Bayesian statistical methods for joint user activity detection, channel estimation, and data decoding in dynamic wireless networks

Fakher Sagheer

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2024IPPAS024

Thèse de doctorat

JOINT USER ACTIVITY DETECTION, CHANNEL ESTIMATION AND DATA DECODING IN DYNAMIC WIRELESS NETWORKS

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à Télécom SudParis

École doctorale n°626 Dénomination (EDIPP)
Spécialité de doctorat: Réseaux, informations et communications

Thèse présentée et soutenue à Palaiseau, le 12/12/2024, par

FAKHER SAGHEER

Composition du Jury :

Pr. Didier Le Ruyet Professeur, CNAM Paris (CEDRIC)	Président/Examinateur
Pr. Jean-Marie Gorce Professeur, INSA Lyon (INRIA)	Rapporteur
Pr. Jean-Pierre Cancès Professeur, Université de Limoges (XLIM)	Rapporteur
Pr. Lina Mroueh Professeure, ISEP (LISITE)	Examinateur
Pr. Frederic Lehmann Professeur, Telecom SudParis (SAMOVAR)	Directeur de thèse
Pr. Antoine Berthet Professeur, Centrale Supelec (L2S)	Co-directeur de thèse

Acknowledgements

The five years of my PhD journey have been challenging and, although longer than the typical duration, they have been filled with growth and valuable experiences. During this time, I have gained both teaching and research expertise, which has greatly enriched my academic career

I would also like to express my deepest gratitude to my supervisors, Prof. Frédéric Lehmann and Prof. Antoine Berthet, for their continuous guidance, insightful feedback, and unwavering support throughout my research. Their expertise and encouragement have been invaluable in overcoming challenges and enriching the quality of this work.

I extend my sincere thanks to the jury members, Prof. Jean-Marie Gorce, Prof. Jean-Pierre Cances, Prof. Didier Le Ruyet, and Prof. Lina Mroueh, for their valuable time in evaluating my report. Their insightful feedback during the thesis defense, along with their thoughtful remarks and genuine interest in my research, have greatly contributed to refining and enhancing the quality of my work.

I would also like to extend my heartfelt thanks to my wife for her patience, encouragement, and constant support during this journey. To my son, Yamaan Khan, your boundless love and joyful spirit,have been a source of endless motivation and happiness.

Contents

List of Figures

*

List of Tables

*

Acronyms

P_{fa} Probability of False Detection. 46

P_{md} Probability of Missed Detection. 46

3GPP 3rd Generation Partnership Project. 7

5GAA 5G Automotive Association. 5

AI Artificial Intelligence. 18

AMP Approximate Message Passing. 19

AoA Angle of Arrival. 23

AR Augmented reality. 2

BER Bit-error Rate. 51

BLER Block Error Rate. 5

BP Belief propagation. 18

BPSK Binary Phase Shift Keying. 29

BS Base Station. 22

CD-NOMA Code-domain non-orthogonal multiple access. 8

CDMA Code Division Multiple Access. 11

CE Channel Estimation. 13

CFR Channel Frequency Response. 24, 51

CIR Channel Impulse Response. 90

CMA Contention-based multiple access. 8

CP Cyclic Prefix. 29

CS Compressive Sensing. 14

CSI Channel State Information. 31

DEC Decoding. 13

DEM Demodulation. 25

DFT Discrete Fourier Transform. 24

DSS Dynamic Spectrum Sharing. 4

EM Expectation Maximization. 32

eMBB Enhanced Mobile Broadband. 2

EP Expectation propagation. 18

FDMA Frequency Division Multiple Access. 11

FEC Forward Error Correction. 22

FER Frame Error Rate. 5

FWA Fixed Wireless Access. 4

GaBP Gaussian Belief Propagation. 41

GF-NOMA Grant-free non-orthogonal multiple access. 8

GF-RA Grant-free random access. 8

GFMA Grant-free Multiple Acces. 8

IBI Inter-block Interference. 29

ICI Inter-carrier Interference. 29

IDMA Interleaved-Division Multiple Access. 12

IoT Internet of Things. 1

IUI Inter-user Interference. 90

KL Kullback-Leibler. 18

LDS Low Density Spreading. 12

LLR Log-Likelihood ratio. 37

LTE-M Long-Term Evolution Machine Type Communication. 6

MAI Multi-access Interference. 42, 49

MAP Maximum-aposteriori Estimator. 25

MF Mean Field. 32

mGF-NOMA Massive Grant-free Access NOMA. 90

MIMO Multiple Input Multiple Output. 3

ML Maximum-Likelihood. 44

MMSE Minimum-mean-squared-error. 93

mMTC Massive machine type communications. 1

MMV Multiple Measurement Vector. 90

mmWave Millimeter Wave. 3

MSBL Multiple Sparse Bayesian Learning. 92

MSE Mean-squared Error. 51

MUD Multi-User Detection. 13

MUSA Multi-User Shared Access. 12

NB-IoT Narrowband Internet of Things. 6

NCMA Non-contention-based multiple access. 8

NOMA Non-orthogonal multiple access. 1

NR New Radio. 7

OFDMA Orthogonal Frequency Division Multiple Access. 11

OMA Orthogonal multiple access. 10

PDMA Pattern Division Multiple Access. 12

PMF Probability Mass Function. 37

QAM Quadrature amplitude modulation. 29

QoS Quality of Service. 10

QPSK Quadrature Phase Shift Keying. 22

RC Reduced-complexity. 14

RE Resource Element. 21

RIP Restricted Isometry Property. 20

RIS Reconfigurable Intelligent Surfaces. 103

RTT Round-Trip Time. 5

SBL Sparse Bayesian Learning. 90

SCMA Sparse Code Multiple Access. 12

SDMA Space-division Multiple-access. 59

SG Scheduling Grant. 9

SIMO Single-Input Multiple Output. vi, 28

SMARTER Study on New Services and Markets Technology Enablers. 4

SNR Signal-to-noise Ratio. 43

SOMP Simultaneous Orthogonal Matching Pursuit. 97

SR Scheduling Request. 9

TA Timing Advance. 9

TDMA Time Division Multiple Access. 11

UAD User Activity Detection. 13

ULA Uniform Linear Array. 22

URLLC Ultra-reliable low latency communications. 1

V2X Vehicle-to-Everything. 8

VB Variational Bayesian. 32

VR Virtual reality. 2, 4

Résumé

L'accès multiple non-orthogonal grant-free (GF-NOMA) s'impose progressivement comme une partie intégrante de la couche physique des systèmes d'accès radio du futur. En permettant d'accéder à une station de base sans allocation explicite de ressources temps/fréquence/code, GF-NOMA permet non seulement d'améliorer l'efficacité spectrale, mais également de rendre possible des communications ultra fiables à faible latence (URLLC). De telles exigences permettront de répondre aux enjeux spécifiques d'applications sans fil telles que l'internet des objets, la réalité virtuelle, les jeux vidéo en ligne, les communications entre machines, véhicules, etc.

Cependant, GF-NOMA introduit un nouveau défi inexistant dans les systèmes de communication classiques, à savoir la détection d'activité des utilisateurs : en plus de l'estimation du canal, de la détection et du décodage des utilisateurs interférant, la station de base réceptrice doit être en mesure de procéder à leur classification en deux catégories : ceux qui sont actifs et transmettent et ceux qui ne le sont pas. La massivité du système, l'absence de contrôle de puissance à l'émission et/ou d'orthogonalité des séquences pilotes des utilisateurs sont autant de caractéristiques qui compliquent les traitements en réception.

Cette thèse a pour thème général l'étude de nouvelles méthodes statistiques basées sur des algorithmes à passage de messages sur des graphes factoriels (factor graphs) appropriés afin de traiter conjointement toutes ces tâches au niveau du récepteur.

Sont étudiées plus précisément : - une méthode (1) d'inférence bayésienne hybride à base de l'algorithme de propagation de croyance (belief propagation algorithm, BP) et de l'algorithme de propagation de l'espérance (expectation propagation algorithm, EP) pour résoudre le problème conjoint de détection d'activité, estimation de canal, et détection multi-utilisateur dans un système GF-NOMA synchrone avec absence de contrôle de puissance à l'émission, séquences pilotes orthogonales et antennes réceptrices multiples. En introduisant un critère d'approximation pour exprimer le passage de messages sous forme de lois gaussiennes, l'estimation du canal et la détection multi-utilisateurs peuvent être traitées efficacement par l'algorithme EP. Ceci s'avérant impossible sous cette forme pour la détection d'activité des utilisateurs, un passage de messages sous forme BP est utilisé à cet effet. La méthode proposée inclut une étape d'estimation des hyperparamètres du modèle que sont l'énergie des signaux reçus et la corrélation spatiale entre les antennes réceptrices. Une variante à complexité réduite ignorant la corrélation spatiale entre antennes réceptrices est également proposée ; - une méthode (2) d'inférence bayésienne à base de l'algorithme EP exploitant des méthodes d'analyse complexe (en utilisant l'approximation de la série de Taylor basée sur le calcul de Wirtinger) permettant de traiter la détection d'activité des utilisateurs également sous la forme d'un algorithme à passage de messages gaussiens ; - une méthode (3) faisant précéder la méthode (2) d'une méthode d'acquisition comprimée bayésienne comme l'apprentissage bayésien clairsemé multiple structuré (S-MSBL), la recherche de correspondance orthogonale simultanée par blocs (B-SOMP) ou la transmission de messages approximatifs (AMP) chargée de l'estimation initiale du canal et de l'activité des utilisateurs dans le contexte complexifié d'un accès massif avec séquences pilotes des utilisateurs non-orthogonales.

L'évaluation par simulations de ces différentes méthodes est effectuée dans le cas particulier d'un système GF-NOMA synchrone par codage, entrelacement et modulation OFDM (GF-OFDM-IDMA). Les performances obtenues (mesurées en termes de taux d'erreur binaire résiduel pour la détection et le décodage, d'erreur quadratique moyenne pour l'estimation de canal, et de probabilités de fausse alarme et de non-détection pour la détection d'activité) se comparent favorablement par rapport à celles obtenues avec des méthodes classiques publiées dans la littérature.

Abstract

The proliferation of wireless communication technologies has led to unprecedented demands for Ultra-reliable low latency communications (URLLC) and Massive machine type communications (mMTC) [54] to support diverse applications ranging from industrial automation to Internet of Things (IoT). Non-orthogonal multiple access (NOMA) has emerged as a promising solution to enhance spectrum efficiency [55] and accommodate the varying requirements of URLLC and mMTC. This thesis explores the application of message passing algorithms in the context of NOMA, coupled with grant-free access, to address the stringent latency and connectivity demands of modern wireless systems.

The thesis begins by presenting a comprehensive overview of the challenges and requirements posed by URLLC and mMTC scenarios. It discusses the limitations of conventional orthogonal multiple access techniques in meeting these demands and introduces NOMA as an alternative solution. Then, theoretical foundations of message passing algorithms are introduced, providing a solid framework for their integration into receiver design for a NOMA-based systems. The thesis further explores the utilization of belief propagation, expectation propagation, and related techniques to to cope with multi-access interference, imperfect power control and antenna correlation in NOMA-based systems.

A timely aspect of this work is the exploration of grant-free access mechanisms in conjunction with NOMA. Grant-free access introduces a paradigm shift in wireless communication, enabling devices to transmit without prior explicit allocation, thus reducing latency and overhead. The thesis investigates the benefits and challenges of incorporating grant-free access into NOMA-based systems and demonstrates its effectiveness in the context of URLLC and mMTC requirements.

The proposed message passing algorithms are evaluated through extensive simulations and performance analysis under various scenarios. The results highlight the improved spectral efficiency, latency reduction, and reliability achievable by the integration of message passing algorithms and grant-free access in NOMA. Furthermore, the thesis discusses practical implementation considerations and provides insights into the design of NOMA-enabled wireless networks to support future 5G and beyond systems.

In conclusion, this thesis contributes to the advancement of wireless communication technologies by showcasing the synergy between message passing algorithms, NOMA, and grant-free access for meeting the stringent demands of URLLC and mMTC. The presented findings will hopefully provide valuable insights for researchers, engineers, and practitioners working on the design and optimization of wireless networks to support a diverse range of emerging applications [55].

Chapter 1

Introduction

Let us start with a summary of the main topics closely tied to the underlying context of this thesis. This chapter begins with an overview of the 5G communication and beyond, which aims to provide high-speed, reliable and low-latency communication for various applications and devices.

One of the key techniques to achieve these goals is grant-free access, which allows users to transmit data without prior reservation of resources. Grant-free access can reduce the signaling overhead and latency [56], and increase the flexibility and scalability of the system. However, grant-free access also poses challenges such as user activity detection, channel estimation, multi-user detection and decoding. Accordingly, this chapter summarizes the fundamentals of grant-free access as a new paradigm for future wireless networks.

The chapter then focuses on existing multi-access schemes with an emphasis on NOMA as a promising physical layer technique that enables grant-free access.

