.

In summary, the main outcome of the thesis is the proposal of signal processing enablers for the XL-MIMO systems. The proposed methods address the aforementioned challenges while providing acceptable performance.

ISSN (online): 2446-1628
ISBN (online): 978-87-7573-986-8

AALBORG UNIVERSITY PRESS