Finally, the chapter concludes with a summary of the challenges addressed in this thesis, along with the contributions and a list of publications.

1.1 5G Communication and beyond

5G communication, the fifth generation of wireless technology, represents a significant leap forward in the realm of connectivity. It enables faster speeds, lower latency, increased network capacity and the ability to support an unprecedented number of connected devices. These advancements pave the way for revolutionary applications and services across various industries [32].

One of the key aspects of 5G is its ability to provide Enhanced Mobile Broadband (eMBB), i.e higher data rates compared to previous generations (see Sec. 1.1.1 for more details). With download speeds reaching several gigabits per second, 5G enables ultra-fast streaming, quick file downloads, and seamless use of bandwidth-intensive applications such as Virtual reality (VR), Augmented reality (AR), and 4K/8K video streaming. The enhanced speed ensures a smoother user experience and opens up opportunities for immersive digital content consumption [49].

Furthermore, 5G incorporates URLLC to cater to mission-critical applications [52] (see Sec. 1.1.2 for more details). Latency refers to the delay between the transmission and reception of data. 5G aims to deliver ultra-low latency, reducing delays to mere milliseconds. This is particularly essential for real-time applications that demand immediate responsiveness. Industries like healthcare, transportation, and emergency services rely on extremely reliable and responsive connectivity. With URLLC, 5G networks provide the necessary infrastructure

for applications like remote surgeries, autonomous transportation, disaster response, and industrial automation. These sectors demand high availability, fast response times, and robust connectivity, all of which 5G communication delivers.

The concept of mMTC is another integral part of 5G (see Sec. 1.1.3 for more details). It allows for seamless communication between a large number of low-power, low-cost devices [53]. This is particularly useful for applications that require vast sensor networks and distributed monitoring, such as environmental monitoring, asset tracking, and precision agriculture. The mMTC aspect of 5G enables energy-efficient and cost-effective data exchange, facilitating the growth of IoT ecosystems and enabling innovative solutions. This capability is pivotal for the proliferation of smart cities, where a vast array of interconnected devices, from traffic sensors to smart streetlights, collaborate to enhance urban life and sustainability. Similarly, industries can leverage this capacity to facilitate efficient automation, predictive maintenance, and real-time monitoring in sectors like manufacturing, energy, and agriculture [57].

To meet diverse requirements across industries, 5G introduces the concept of network slicing [59]. Network slicing allows network resources to be dynamically allocated and customized to suit specific applications or use cases. It enables service providers to create virtual network slices that cater to specific demands, such as varying bandwidth, latency, security levels, and quality of service. This flexibility enables efficient resource utilization and ensures that different industries or applications receive tailored connectivity that matches their unique needs. For example, a smart city network slice may prioritize low-latency communication for traffic management, while an industrial network slice may emphasize high reliability and data security for critical machine-to-machine communication.

Antenna technologies play a crucial role in realizing the potential of 5G communication. Two prominent techniques employed in 5G are beamforming [60] and Multiple Input Multiple Output (MIMO) [61],[62]. Beamforming enables targeted signal transmission and reception by focusing the energy in a specific direction. It utilizes an array of antennas and adjusts the phase and amplitude of signals to create constructive interference in the desired direction while minimizing interference in other directions. Beamforming enhances signal strength, extends coverage, and improves spectral efficiency, making it vital for achieving reliable and high-speed connectivity in 5G networks. MIMO, on the other hand, leverages multiple antennas to transmit and receive multiple data streams simultaneously. It enables spatial multiplexing, where data streams are separated in space, allowing for increased data rates and spectral efficiency. By utilizing multiple antennas, MIMO enhances the overall capacity and throughput of 5G networks. It is particularly beneficial in environments with high user density and multipath propagation, where signals bounce off obstacles and arrive at the receiver via multiple paths.

In terms of frequency spectrum, 5G utilizes a range of bands, including both sub-6 GHz frequencies and Millimeter Wave (mmWave) frequencies. Sub-6 GHz bands offer wider coverage and better penetration through obstacles, making them suitable for delivering 5G services across larger areas. On the other hand, mmWave frequencies provide extremely high bandwidth and capacity but have shorter range and are more susceptible to blockage by obstacles. Deploying mmWave-based 5G networks requires the deployment of small cells and denser network infrastructure to ensure consistent coverage. These higher frequency bands are essential for meeting the ever-increasing demand for data and supporting emerging applications that rely on ultra-fast speeds.

Overall, 5G communication represents a transformative shift in wireless connectivity. Its advancements in speed, latency, capacity, and device connectivity unlock a vast array of possibilities in various sectors. Industries such as healthcare, transportation, manufacturing, entertainment, and smart cities can harness the power of 5G to revolutionize their operations, deliver innovative services, and improve the quality of life for individuals worldwide. With its potential to connect billions of devices, enable real-time interactions, and empower new applications, 5G communication holds the key to a more connected and technologically advanced

future.

1.1.1 Enhanced Mobile Broadband (eMBB)

Enhanced Mobile Broadband (eMBB) is another feature defined for 5G [34] by the 3GPP as part of its Study on New Services and Markets Technology Enablers (SMARTER) project [35]. The goal of SMARTER was to develop high-level use cases and identify the features and capabilities that 5G would need to provide to enable them.

eMBB works by using advanced technologies such as:

- Higher spectrum bands: eMBB can utilize mmWave frequencies above 24 GHz, which offer much more bandwidth than the lower bands used by 4G. However, mmWave signals have limited range and penetration, so they require more base stations and antennas to provide coverage.

- Massive MIMO: eMBB can employ massive MIMO technology, which uses hundreds of antennas at each base station to transmit and receive signals simultaneously. This increases the capacity and efficiency of the network, as well as the signal quality and reliability.

- Beamforming: eMBB can leverage beamforming technology, which directs the radio signals towards the intended users, rather than broadcasting them in all directions. This improves the signal strength and reduces interference and power consumption.

- Dynamic Spectrum Sharing: eMBB can benefit from Dynamic Spectrum Sharing (DSS) technology, which allows 4G and 5G to coexist on the same spectrum band. This enables a smooth transition from 4G to 5G without requiring additional spectrum allocation or network deployment [36].

eMBB can deliver several benefits for users and businesses, such as:

- Faster data transfer rates and faster network experience: eMBB can support peak data rates of up to 20 Gbps for downlink and 10 Gbps for uplink, as well as user experienced data rates of 100 Mbps for downlink and 50 Mbps for uplink. This means that users can enjoy seamless streaming of high-definition video [37], cloud gaming, virtual reality and other bandwidth-intensive applications.

- Truly immersive VR, AR and 360-degree video experiences: eMBB can provide the high bandwidth and low latency needed to create realistic and interactive virtual environments. For example, users can watch live sports or concerts in 360-degree video, or explore distant places in VR. Businesses can also use eMBB to enhance their training, education, entertainment and marketing offerings with VR and AR.

- Broadband everywhere: Technologies like Fixed Wireless Access (FWA) can offer consistent coverage around the world with minimum speeds of 50 Mbps. This can enable access to broadband services in rural areas, remote locations and developing regions, where wired infrastructure is lacking or costly. FWA can also provide a backup solution for wired broadband in case of network failures or natural disasters.

- Public transportation: Broadband access on high-speed trains and other modes of public transport are examples of eMBB use cases. eMBB can provide reliable connectivity for passengers who want to work, play or communicate on the go. It can also enable smart transportation systems that can improve safety, efficiency and sustainability.

1.1.2 Ultra reliable and low latency communication (URLLCs)

Ultra-Reliable and Low-Latency Communications (URLLC) is a key feature of 5G communication that aims to provide extremely reliable connectivity with minimal latency [33]. URLLC is crucial for applications that require high availability, fast response times, and robust communication links, particularly in mission-critical scenarios.

When we talk about "ultra-reliable," it means that the communication link is highly dependable, with a very low probability of failure or disruption. In the context of URLLC, this reliability is typically quantified using metrics such as the Block Error Rate (BLER) or the Frame Error Rate (FER). These metrics measure the probability of errors occurring in the received data blocks or frames. For URLLC applications, the target is to achieve an extremely low error rate, often in the range of 10^{-5} to 10^{-9}, ensuring high reliability.

Low-latency refers to the minimized delay in transmitting and receiving data packets over the network. In the context of URLLC, latency is often measured as the Round-Trip Time (RTT), which is the time taken for a packet to travel from the source to the destination and back. URLLC applications require very low-latency connections to ensure real-time responsiveness. The target for URLLC is to achieve ultra-low latency, typically in the range of 1-10 milliseconds, enabling near-instantaneous communication and rapid data exchange.

To understand the significance of URLLC, let us consider a few statistics and examples:

1. Autonomous Vehicles: Autonomous driving relies heavily on real-time communication for instant decision-making and coordination. With URLLC, vehicles can exchange critical safety information, such as collision warnings or traffic updates, with ultra-low latency. This helps enhance road safety and enables seamless cooperation among vehicles. For instance, a study conducted by Thales [38] showed that URLLC reduced the latency of vehicle-to-vehicle communication from about 200 milliseconds (in 4G) to just 1 millisecond (in 5G), improving responsiveness and enabling faster reaction times.

2. Industrial Automation: In industries like manufacturing and robotics, URLLC plays a vital role in enabling precise and time-sensitive control systems. With ultra-reliable and low-latency connections, robots can receive instructions and respond in real-time, leading to more efficient and safe automation processes.

3. Emergency Services: When it comes to emergency situations, rapid and reliable communication is critical. URLLC enables emergency responders to exchange vital information, such as location data or medical records, in real-time with low latency. This facilitates swift decision-making and enhances the effectiveness of emergency response efforts. A report by the 5G 5G Automotive Association (5GAA) highlights the importance of URLLC in emergency services, stating that the ultra-reliable and low-latency capabilities of 5G can significantly reduce emergency response times, potentially saving lives [40].

4. Remote Surgeries: Telemedicine and remote surgeries rely on secure and ultra-reliable connections to enable doctors to remotely control robotic surgical systems. URLLC ensures that the surgeon's commands are transmitted with minimal delay and without errors, allowing for precise and real-time control of surgical procedures.

These examples highlight the significance of ultra-reliable and low-latency communications in various sectors. URLLC's ability to provide dependable and real-time connectivity opens up new possibilities for applications that require high reliability, fast response times, and seamless interaction. By leveraging the capabilities of URLLC, 5G communication enables transformative use cases, enhances safety, improves efficiency, and facilitates innovation across industries.

1.1.3 Massive machine-type communications (mMTC)

Massive Machine-Type Communication (mMTC) is a fundamental concept in 5G communication that focuses on enabling seamless connectivity for a massive number of low-power, low-cost devices. It is a key feature of the Internet of Things (IoT) ecosystem, where billions of devices are interconnected to enable smart applications and services.

To understand the significance of mMTC, let us delve into its details and explore some relevant statistics:

1. Scale of Connectivity: mMTC aims to support an unprecedented scale of device connectivity. According to Ericsson's Mobility Report [39], it is estimated that by 2025, there will be around 5 billion IoT devices connected globally. These devices span various sectors, including smart homes, smart cities, industrial automation, agriculture, healthcare, and transportation. mMTC enables efficient communication between these devices, facilitating data exchange, monitoring, and control on an enormous scale.

2. Device Density: mMTC addresses the challenge of connecting a large number of devices within a limited area. For instance, in smart city applications, numerous sensors, actuators, and smart infrastructure components need to communicate and coordinate with each other. mMTC ensures that these devices can efficiently transmit and receive data, enabling effective smart city management and services.

3. Low-Power and Low-Cost Devices: mMTC focuses on connecting devices that have limited power and computational capabilities, and often operate on batteries for extended periods. These devices are typically cost-sensitive, making it essential to optimize their energy consumption and minimize their communication overhead. For instance, sensors used in environmental monitoring, asset tracking, or agriculture may need to operate autonomously for long durations. mMTC provides energy-efficient communication protocols and mechanisms that allow these devices to exchange data while preserving battery life.

4. Data Traffic Volume: The scale of mMTC results in a massive volume of data traffic generated by interconnected devices. According to a study by Cisco [41], it is estimated that IoT devices will generate approximately 175 zettabytes (1 zettabyte = 1 trillion gigabytes) of data per year by 2025. This data encompasses sensor readings, status updates, telemetry data, and more. mMTC enables efficient data transmission, compression, and aggregation techniques to handle this exponential growth in data traffic, ensuring that network resources are utilized effectively.

5. Application Areas: mMTC has numerous applications across various economic sectors. For example, in agriculture, mMTC enables farmers to monitor soil conditions, crop health, and irrigation systems through a network of low-cost sensors, improving crop yield and resource efficiency. In healthcare, mMTC facilitates remote patient monitoring, wearable devices, and telemedicine solutions, allowing healthcare providers to deliver personalized and timely care. In transportation, mMTC enables vehicle-to-vehicle communication, traffic management systems, and autonomous driving capabilities, enhancing safety and efficiency on the roads.

6. Network Efficiency: mMTC requires efficient utilization of network resources to accommodate the massive number of connected devices. Technologies like Narrowband Internet of Things (NB-IoT) [42] and Long-Term Evolution Machine Type Communication (LTE-M) [43] have been developed specifically for mMTC to optimize power consumption, coverage, and capacity. These technologies provide extended coverage, increased device density, and improved battery life, enabling reliable and cost-effective connectivity for mMTC applications.

In summary, mMTC in 5G communication addresses the challenges of connecting a vast number of low-power, low-cost devices in diverse industries. Its ability to handle the scale,

density, and unique requirements of IoT deployments is crucial for enabling smart applications, improving operational efficiency, and driving innovation. With mMTC, the vision of a fully interconnected and intelligent world becomes a reality.

1.1.4 The 3GPP

The 3rd Generation Partnership Project (3GPP) is a global organization responsible for standardizing cellular communication technologies, including the development of 5G networks. The 3GPP standardization process plays a crucial role in defining the specifications, protocols, and requirements that enable interoperability and compatibility among different vendors and network operators.

The standardization of 5G within 3GPP has involved multiple phases and releases. Let's delve into the details of these phases and their key components, supported by relevant statistics and references:

1. Release 15:

3GPP's Release 15 [44], finalized in December 2017, focused on the initial specifications for 5G, primarily targeting Non-Standalone (NSA) 5G deployments. NSA 5G refers to the integration of 5G with existing 4G LTE networks. The key components of Release 15 included:

- **New Radio (NR)**: Release 15 introduced the 5G NR air interface, which operates in both sub-6 GHz frequency bands and mmWave frequencies. NR offered significant improvements over previous generations, including higher data rates, reduced latency, and increased network capacity.
- **Non-Standalone (NSA) and Standalone (SA) Architecture:**
  - **NSA Architecture:** 5G is initially deployed alongside 4G LTE infrastructure, where LTE handles control signaling, and 5G focuses on data transmission.
  - **SA Architecture:** Fully independent 5G networks allow for complete deployment of 5G capabilities such as low latency and network slicing.

According to Ericsson's Mobility Report [45], by the end of 2021, there were around 230 commercial 5G NR networks deployed worldwide. The report also projected that by the end of 2026, the number of 5G subscriptions would reach 5.8 billion globally, covering approximately 60% of the world's population.

2. Release 16:

3GPP's Release 16 [46], finalized in July 2020, built upon the initial specifications and introduced SA 5G. SA 5G refers to a fully independent 5G network architecture that does not rely on existing 4G infrastructure. Release 16 introduced several key features, including:

- **Network Slicing:** Release 16 standardized network slicing, allowing the creation of virtualized networks tailored to specific requirements. Network slicing enables efficient resource allocation and support for diverse use cases, such as low-latency applications, high-bandwidth applications, and massive IoT deployments.
- **URLLC Enhancements:** Release 16 further improved URLLC capabilities, enabling critical and delay-sensitive applications. It focused on reducing latency, increasing reliability, and ensuring seamless connectivity for applications such as autonomous vehicles,

industrial automation, and remote surgery.

- **mMTC Enhancements:** Release 16 addressed the connectivity needs of a massive number of low-power IoT devices. It introduced enhancements to support efficient connectivity, energy optimization, and scalability for IoT deployments across industries like agriculture, smart cities, and healthcare.

- **Vehicle-to-Everything (V2X) Communications:** Release 16 standardized V2X communications, facilitating direct communication between vehicles, infrastructure, pedestrians, and other road users. V2X enables advanced driving assistance systems, improved road safety, and cooperative traffic management.

1.2 Grant-free access: a new paradigm for 5G and beyond

Grant-free access in wireless communication is a scheme that allows users to transmit data without waiting for a grant or a scheduling request from the network. This can reduce the latency and the signaling overhead, especially for small packet transmissions, URLLC and mMTC. There are different ways of implementing grant-free access in wireless communications, such as:

- Grant-free random access (GF-RA): This is a Contention-based multiple access (CMA) scheme that uses a slotted structure and preambles to access the channel. GF-RA can support a large number of devices with sporadic traffic, but it suffers from low throughput and high collision rate [47].

- Grant-free non-orthogonal multiple access (GF-NOMA): This is a Non-contention-based multiple access (NCMA) scheme that uses Code-domain non-orthogonal multiple access (CD-NOMA) to spread user data with non-orthogonal signatures (see Sec. 1.3.2 for more details). A signature is a sequence of bits or symbols that identifies a device or a user in the channel. For example, a signature can be a Gold code, a Walsh code, or a Zadoff-Chu sequence, etc. GF-NOMA can achieve higher throughput and overloading performance than GF-RA, but it requires advanced receivers and channel estimation techniques.

- Grant-free Multiple Acces (GFMA): This is a hybrid scheme that combines CMA and NCMA features. GFMA uses a random signature assignment mechanism to allocate non-orthogonal signatures to devices [47]. GFMA can achieve better trade-off between throughput and collision rate than GF-RA and GF-NOMA, but it requires coordination among devices and base station.

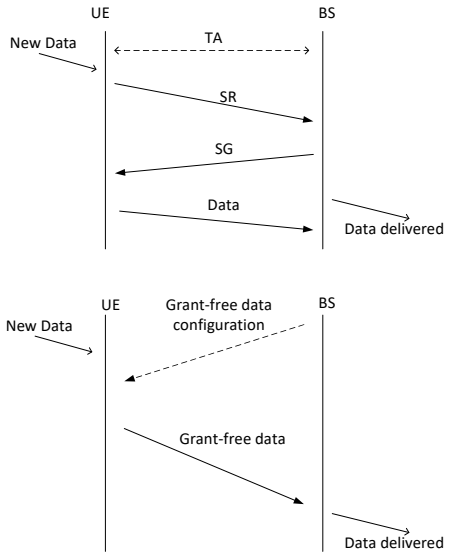

Figure 1.1: Grant-based acces on top and grant-free access on bottom.

Both grant-based access and grant-free access are shown in Fig. 1.1). In grant based access, multiple messages have to be transmitted between the devices and base station in order to allocate resources before the data transmission takes place. It involves the following steps.

- The user buffers data for the transmission
- It then waits for the prescheduling opportunity to send the Scheduling Request (SR)
- User equipment then sends the scheduling request
- Base station sends the Scheduling Grant (SG)
- Timing Advance (TA) is adjusted, based on the distance of the user from the base station.
- Finally the user equipment responds to the scheduling grant and transmits the data.

This results in high delay in data transmission and high energy consumption. That is why, we focus on grant free access, which reduces the signaling overhead caused by the hand-shaking protocols.

If we make a comparison between the pros and cons of grant-based access and grant-free access, they can be summarized as follows.

Grant-based access has the following pros.

- **Guaranteed bandwidth:** It ensures that each user is allocated a certain amount of bandwidth.
- **Lower interference:** As each user a priori demands the allocation of resources, so only one user is transmitting the data at a time which results in low to no interference.
- **Works well in stable channel conditions:** Grant-based access works very well in stable and predictiable channel conditions because it relies on coordinated scheduling mechanism thus optimizing resource allocation which is useful for the applications that have to be reliable [144].

Some of the cons of grant based access are:

- **Overhead and delay:** It introduces additional overhead and latency, since the user has to send a request and wait for a response from the network before using a resource.
- **Unsuitable for sporadic traffic:** It may not be suitable for bursty or sporadic traffic, since the user may not be able to predict when and how much resource they need in advance.
- **Unsuitable for dynamic channel conditions:** It may not be flexible enough to adapt to dynamic network conditions, since the resource allocation is fixed once granted.

Some of the pros of grant free access are:

- **Overhead and latency:** It reduces overhead and latency, since the user can use a resource without any prior reservation or signaling.
- **Accommodation of bursty or sporadic traffic:** It can accommodate bursty or sporadic traffic, since the user can use a resource whenever needed.

Some of the cons of grant free access are:

- It may cause collisions and interference among users, since multiple users may use the same resource at the same time without coordination.
- It may degrade the network efficiency and throughput, since the resource utilization may not be optimal due to random access.
- Grant-free access may struggle to support Quality of Service (QoS) guarantees because users may not obtain adequate resources for optimal performance. This issue arises from the dynamic nature of network loads, which fluctuate over time. Resource allocation may fail to adapt effectively to these variations, leading to inefficiency. Under light network loads, resources may be underutilized, while during heavy loads, there is a higher risk of collisions and degraded performance, as the system cannot allocate resources efficiently to meet the needs of all users [48].

In conclusion, grant based access and grant free access have different advantages and disadvantages, and there is no one-size-fits-all solution for all network scenarios and user requirements. In general, a trade-off analysis is needed to determine which method is more suitable for a specific application.

1.3 Overview of multiple access schemes

Orthogonal multiple access (OMA) and NOMA are two different multiple access schemes used in wireless communication systems.

OMA: In OMA, each user is assigned a separate, non-overlapping portion of the frequency, time, or code resources. The key idea is that users' transmissions are orthogonal to each other, meaning they do not interfere with one another (see Sec. 1.3.1 for more details).

NOMA: NOMA takes a different approach by allowing multiple users to share the same frequency, time, or code resources simultaneously and non-orthogonally (see Sec. 1.3.2 for more details).

1.3.1 Types of OMA schemes

There are different types of OMA schemes that are detailed below.

1. Frequency Division Multiple Access (FDMA): FDMA is a multiple access scheme that divides the available frequency spectrum into non-overlapping frequency bands. Each user is allocated a unique frequency band for communication [50]. FDMA provides a fixed allocation of frequency resources to each user, ensuring that they do not interfere with one another. This scheme is commonly used in analog cellular systems and early digital systems like 2G (GSM). FDMA offers simplicity and deterministic resource allocation, but it may suffer from inefficiency when the traffic load is low or when users have varying data rate requirements.

2. Time Division Multiple Access (TDMA): TDMA is a digital multiple access scheme that divides the available frequency band into discrete time slots [50]. Each user is assigned a specific time slot for transmission. Users take turns to transmit in their allocated time slots, thereby sharing the same frequency band. TDMA allows for efficient utilization of the spectrum by dividing it into time intervals. Multiple users can transmit simultaneously by using different time slots. This scheme is used in systems like 2G (GSM) and the North American Digital Cellular (IS-136) standard. TDMA offers flexibility in allocating time slots, support for various data rates, and compatibility with circuit-switched and packet-switched services. However, synchronization among users is crucial for successful transmission, and the capacity may be limited when the number of users increases.

3. Code Division Multiple Access (CDMA): CDMA is a spread spectrum-based multiple access scheme where each user is assigned a unique code or signature [51]. CDMA allows multiple users to transmit simultaneously on the same frequency band and time slot. At the receiver, the desired user's signal is distinguished by the unique code assigned to that user, while signals from other users appear as noise. CDMA offers increased capacity, improved resistance to interference, enhanced security, and support for variable data rates. It is used in 3G systems like CDMA2000 and the Universal Mobile Telecommunications System (UMTS). CDMA allows for flexible allocation of resources and supports a larger number of users compared to FDMA and TDMA.

4. Orthogonal Frequency Division Multiple Access (OFDMA): OFDMA is a multiple access scheme based on multi-carrier modulation. It divides the available frequency band into multiple orthogonal subcarriers. Each user is assigned a subset of subcarriers for transmission, and multiple users can transmit simultaneously by using different subsets of subcarriers. OFDMA allows for flexible allocation of subcarriers, enabling efficient adaptation to varying channel conditions and user requirements. It provides a balance between frequency-selective fading mitigation, spectral efficiency, and robustness against interference. OFDMA is a key technology used in 4G (LTE) and 5G wireless systems. It offers high data rates, support for low-latency applications, improved spectral efficiency, and scalability for a large number of users and diverse services. OFDMA can accommodate users with different bandwidth requirements and allows for dynamic resource allocation based on channel conditions and user demands.

These OMA schemes provide different approaches to multiple access in wireless communication systems, enabling efficient utilization of the available spectrum and supporting simultaneous communication among multiple users. Each scheme has its own characteristics and advantages, making them suitable for different network scenarios and technologies.

1.3.2 Types of NOMA schemes

NOMA schemes can be broadly classified into two types: power-domain NOMA and code-domain NOMA. In power-domain NOMA, users are multiplexed in the power domain by using superposition coding at the transmitter and successive interference cancellation at the receiver.

In code-domain NOMA, users are multiplexed in the code domain by using low-density spreading codes, interleavers, sparse codes, or pattern codes.

At the receiver side of NOMA schemes, advanced signal processing techniques are employed to decode the signals from different users. One commonly used technique is successive interference cancellation (SIC). The receiver begins by decoding the signal from the user with the strongest received power level. At this stage, the signals from other users are treated as interference. Once the signal from the first user is decoded, it is subtracted from the received signal, effectively removing the interference caused by that user. This interference cancellation step reduces the impact of the strongest user's signal on the subsequent decoding process. The receiver then proceeds iteratively, decoding the signals from users in decreasing order of their power levels and subtracting the contributions of previously decoded users from the received signal. This iterative process allows the receiver to separate the signals from different users, even though they are transmitted simultaneously and share the same time-frequency resource. By iteratively removing the decoded signals, the receiver progressively uncovers the signals of the remaining users until all user signals are successfully decoded.

Power domain NOMA

In power domain NOMA [58], the base station or access point allocates different power levels to individual users based on factors such as their channel conditions, quality-of-service requirements, or priority levels. This power allocation strategy ensures that users with better channel conditions or higher priority receive higher power allocations, while users with poorer channel conditions or lower priority are assigned lower power levels. By assigning different power levels, NOMA exploits the varying channel conditions among users to enhance the system's overall performance.

Code domain NOMA

There are different types of code domain NOMA schemes. We describe the most common ones below (see [63] for a recent survey).

1. Low Density Spreading (LDS) systems [64]: multiple users are superimposed using sparse spreading sequences (i.e. containing well-designed zeros) in order to facilitate overloading. Well-known examples include LDS-CDMA and LDS-OFDM, where the sequences are transmitted over time and frequency resources, respectively.

2. Sparse Code Multiple Access (SCMA) [65]: SCMA is an improved version of LDS systems assigning to each user a different codebook generated by multi-dimensional constellations in order to enhance the interference cancellation capability.

3. Pattern Division Multiple Access (PDMA) [66]: each user's transmitted data is repeated over a resource group. The resource groups assigned to different users are designed to minimize the overlap in the time, frequency or spatial domain.

4. Multi-User Shared Access (MUSA) [67]: MUSA is a CDMA-style system using a large number of complex M-ary spreading sequences, designed to reach low cross-correlation so as to reach good user overloading performance.

5. Interleaved-Division Multiple Access (IDMA) [68]: each user's transmitted bit is first multiplied by a binary spreading sequence before undergoing user-specific bit-interleaving to separate the users at the receiver side. After modulation, the obtained complex symbol sequences of all users are multiplexed either in the time (basic IDMA) or frequency (OFDM-IDMA) domain.

1.3.3 Comparison of OMA and NOMA schemes

We present some elements of comparison between OMA and NOMA:

1. Spectral Efficiency: OMA schemes share the resources among the users in an orthogonal way potentially leading to underutilization. In contrast, NOMA allows for non-orthogonal resource allocation, enabling multiple users to share the same time-frequency-code resources, resulting in higher spectral efficiency.

2. Capacity and User Scaling: in OMA schemes, as the number of users increases, the available resources must be divided among more users, which can lead to reduced capacity per user. Since this is not the case in NOMA schemes, a larger number of users can be supported with improved capacity.

3. Interference Management: In OMA, interference can arise from users operating in the same frequency band or time slot, potentially degrading the quality of communication. NOMA, on the other hand, exploits the interference by allowing users to decode their intended signals and treat the interference as useful signals. This interference management capability of NOMA can significantly improve the overall system performance.

It is important to note that while NOMA offers advantages over OMA, it also introduces additional complexities in terms of resource allocation, power control, and interference cancellation techniques.

Overall, NOMA is a promising technique for future wireless communication systems, especially in scenarios with a high number of users and limited bandwidth resources. However, its implementation can be challenging due to the need for advanced signal processing techniques and the potential for interference among users. The advantages of NOMA are massive connectivity, fulfilling low latency requirements and high spectral efficiency.

1.4 Challenges

In this thesis, we consider the challenges in designing receiver structures for grant-free access with code-domain NOMA as the physical layer:

Channel Estimation (CE): Channel estimation is the process of determining the characteristics of the communication channel between the transmitters and the receiver. Accurate channel estimation is essential for successful data decoding. In the context of grant-free access, where nonorthogonal users may transmit sporadically, estimating the channel conditions becomes a complex task.

Multi-User Detection (MUD): Multi-user detection involves distinguishing and decoding signals from multiple users sharing the same communication resources. In grant-free scenarios, where multiple users transmit simultaneously, MUD is crucial for separating and decoding the signals from different users.

Decoding (DEC): Data decoding involves recovering the transmitted information digits from the received signal. In grant-free access scenarios, where devices transmit without explicit scheduling, decoding makes the receiver robust to residual interference.

User Activity Detection (UAD): User activity detection is the process of determining at the receiver side, which users are actively transmitting data over the considered resource elements. This problem, inexistent by design in grant-based systems, is due to the absence of explicit scheduling in grant-free access. Accurately estimating user activity is essential for

accurate CE and MUD.

Massive Access: Massive connectivity with a large number of users is one of the key features of 5G and future wireless networks, with important applications such as IoT with potentially millions of users per square kilometer. Since the behavior of such users takes the form of small and sporadic packets transmitted in an uncoordinated way, grant-free access is a natural framework to design wireless networks for this kind of traffic. An important and yet largely unsolved issue consists in the design of receivers ensuring the scalability of joint CE/MUD/UAD/DEC with a large number of users and a low number of receive antennas. This open problem will be addressed assuming devices having low and unknown power transmitting with a low activity rate during a given time slot.

We address all aforementioned aspects jointly by proposing new powerful iterative receiver designs based on message passing.

1.5 Contributions

After a comprehensive review of message passing algorithms for Bayesian estimation in chapter 2 and after a generic system model for sporadic IoT-like traffic over multi-antenna frequency-selective channels having short coherence time is developed in chapter 3, the next chapters introduce novel joint receivers in various grant-free access contexts:

- Leveraging a principled Gaussian approximation, an hybrid message passing receiver is introduced in chapter 4. Importantly, receive antenna correlation is exploited and a pilot-based method is provided to estimate the hyperparameters of the dynamic channel model. The effect of unequal and unknown receive energy is assessed
- A Reduced-complexity (RC) version of the receiver in chapter 4 is developed in chapter 5 by ignoring receive antenna correlation.
- Leveraging a new principled Gaussian approximation using on a Wirtinger calculus-based approximation, an improved version of the receiver in chapter 4 is introduced in chapter 6.
- The receiver in chapter 6 becomes suitable for massive access when initialized with the pilot-based joint UAD/CE estimation Bayesian Compressive Sensing (CS) technique derived in chapter 7.

1.6 List of publications

1.6.1 Journal papers

F. Sagheer, F. Lehmann and A.O. Berthet. A new hybrid message passing algorithm for joint user activity detection, channel estimation and data decoding in grant-free OFDM-IDMA. IEEE Transactions on Vehicular Technology, 2024, 73 (7), pp.10365-10380.

F. Sagheer, F. Lehmann and A. O. Berthet, ”Wirtinger calculus-based expectation propagation in latent variable models applied to grant-free NOMA,” IEEE Signal Processing Letters, vol. 31, pp. 2360-2364, 2024.

F. Sagheer, F. Lehmann and A. O. Berthet, ”Sparse Bayesian Learning for Initial Joint User Activity, Channel and Parameter Estimation for Massive Grant-Free with Non-orthogonal Pilots,” submitted to IEEE access, November 2024.

1.6.2 Conference papers

F. Sagheer, F. Lehmann and A.O. Berthet, “Low-complexity dynamic channel estimation in multi-antenna grant-free NOMA,” Proc. IEEE 95th Vehicular Technology Conference (VTC2022-Spring), Helsinki, Finland, June 2022.

Chapter 2

Background: Message passing algorithms

This chapter provides a brief overview of message-passing algorithms between nodes in a graphical model for the sake of distributed inference. These methods will serve as a workhorse in subsequent chapters to design efficient methods for multi-user detection, decoding, channel estimation and user activity detection in NOMA systems.

2.1 Introduction

Message passing algorithms are a class of techniques that can be used to solve inference problems in graphical models, such as Bayesian networks or Markov random fields. Compared to other techniques, message passing algorithms offer the potential of:

1. Flexibility: Message passing algorithms can handle inference of discrete, continuous and mixed discrete-continuous variables.

2. Scalability: Message passing algorithms can handle inference over graphical models with an increasing number of nodes, while conceptually the same simple computational rules apply.

3. Computational efficiency: Since the essence of message passing is to perform inference using distributed algorithms, they usually require less computational resources than other techniques.

Some of the disadvantages are:

1. They may not converge or may converge to incorrect results on loopy graphs, due to the presence of cycles or feedback loops in the graph. This happens because they assume that the messages are independent and consistent, which is not true for graphs with loops. As a result, they may oscillate between different values, or converge to a fixed point that does not correspond to the true marginal distribution.

2. They may suffer from numerical instability or overflow/underflow issues, especially when dealing with high-dimensional or multimodal distributions.

3. They may have difficulty in representing complex distributions that have multiple modes or peaks, which can lead to inaccurate or misleading results.

4. They may require a large number of iterations or messages to reach a satisfactory level of accuracy or convergence. This depends on the structure and complexity of the graphical

model, as well as the desired precision of the inference. In some cases, they may need to send and receive hundreds or thousands of messages before they converge to a stable solution, which can be costly in terms of time and resources.

2.2 The factor graph formalism

We use the elegant and convenient formalism of factor graphs [76] to represent graphical models. From a probabilistic modeling point of view, we aim at visualizing conditional independencies among subsets of hidden random variables.

Consider a generic Bayesian inference problem, where X and $Y_1, Y_2, \dots$ denote the hidden random variables (or vectors) of interest. Assume the associated posterior distribution admits a factorization of the form

$$p(x, y_1, y_2, \dots) \propto f(x, y_1, y_2, \dots) \times h_1(x, \dots) \times h_2(x, \dots) \times \dots, \tag{2.1}$$

$x, y_1, y_2, \dots$ denote the arguments of the posterior distribution corresponding to the hidden variables $X, Y_1, Y_2, \dots$. $f(.)$, $h_1(.)$, $h_2(.)$ are local functions depending on a subset of the arguments corresponding to hidden variable nodes, and the notation $\propto$ means "is proportional to."

A factor graph, is by definition a triplet of the form (V, L, E), where $V = (x, y_1, y_2, \dots)$ (resp. $L = (f, h_1, h_2, \dots)$) is the set of arguments (resp. the set of local functions) in (2.1) and E is a set of edges where every edge connects a vertex in V to one or more in L. A variable node (depicted by a circle) is a node in the graph representing an element of V. A function node (depicted by a square) is a node in the graph representing a local function in L. The construction of E consists in drawing a connection between variable node x and function node f iif x is an argument of f. Consequently, the factor graph corresponding to in (2.1) can be represented by Fig. 2.1, where the notation $n(v)$ is used for the set of neighbors of node v in the graph.

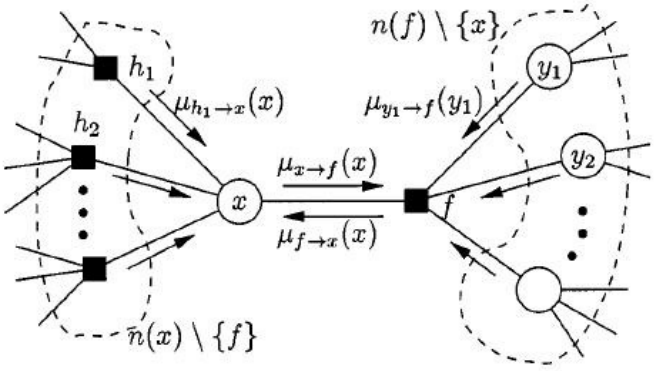

Figure 2.1: Portion of the factor graph corresponding to the posterior (2.1) along with the exchange of messages [76].

The obtained factor graph can in turn be used as the support of a distributed Bayesian inference algorithm, by passing messages over the edges of the graph (see Sec. 2.3). The message sent by node a to node b will be denoted by $\mu_{a \to b}(.)$. Such a message passing algorithm works by passing real valued functions called messages along the edges between the nodes. More precisely, considering x as a variable node and f as a factor node connected to x in the factor graph, then the messages from x to f and the messages from f to x are real-valued functions $\mu_{x \to f}$ and $\mu_{f \to x}$, whose domain is the set of values that can be taken by the random variable (or vector) associated with x, denoted by D_x.

2.3 Families of message passing algorithms

We present three families of message passing algorithm that will be used throughout this thesis. Other well-known families, including mean field and variational Bayes methods, can be found in the survey [27].

2.3.1 Belief propagation

Belief propagation (BP), also known as sum-product message passing, is a message-passing algorithm for performing inference on graphical models, such as Bayesian networks and Markov random fields. It calculates the marginal distribution for each unobserved node (or variable), conditional on any observed nodes (or variables). Belief propagation is commonly used in Artificial Intelligence (AI) and information theory, and has demonstrated empirical success in numerous applications, such as low-density parity-check codes, turbo codes, free energy approximation, and satisfiability [69].

The algorithm was first proposed by Judea Pearl in 1982 [70], who formulated it as an exact inference algorithm on trees, and later extended it to polytrees [71]. While the algorithm is not exact on general graphs with cycles, it has been shown to be a useful approximate algorithm [72].

The two basic rules of the belief propagation (also called sum-product) algorithm [76] when exchanging messages between factor node f and variable node x are recalled below:

- **Message from a factor node to a variable node**

$$\mu_{f\to x}(x) = \sum_{\sim\{x\}} \left(f(n(f)) \prod_{y\in n(f)\setminus\{x\}} \mu_{y\to f}(y) \right) \tag{2.2}$$

  where the summary function $\sum_{\sim\{x\}}$ denotes discrete summation (resp. integration) over all discrete (resp. continuous) variables except x

- **Message from a variable node to a factor node**

$$\mu_{x\to f}(x) = \prod_{h\in n(x)\setminus\{f\}} \mu_{h\to x}(x). \tag{2.3}$$

2.3.2 Expectation propagation

Expectation propagation (EP) is originally a technique in Bayesian machine learning that finds approximations to a probability distribution. [73] It uses an iterative approach that uses the factorization structure of the target distribution.

EP achieves this approximation usually by minimizing the Kullback-Leibler (KL) divergence between the true and target distribution in some family Φ. In the special case where the target distribution is Gaussian, this boild down to computing the mean and covariance of the true distribution.

Interestingly, EP has been extended to message passing over graphical models in [74]. The difference between EP and BP in terms of accuracy can be explained as follows: EP aims to minimize a divergence between the approximate distribution and the exact posterior distribution, which is a measure of how close they are (e.g. the KL divergence). BP, on the other hand, tries to satisfy local consistency conditions between neighboring nodes in the graphical model, which may not guarantee global consistency or optimality. Therefore, EP can achieve higher accuracy than BP by using a global criterion that takes into account all the information

from the model and the data.

The two basic rules of the expectation propagation algorithm [74] when exchanging messages between factor node f and variable node x are recalled below:

- **Message from a factor node to a variable node**

$$\mu_{f\to x}(x) = \frac{\text{proj}_{\Phi}\left(k \cdot \mu_{x\to f}(x) \sum_{\sim\{x\}} f(n(f)) \prod_{y\in n(f)\setminus\{x\}} \mu_{y\to f}(y)\right)}{\mu_{x\to f}(x)}, \tag{2.4}$$

  where k is a generic notation for normalization constants

- **Message from a variable node to a factor node**

$$\mu_{x\to f}(x) = \frac{\text{proj}_{\Phi}\left(k \cdot \mu_{f\to x}(x) \prod_{h\in n(x)\setminus\{f\}} \mu_{h\to x}(x)\right)}{\mu_{f\to x}(x)}, \tag{2.5}$$

  where $\text{proj}_{\Phi}(.)$ is a projection operator over the family of distributions Φ.

A common (but non-unique as emphasized in [74]) way to defined $\text{proj}_{\Phi}(.)$ is to minimize the KL divergence of the argument wrt the family Φ. The KL divergence between two probability distributions p and q is defined as:

$$KL(p \parallel q) = \sum_{x} p(x) \ln \frac{p(x)}{q(x)} \tag{2.6}$$

where,

- $p(x)$ represents the true distribution of the data.
- $q(x)$ represents the approximating distribution in Φ.

The term $\ln \frac{p(x)}{q(x)}$ measures the logarithmic difference between the probability of x under the true distribution p and the approximating distribution q. This difference is then weighted by $p(x)$, reflecting the importance of x under the true distribution. Finally, the sum operator stands for discrete summation or integration, depending on the context.

2.3.3 Approximate message passing

Approximate Message Passing (AMP) [120] is an advanced iterative algorithm designed to solve high-dimensional signal processing problems, particularly those involving compressed sensing and sparse linear regression. The core idea behind AMP is to efficiently recover sparse signals from a relatively small number of linear measurements. This is achieved by leveraging principles from belief propagation and expectation propagation, which are powerful techniques from statistical inference and graphical models. AMP stands out for its computational efficiency and scalability, making it suitable for large-scale applications where traditional methods may falter.

The AMP algorithm begins with an initialization step, where the initial estimate of the signal and the residuals are set. The iterative process involves two main updates in each iteration: the estimate update and the residual update. The estimate update uses the current residuals and applies a nonlinear function, often a denoising function such as soft-thresholding, to refine the signal estimate. The residual update then adjusts the residuals based on the difference between the observed measurements and the current signal estimate. A damping factor and

a correction term are included in these updates to ensure stability and convergence of the algorithm.

One of the remarkable features of AMP is its theoretical foundation, which provides guarantees for the accuracy and convergence of the algorithm under certain conditions. Specifically, the Restricted Isometry Property (RIP) of the measurement matrix plays a crucial role in these guarantees. RIP ensures that the measurement matrix approximately preserves the Euclidean norms of sparse signals, enabling accurate reconstruction. If the measurement matrix satisfies RIP for a sufficiently small constant, AMP can reliably recover the sparse signal even in the presence of noise.

AMP has found extensive applications in various fields due to its robustness and efficiency. In compressed sensing, it is used to reconstruct signals from fewer measurements than traditionally required. In high-dimensional statistics, AMP is applied to sparse regression problems, where the number of predictors can exceed the number of observations. Moreover, its utility extends to image processing tasks, such as denoising and reconstruction, where the signals of interest are often sparse in some domain. The simplicity and effectiveness of AMP make it a valuable tool for tackling complex, high-dimensional problems in modern signal processing and statistical inference.

Chapter 3

System Model

In this chapter, a generic transceiver model for grant-free access in introduced in Sec 3.1, with a complete description of the emitter, wireless channel and receiver.
Apart from the fact that the physical layer is restricted to CD-NOMA, this transceiver model is general enough so that any of the newly proposed receiver designs in the subsequent chapters can be applied.

Then in Sec. 3.2, the problem of estimating hidden variables at the receiver is cast in a graphical model, applying the factor graph framework recalled in Sec. 2.2.

Without loss of generality, Sec. 3.3 will particularize the generic transceiver of Sec 3.1 to an instance CD-NOMA, namely OFDM-IDMA that will be used as a running example for the sake of fair comparison of simulation results throughout the following chapters.

3.1 NOMA-based grant-free access model

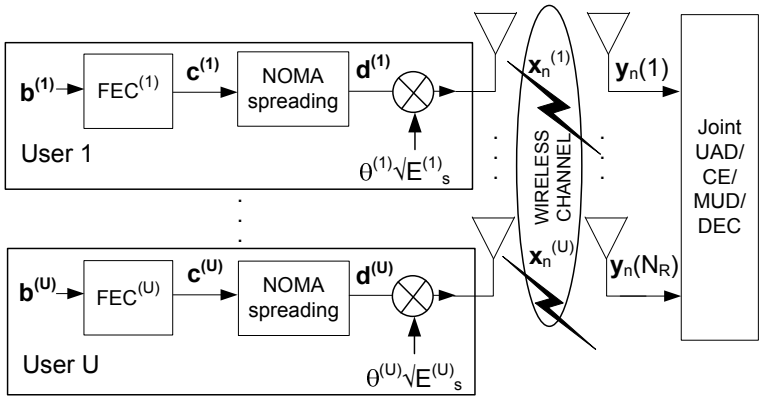

Figure 3.1: NOMA-based grant-free system model.

3.1.1 Transmitter

We consider the single-cell synchronous multi-antenna transmission system depicted in Fig 3.1. Note that any CD-NOMA scheme [63] (see Sec. 1.3.2 for a description of the most common ones) can be used for signalling in the physical layer. U denotes the maximum number of single-antenna users, N denotes the number of Resource Element (RE)s in the time or frequency domain, while N_R denotes the number of receive antennas.

The u-th user information bits $\mathbf{b}^{(u)} \in \{0,1\}^{N_b}$ are uniformly, independently and identically distributed (u.i.i.d.) and transformed to $\mathbf{c}^{(u)} \in \{0,1\}^{N_c}$ after Forward Error Correction (FEC). $\mathbf{c}^{(u)}$ is subsequently converted to a vector of N complex symbols $\mathbf{d}^{(u)} = [d_0^{(u)}, \ldots, d_{N-1}^{(u)}]$ after CD-NOMA spreading and pilot symbol insertion (depending on the chosen CD-NOMA scheme, some coordinates in $\mathbf{d}^{(u)}$ may be *a priori* equal to zero due to codebook sparsity).

3.1.2 Pilot sequences

For the sake of estimating dynamic channels varying over successive REs, scattered pilot sequences are inserted in the transmitted complex symbol vectors. In the sequel, $\mathcal{P}^{(u)} \subset \{0, 1, \ldots, N-1\}$ will denote the subset of pilot REs indices devoted to the u-th user.
Complex pilot symbols are selected as u.i.i.d. from a Quadrature Phase Shift Keying (QPSK) constellation.

Unless otherwise specified, orthogonal scattered pilot sequences (see Fig. 3.2) will be used to force zero inter-user interference over the pilot REs. While pilot orthogonality is useful for ease of initial channel estimation at the receiver side, it comes at the expense of a limitation in the maximum number of users U, that the transceiver can accommodate. In chapter 7, we will allow non-orthogonal scattered pilot sequences over a subset of REs $\mathcal{P}$ that is common to all users, to address massive grant-free access (see Fig. 3.3).

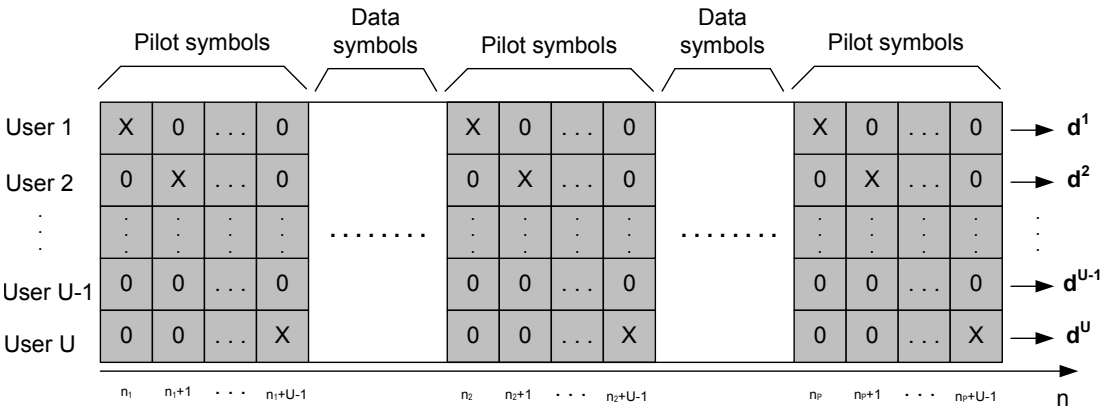

Figure 3.2: Insertion of known orthogonal scattered pilot symbols between data symbol. X denotes a non-zero pilot symbol. A RE (whether in time or frequency) is indexed by the integer n.

3.1.3 Uplink wireless propagation channel

Consider a Base Station (BS) equipped with a Uniform Linear Array (ULA) having N_R antenna elements with antenna separation d and wavelength λ.

In the sequel, the uplink wireless channel between the u-th user (equipped with a single-antenna) and the BS (equipped with a N_R antennas) is considered. As a consequence, the channel coefficients are correlated over the receive antenna elements due to the:

- **Antenna Spacing**: When the distance between antennas is small (usually less than half a wavelength), the signals received by adjacent antennas become more similar due to less spatial diversity, resulting in higher correlation. On the other hand, increasing the antenna spacing reduces correlation and improves the system's diversity.
- **Propagation Environment**: In rich scattering environments, such as urban areas with multiple reflective surfaces, signals arriving at different antennas undergo multiple reflections and diffractions, leading to lower correlation. Conversely, in environments with

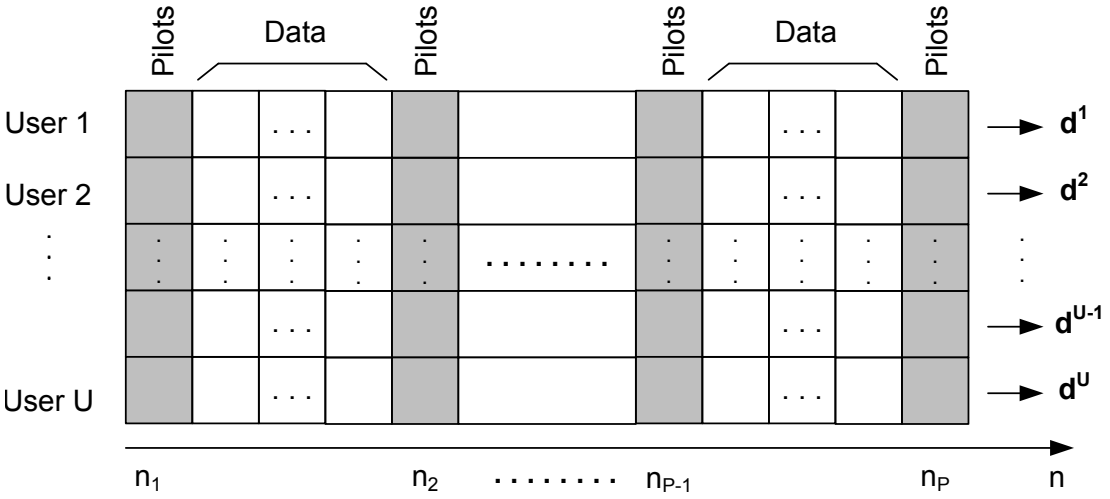

Figure 3.3: Insertion of known non-orthogonal scattered pilot symbols between data symbol. A RE (whether in time or frequency) is indexed by the integer n.

limited scattering (such as open areas), signals take fewer propagation paths, which increases correlation.

- **Angle of Arrival (AoA)**: If signals arrive from a narrow range of directions, this increases correlation between the signals received at different antennas, as they experience similar propagation conditions. A wider AoA distribution leads to less correlation.

Assuming angle-of-arrivals with zero-mean Gaussian spread and small standard deviation $s^{(u)}$, the antenna correlation coefficient is $\rho^{(u)} = \exp[-(2\pi s^{(u)} d/\lambda)^2]$ [75]. A simple example of the receive antenna correlation matrix has the form of a Toeplitz matrix [77], i.e.

$$\mathbf{\Gamma}^{(u)} = \begin{bmatrix} 1 & \rho^{(u)} & \dots & \rho^{(u)N_R-1} \\ \rho^{(u)} & 1 & \dots & \rho^{(u)N_R-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{(u)N_R-1} & \rho^{(u)N_R-2} & \dots & 1 \end{bmatrix}. \tag{3.1}$$

In the sequel, $E_s^{(u)}$ denote the u-th user receive energy accounting for the combined effect of path loss, large-scale fading and power allocation. Also, T_s denotes the sampling period needed to represent continuous-time signals with discrete-time signals without loss of information.

a) Flat fading model

In case the CD-NOMA spread complex symbols are transmitted over consecutive time-domain REs, a flat fading model is considered for the wireless channel.

The u-th user complex baseband equivalent flat fading time-domain channel at instant $t = nT_s$ for rich scattering, can be modeled under Jakes' Doppler power spectrum with Doppler spread f_m as a length-N_R vector [78]

$$\mathbf{h}_n^{(u)} = \sqrt{E_s^{(u)}}\mathbf{\Gamma}^{(u)1/2}\mathbf{g}_n^{(u)}, \tag{3.2}$$

where $\mathbf{g}_n^{(u)}$ is a zero-mean Gaussian process with temporal correlation $E[\mathbf{g}_n^{(u)}{\mathbf{g}_{n-l}^{(u)}}^H] = J_0(2\pi f_m l T_s)\mathbf{I}_{N_R}$, so that $E[(\mathbf{h}_n^{(u)} - \mathbf{h}_{n-1}^{(u)})(\mathbf{h}_n^{(u)} - \mathbf{h}_{n-1}^{(u)})^H] = \sigma^2 E_s^{(u)}\mathbf{\Gamma}^{(u)}$, where $\sigma = \sqrt{2(1 - J_0(2\pi f_m T_s))}$ is the temporal correlation coefficient between consecutive sampling instants.

b) Multipath block fading model

In case the CD-NOMA spread complex symbols are transmitted over consecutive frequency-domain REs, a multipath block fading model is considered for the wireless channel, that is channel realizations are assumed independent each time a user has a new CD-NOMA block to transmit.

The complex baseband equivalent wide-band multipath channel impulse response for the u-th user over the N_R receive antennas has the form [79]

$$\mathbf{h}^{(u)}(\tau) = \sqrt{E_s^{(u)}} \sum_{l=0}^{L-1} \sqrt{p_l} \mathbf{\Gamma}^{(u)1/2} \mathbf{g}_l^{(u)} \delta(\tau - lT_s), \tag{3.3}$$

where p_l is the average power at the l-th delay and the $\mathbf{g}_l^{(u)}$'s are i.i.d. random vectors with distribution $\mathcal{CN}(0, \mathbf{I}_{N_R})$. Assuming a standard exponentially decaying power delay profile [79] is in order, then $p_l = Ae^{-lT_s/\bar{\sigma}_\tau}$, for $l = 0, \ldots, L-1$, where $\bar{\sigma}_\tau$ is the rms delay spread and A is a normalization constant such that $\sum_{l=0}^{L-1} p_l = 1$. Applying an element-wise N-point Discrete Fourier Transform (DFT) with zero-padding to the coefficients of $\mathbf{h}^{(u)}(\tau)$, we obtain the u-th user Channel Frequency Response (CFR) at the r-th receive antenna $H_{n,r}^{(u)}$, where n now denotes the index of the discrete frequency $n/(NT_s)$, $n = 0, \ldots, N-1$. Thus, the CFR in vector form $\mathbf{H}_n^{(u)} = [H_{n,1}^{(u)}, \ldots, H_{n,N_R}^{(u)}]^T$, can be written as

$$\mathbf{H}_n^{(u)} = \sqrt{E_s^{(u)}} \sum_{l=0}^{L-1} \sqrt{p_l} \mathbf{\Gamma}^{(u)1/2} \mathbf{g}_l^{(u)} e^{-j2\pi nl/N}, \tag{3.4}$$

so that $E[(\mathbf{H}_n^{(u)} - \mathbf{H}_{n-1}^{(u)})(\mathbf{H}_n^{(u)} - \mathbf{H}_{n-1}^{(u)}))^H] = \sigma^2 E_s^{(u)} \mathbf{\Gamma}^{(u)}$, where $\sigma = \sqrt{2\sum_{l=0}^{L-1} p_l(1 - \cos(2\pi l/N))}$ is now the frequency correlation coefficient between consecutive discrete frequencies.

3.1.4 Unified Markovian model for dynamic channels

We now seek simple a first order Markovian models to track the channel variations valid for all wireless propagation channels introduced in Sec. 3.1.3.

We let n denote the discrete time index (resp. discrete frequency index) for a flat (resp. multipath) fading channel and σ is the correlation coefficient in the corresponding domain. Let $x_{n,r}$ be the generic notation for a channel coefficient at the r-th receive antenna (either in the time-domain as in Eq. (3.2) or frequency-domain as in Eq. (3.4)) channel in Sec. 3.1.3. Let us propose unified models for the random vector $\mathbf{x}_n^{(u)} = [x_{n,1}^{(u)}, \ldots, x_{n,N_R}^{(u)}]^T$.

Vector channel model We first propose a vector Gaussian random walk explicitly modeling antenna correlation

$$\mathbf{x}_n^{(u)} = \mathbf{x}_{n-1}^{(u)} + \mathbf{\Delta}_n^{(u)}, \tag{3.5}$$

with i.i.d. process noise $\mathbf{\Delta}_n^{(u)} \sim \mathcal{CN}(\mathbf{\Delta}_n^{(u)}; \mathbf{0}_{N_R \times 1}, \zeta\sigma^2 E_s^{(u)} \mathbf{\Gamma}^{(u)})$ and ζ is a tuning parameter controlling the modeling error.

Scalar channel model Alternatively, the receiver can deliberately choose to ignore antenna correlation for the sake of complexity reduction in the implementation. This results in the following independent per-antenna Gaussian random walk models for $r = 1, \ldots, N_R$

$$x_{n,r}^{(u)} = x_{n-1,r}^{(u)} + \Delta_{n,r}^{(u)}, \tag{3.6}$$

with i.i.d. process noise $\Delta_{n,r}^{(u)} \sim \mathcal{CN}(\Delta_{n,r}^{(u)}; 0, \zeta\sigma^2 E_s^{(u)})$ and ζ is a tuning parameter controlling the modeling error. In vector form this can be written as (3.5), under the crude approximation $\mathbf{\Gamma}^{(u)} = \mathbf{I}_{N_R}$ by dropping all off-diagonal elements.

3.1.5 Receiver observation model

Let $\mathbf{x}_n^{(u)}$ in (3.5) be the u-th user multi-antenna channel over the n-th RE. Also, the u-th user has an associated binary user existence variable $\theta^{(u)} \in \{0, 1\}$, where $p_a^{(u)} = P(\theta^{(u)} = 1)$, will account for grant-free access over a block of N consecutive REs. Collecting the received discrete-time signal at RE n and over N_R receive antenna elements, we obtain the vector $\mathbf{y}_n = \begin{bmatrix} y_{n,1} & y_{n,2} & \dots & y_{n,N_R} \end{bmatrix}^T$ as

$$\mathbf{y}_n = \sum_{u=1}^{U} \theta^{(u)} d_n^{(u)} \mathbf{x}_n^{(u)} + \mathbf{w}_n, \qquad n = 0, .., N-1 \tag{3.7}$$

under the assumption that the N consecutive REs are orthogonal.

Moreover, $\mathbf{w}_n = \begin{bmatrix} w_{n,1} & w_{n,2} & \dots & w_{n,N_R} \end{bmatrix}^T$ is a white Gaussian noise vector with zero mean and covariance matrix parameterized by $\mathbf{R} = N_0 \mathbf{I}_{N_R}$.

If this Eq. (3.7) is written in scalar form over receive antenna r, then

$$y_{n,r} = \sum_{u=1}^{U} \theta^{(u)} d_n^{(u)} x_{n,r}^{(u)} + w_{n,r}, \tag{3.8}$$

for $1 \leq r \leq N_R$.

3.2 Bayesian framework and associated factor graph

Define $\mathbf{b} = [\mathbf{b}^{(1)}, \mathbf{b}^{(2)}, \dots, \mathbf{b}^{(U)}]$, $\mathbf{c} = [\mathbf{c}^{(1)}, \mathbf{c}^{(2)}, \dots, \mathbf{c}^{(U)}]$, and $\mathbf{d} = [\mathbf{d}^{(1)}, \mathbf{d}^{(2)}, \dots, \mathbf{d}^{(U)}]$ as the matrices of information bits, coded bits and modulated symbols. Similarly, define the vector of complex symbols transmitted by all users over the n RE as $\mathbf{d}_n = [d_n^{(1)}, \dots, d_n^{(U)}]^T$.

Assuming the users are both far apart from each other and with independent traffic, the random process for the u-th user channel over all REs $\{\mathbf{x}_0^{(u)}, \dots, \mathbf{x}_{N-1}^{(u)}\}$ and user activity variable $\theta^{(u)}$ are independent from all other hidden variables. Also, let $\mathbf{x} = \{\mathbf{x}_0^{(u)}, \dots, \mathbf{x}_{N-1}^{(u)}\}_{u=1}^{U}$ denote the hidden channel random processes corresponding to all users over all REs. Similarly, we let $\boldsymbol{\theta} = [\theta^{(1)}, \theta^{(1)}, \dots, \theta^{(U)}]^T$ (resp. $\mathbf{y} = [\mathbf{y}_0^T, \mathbf{y}_1^T, \dots, \mathbf{y}_{N-1}^T]^T$) denote the vector of activity variables for all users (resp. the vector of observations over all REs) during the current transmission block.

Estimation of $\mathbf{b}^{(u)}$ corresponds to DEC, of $\mathbf{d}^{(u)}$ to Demodulation (DEM), of $\mathbf{x}_n^{(u)}$ to CE, and of $\theta^{(u)}$ to UAD, respectively. In a Bayesian setting, these random vectors corresponds to hidden variables that the receiver wants to infer. For instance, finding the Maximum-aposteriori Estimator (MAP) estimator of $\theta^{(u)}$ corresponds to the Bayesian estimation problem

$$\hat{\theta}^{(u)} = \arg\max_{\theta^{(u)}} \; p(\theta^{(u)}|\mathbf{y}) = \arg\max_{\theta^{(u)}} \sum_{\sim\theta^{(u)}} \int p(\mathbf{b}, \mathbf{c}, \mathbf{d}, \mathbf{x}, \boldsymbol{\theta}|\mathbf{y}) d\mathbf{x}, \tag{3.9}$$

where we have employed marginalization of the joint a posteriori distribution. This method is crucial for simplifying the problem, which otherwise involves dealing with a high-dimensional probability distribution. The process of marginalization allows us to focus on a subset of the variables of interest, reducing the overall complexity of the problem.

However, even after marginalization, solving the resulting maximization involves navigating through a vast and intricate solution space, which makes the direct optimization approaches infeasible. To address this issue, we adopt a strategy based on factor graphs, a powerful tool in probabilistic graphical models.

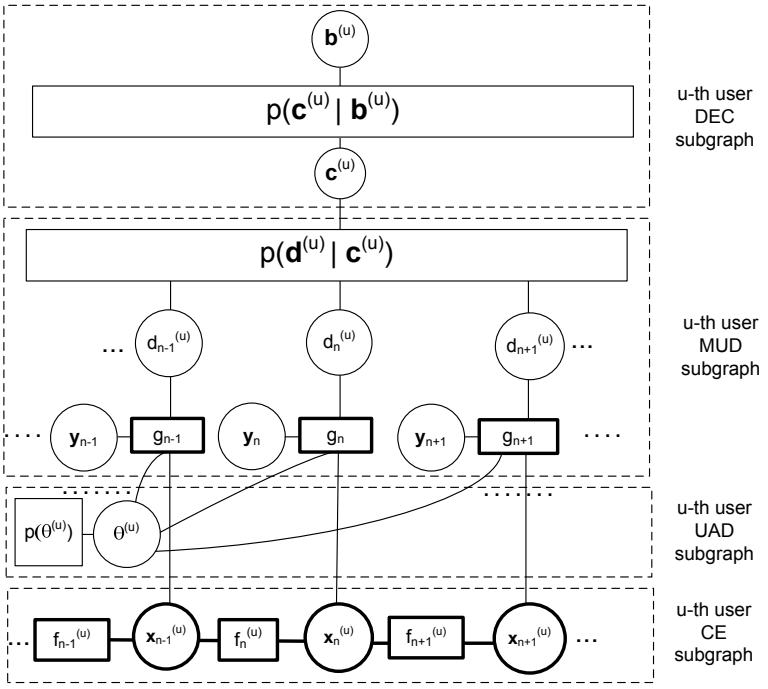

Figure 3.4: Fraction of the factor graph corresponding to the u-th user in the proposed grant-free NOMA system model for a vector channel model.

Factor graphs provide a structured way to represent and manipulate the joint posterior distribution by breaking it down into simpler components. This decomposition is achieved through factorization, which expresses the joint posterior as a product of several smaller and more manageable functions, known as factors. Each factor typically depends on a subset of the total variables, which significantly simplifies the overall analysis (see Sec. 2.2 for more details).

As mentioned earlier, the purpose of a factor graph is to lower the complexity of computing the marginal distribution of a hidden variable (marginal *a posteriori* distribution in this case) via distributed inference algorithms in the form of message passing, as recalled in Sec. 2.3.

3.2.1 Factor graph considering the vector channel model

Using the conditional independence assumptions in our transceiver model in Sec. 3.1 along with the vector channel model in Sec. 3.1.4, we obtain

$$p(\mathbf{b},\mathbf{c},\mathbf{d},\mathbf{x},\boldsymbol{\theta}|\mathbf{y}) \propto \left(\prod_{n=0}^{N-1} p(\mathbf{y}_n|\mathbf{d}_n,\mathbf{x}_n,\boldsymbol{\theta})\right)\prod_{u=1}^{U}\Big\{p(\theta^{(u)})p(\mathbf{x}_0^{(u)}) \times\left(\prod_{n'=1}^{N-1} p(\mathbf{x}_{n'}^{(u)}|\mathbf{x}_{n'-1}^{(u)})\right)p(\mathbf{d}^{(u)}|\mathbf{c}^{(u)})p(\mathbf{c}^{(u)}|\mathbf{b}^{(u)})\Big\}, \quad (3.10)$$

where $\mathbf{x}_n = \{\mathbf{x}_n^{(1)},\ldots,\mathbf{x}_n^{(U)}\}$.

In this factorization of the joint posterior distribution, the components are defined as follows:

1. The first factor, $p(\mathbf{y}_n|\mathbf{d}_n,\mathbf{x}_n,\boldsymbol{\theta})$, accounts for the likelihood of data symbols, channel vector, and user activity variable over the n-th RE.

2. $p(\theta^{(u)})$ represents the prior distribution of user activity variable for the u-th user.

3. $p(\mathbf{x}_0^{(u)})$ denotes the prior distribution of the u-th user channel vector over the initial RE.

4. $p(\mathbf{x}_{n'}^{(u)} \mid \mathbf{x}_{n'-1}^{(u)})$ describes the transition distribution of the u-th user channel vector from one RE to the next, capturing the correlation across different receive antennas.

5. $p(\mathbf{d}^{(u)}|\mathbf{c}^{(u)})$ represents the deterministic mapping induced by the chosen CD-NOMA scheme.

6. $p(\mathbf{c}^{(u)} \mid \mathbf{b}^{(u)})$ details the deterministic mapping induced by the chosen FEC scheme.

Introducing the shorthand notations

$$\begin{aligned} f_n^{(u)} &= p(\mathbf{x}_n^{(u)}|\mathbf{x}_{n-1}^{(u)}) \\ g_n &= p(\mathbf{y}_n|\mathbf{d}_n, \mathbf{x}_n, \boldsymbol{\theta}) \end{aligned} \tag{3.11}$$

leads to the portion of the factor graph associated with the u-th user depicted in Fig. 3.4, where it is implicit that there are $U-1$ similar subgraphs (corresponding to all other users) stacked in parallel.

3.2.2 Factor graph considering the scalar channel model

In fact, under some conditions further factorizations can be exploited. Indeed, noting that the coordinates of the noise vector affecting the observation model in Eq. (3.7) are independent, we have

$$p(\mathbf{y}_n|\mathbf{d}_n, \mathbf{x}_n, \boldsymbol{\theta}) = \prod_{r=1}^{N_R} p\left(y_{n,r}|\{\mathbf{d}_n, \{x_{n,r}\}_{u=1}^{U}\right). \tag{3.12}$$

Now, under the scalar channel model ignoring antenna correlation in Sec. 3.1.4, we also get

$$p(\mathbf{x}_{n'}^{(u)}|\mathbf{x}_{n'-1}^{(u)}) = \prod_{r=1}^{N_R} p(x_{n',r}^{(u)}|x_{n'-1,r}^{(u)}). \tag{3.13}$$

Introducing the shorthand notations

$$\begin{aligned} f_{n,r} &= p(x_{n,r}^{(u)}|x_{n-1,r}^{(u)}) \\ g_{n,r} &= p\left(y_{n,r}|\{\mathbf{d}_n, \{x_{n,r}\}_{u=1}^{U}\right). \end{aligned} \tag{3.14}$$

leads to the portion of the factor graph associated with the u-th user depicted in Fig. 3.5, where it is implicit that there are $U-1$ similar subgraphs (corresponding to all other users) stacked in parallel.

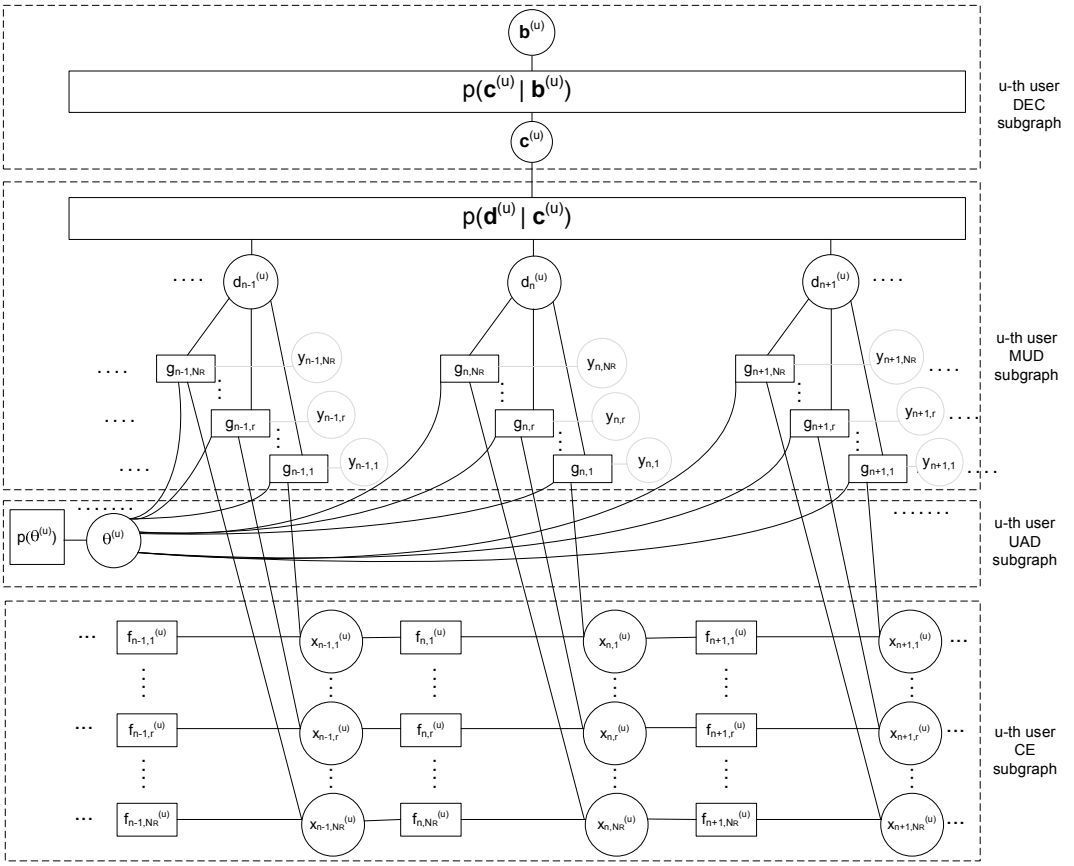

Figure 3.5: Fraction of the factor graph corresponding to the u-th user in the proposed grant-free NOMA system model for a scalar channel model.

3.3 Running example: OFDM-IDMA transceiver

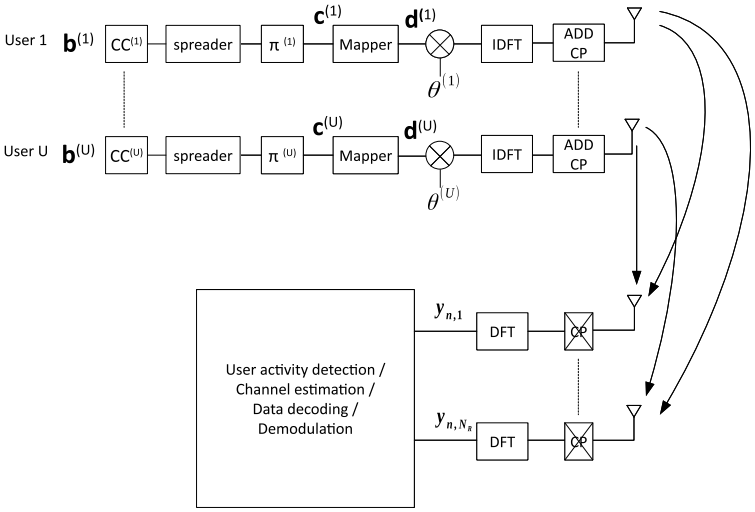

Figure 3.6: OFDM-IDMA-SIMO system model.

Let us introduce the particular physical layer, that will be used over and over as an illustrative example for the sake of obtaining simulation results in a same context in the next chapters. We entirely specify the transceiver when OFDM-IDMA [80] is the CD-NOMA scheme of interest, represented in Fig. 3.6. We wish to emphasize that generic notations for any CD-NOMA-based grant-free access system (resp. for the wireless channel model) under consideration have been defined in Sec. 3.1 (resp. in Sec. 3.1.3). Besides being the simplest CD-NOMA scheme, other reasons for selecting IDMA is that it is easily combined with off-the-shelf building blocks such as channel coding, standard modulation formats (see Sec. 3.3.1), OFDM, multi-antenna reception (see Sec. 3.3.2) [80, 81]. While message-passing (see chapter 2) has been used in [82] to address most of the challenges listed in Sec. 1.4, the problem of UAD for OFDM-IDMA has been mostly overlooked (with the notable exception of [83]).

3.3.1 Interleaved Division Multiple Access (IDMA)

We first briefly recall the essential building blocks of IDMA [68].

The vector of information bits denoted by $\mathbf{b}^{(u)}$ for the u-th user undergoes FEC encoding with a recursive systematic convolution encoder $CC^{(u)}$, then spreading with a repetition encoder. These encoded bits of each user are then passed through the user-specific interleavers $\pi^{(u)}$. The resulting bits are denoted by the vector $\mathbf{c}^{(u)} = \mathcal{C}^{(u)}(\mathbf{b}^{(u)})$, where the one-to-one function $\mathcal{C}^{(u)}(.)$ denotes the combined effect of encoding and interleaving. Then, modulation uses a Q-ary mapping function χ to generate the n-th complex symbol $d_n^{(u)} = \chi(\mathbf{c}_n^{(u)})$, where $\mathbf{c}_n^{(u)} = \begin{bmatrix} c_{n,0}^{(u)} & c_{n,1}^{(u)} & \dots & c_{n,Q-1}^{(u)} \end{bmatrix}$ is the corresponding vector of binary labels. Without loss of generality, it is further assumed that the modulation constellation is normalized to unit energy, i.e. $E[|d_n^{(u)}|^2] = 1$.

Examples of modulation used in this thesis are Binary Phase Shift Keying (BPSK), QPSK and Quadrature amplitude modulation (QAM).

3.3.2 Orthogonal frequency division multiplexing (OFDM)

Orthogonal Frequency Division Multiplexing (OFDM) [78] is a multi-carrier modulation technique that is widely used in modern digital communication systems such as Wi-Fi, 4G, 5G and digital broadcasting. OFDM divides a high-speed data stream into multiple substreams, each transmitted over a separate subcarrier frequencies, by applying an inverse discrete Fourier transform (IDFT) to the complex modulated symbols.

Importantly, we will consider that the following assumptions are satisfied, so that the subcarriers can be considered as orthogonal REs:

- at the transmitters, a sufficiently long Cyclic Prefix (CP) is inserted in front of each OFDM block to absorb the joint effect of Inter-block Interference (IBI) and user asynchronism
- users' frequency selective channels are block fading (i.e. approximately constant over the duration of each OFDM block) and the frequency offsets between users and the BS are small enough to avoid Inter-carrier Interference (ICI).

The main reasons for considering OFDM is that it essentially converts a per-user multi-antenna frequency selective block fading channel of the form (3.3) to simpler per-subcarrier dynamic CFR of the form (3.4) after suppressing the CP and applying a discrete Fourier transform (DFT).

3.3.3 Factor graph representation for grant-free OFDM-IDMA

In case OFDM-IDMA is selected as the CD-NOMA scheme, the factor graph in Fig. 3.4 can be particularized to Fig. 3.7, where all the details of the retained CD-NOMA spreader in Sec. 3.3.1 have been taken into account.

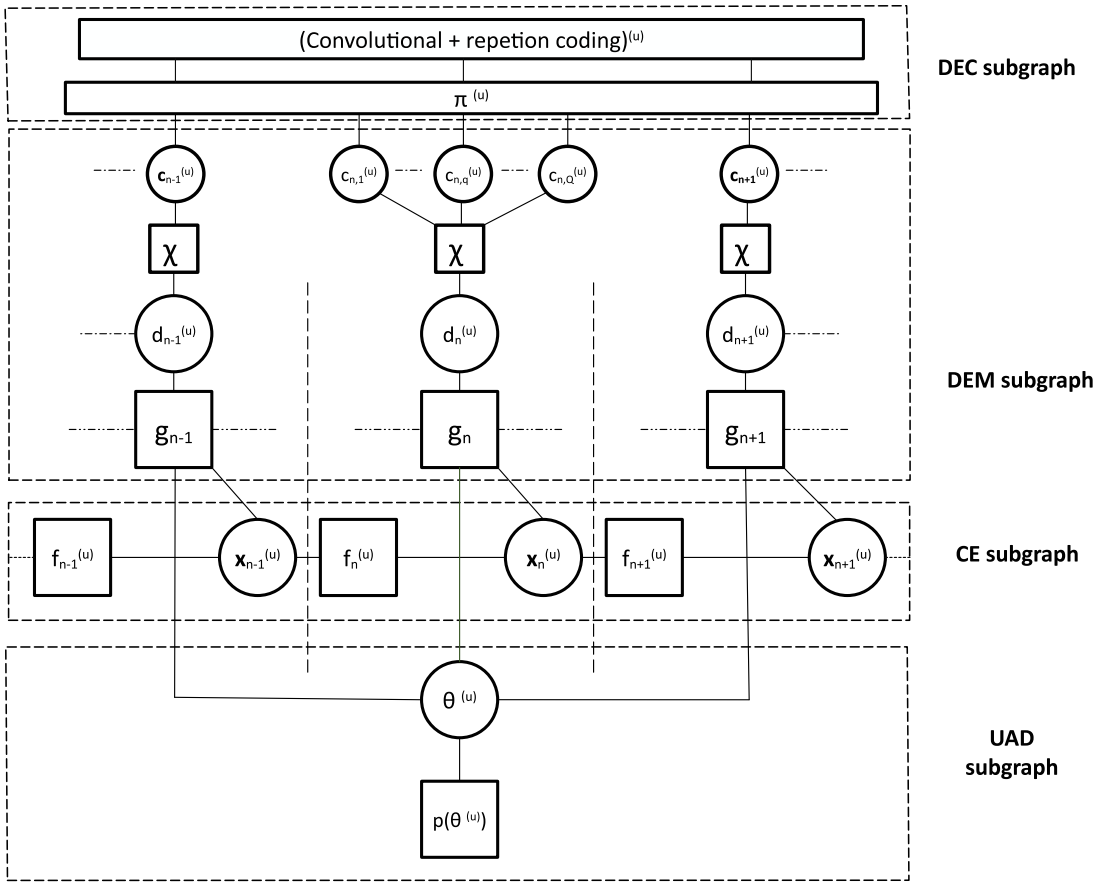

Figure 3.7: Fraction of the factor graph corresponding to the u-th user hidden variables in grant-free coded OFDM-IDMA.

While the receiver introduced in subsequent chapters are designed for the general factor graph in Fig. 3.4, Fig. 3.7 will serve as the particular support of message passing in the corresponding simulation results sections.

Chapter 4

Hybrid EP-BP grant-free receiver exploiting antenna correlation

This chapter introduces a novel grant-free receiver design based on Gaussian message-passing applicable over the factor graph in Fig. 3.4 (except over the decoding subgraph, where this is irrelevant) suitable for the vector channel model in Eq. (3.5) and the vector observation model in Eq. (3.7). This design is motivated by the need for reduced-complexity inference obtained by projecting the messages incoming and outcoming from all hidden variables to the family of circularly symmetric Gaussian distributions. This circumvents the need for cumbersome multiple summations and integrations when explaining away discrete-valued hidden variables such as the UAD or modulated symbol variables.

While EP can be used for that purpose over the portions of the factor graph corresponding to channel estimation, multi-user detection and decoding, this grossly fails over the user activity detection subgraph. This justifies the adoption of an hybrid message-passing scheme, where BP is used for the sake of decoding and user activity detection.

Unless otherwise specified, in the sequel we will refer to messages exchanged over the graphical model for grant-free NOMA access in Fig. 3.4.

4.1 Related work

Grant-free access introduces several challenges. Since the receiver is unaware of which users are transmitting, UAD must be performed in addition to CE, MUD, DEM, and DEC. Both separate and joint UAD and MUD methods have shown satisfactory performance, even in massive grant-free access systems [1], [2], [3], [4], [5], [6].

Many approaches rely on AI and/or CS, which need large datasets for training, making them impractical in some cases. They also assume Channel State Information (CSI) is known, which introduces the additional challenge of CE. Preamble-based CE methods [7], [8], [9], [10], [11] increase signaling overhead, making them less feasible. Limited or no preamble methods [12], [13] have been developed, primarily for massive MIMO.

NOMA's ability to superimpose signals over the same resources with controlled interference, rather than using RA techniques [14], [15], [16], is an attractive approach for grant-free access, improving resource use and system capacity.

We aim to tackle the challenge of solving the multiple access problem on the receiver side in grant-free NOMA systems using message-passing algorithms. These algorithms are well-suited due to their adaptability in performing inference within probabilistic graphical models,

as outlined in [17], which provides a unified perspective on techniques such as BP, EP, Mean Field (MF), and Variational Bayesian (VB) inference.

Earlier research explored separate CS-based UAD with discrete BP for MUD under perfect CSI [18]. Furthermore, joint UAD and CE have been examined using either hybrid BP/MF [19] or hybrid BP/AMP approaches [20], before performing separate MUD.

Attempts at joint UAD, CE, MUD, and DEC can be categorized based on the underlying models for user activity and transmitted symbols. First, BP-EP-VB methods, such as in [25], treat user activity as continuous-valued precision parameters, while symbols are discrete-valued. Second, hybrid BP-EP approaches [22] model user activity with binary variables, considering symbols as continuous-valued. Finally, works like BP [21] and auxiliary variable hybrid BP-EP-MF [23, 24] treat both user activity and symbols as discrete-valued.

A key challenge in these models arises from the mixed discrete-continuous probabilistic framework. This combination leads to some messages being treated as p.m.f.s and others as p.d.f.s, complicating the marginalization over user activity variables. In [21] and [24], this involves tedious discrete summations, while [22] requires an additional Expectation Maximization (EM) procedure to integrate the stages of UAD, CE, and MUD.

We address this challenge by proposing a unified approach where all messages, even those related to user activity variables, are modeled as Gaussian p.d.f.s. This shift simplifies marginalization, reducing it to straightforward integration, thus avoiding computationally expensive mixed summation-integral operations and maintaining consistency within the message-passing framework.

4.2 Main contributions

The main contributions in this chapter are summarized as follows:

1. A message-passing receiver design based on propagating Gaussian messages, even for discrete-valued hidden variables, is introduced for the sake of complexity-reduction
2. A principled projection operator is developed to approximate continuous mixtures of Gaussians as a single Gaussian distribution for the sake of EP message-passing implementing DEM and CE
3. BP message-passing with Gaussian approximation is developed for the sake of UAD
4. Antenna correlation is explicitly taken into account, thus avoiding suboptimal CE
5. A pilot-only hyperparameter estimator of users' receive energy and inter-antenna correlation is devised, as a step forward to fully grant-free transmission avoiding handshake protocols between users and the BS.

4.3 Novel projection operator

In the observation constraint node g_n in Eq. (3.11), the hidden variables that need to be estimated are the data symbols $d_n^{(u)}$, the channels $\mathbf{x}_n^{(u)}$, and the user activity variables $\theta^{(u)}$. These variables represent different aspects of the system: $d_n^{(u)}$ corresponds to the transmitted data symbols for the u-th user on RE n, $\mathbf{x}_n^{(u)}$ represents the channel of the u-th user for the same RE, and $\theta^{(u)}$ denotes the binary user activity variables indicating whether a user is active or inactive.

Let $\boldsymbol{\pi}$ be a parameter vector containing one of the aforementioned hidden variables, with prior distribution $p(\boldsymbol{\pi})$. As we shall see in the next sections, EP message-passing will involve conditional Gaussian continuous mixtures of the form [85, p. 239]

$$p(\mathbf{z}|\boldsymbol{\pi}) = \int_{\boldsymbol{\theta}} \boldsymbol{\omega}(\boldsymbol{\theta}) \mathcal{CN}(\mathbf{z}; \mathbf{m}(\boldsymbol{\theta}|\boldsymbol{\pi}), \boldsymbol{\Sigma}(\boldsymbol{\theta}|\boldsymbol{\pi})) d\boldsymbol{\theta}, \tag{4.1}$$

where $\mathbf{z} \in \mathbb{C}^d$ (d is a strictly positive integer) and $\int_{\boldsymbol{\theta}} \boldsymbol{\omega}(\boldsymbol{\theta}) d\boldsymbol{\theta} = 1$.

The set of densities Φ over which EP will repeatedly project messages will be the family of circularly symmetric Gaussian densities $\mathcal{G}$. Thus, a projection operator over Φ, $\text{proj}_{\Phi}(.)$ capable of projecting Eq. (4.1) to a target Gaussian density of the form $q(\mathbf{z}|\boldsymbol{\pi}) = \mathcal{CN}(\mathbf{z}; \mathbf{m}(\boldsymbol{\pi}), \boldsymbol{\Sigma})$ is needed. Using the common projection operator minimizing the KL divergence $KL(p||q)$ [74] is unsuited for that purpose, as the covariance of the target distribution would depend on $\boldsymbol{\pi}$. However, as emphased in [74], other divergence measures such as the α-divergence can be used, as long as they are convex in both p and q.

To solve this issue, we define a new projection operator that will be used in the sequel for any continuous mixture of Gaussian distributions

$$\text{proj}_{\Phi}(p) = \arg\min_{q \in \Phi} E_{p(\boldsymbol{\pi})}[KL(p||q)], \tag{4.2}$$

which is the expected KL divergence with respect to the prior of $\boldsymbol{\pi}$. The mean and covariance of the target Gaussian density are then found in closed form as

$$\begin{aligned} \mathbf{m}(\boldsymbol{\pi}) &= \int_{\boldsymbol{\theta}} \boldsymbol{\omega}(\boldsymbol{\theta}) \mathbf{m}(\boldsymbol{\theta}|\boldsymbol{\pi}) d\boldsymbol{\theta} \\ \boldsymbol{\Sigma} &= \int_{\boldsymbol{\pi}} \left[\int_{\boldsymbol{\theta}} \boldsymbol{\omega}(\boldsymbol{\theta}) \big((\mathbf{m}(\boldsymbol{\pi}) - \mathbf{m}(\boldsymbol{\theta}|\boldsymbol{\pi}))(\mathbf{m}(\boldsymbol{\pi}) - \mathbf{m}(\boldsymbol{\theta}|\boldsymbol{\pi}))^H + \boldsymbol{\Sigma}(\boldsymbol{\theta}|\boldsymbol{\pi}) \big) d\boldsymbol{\theta} \right] \cdot p(\boldsymbol{\pi}) d\boldsymbol{\pi}. \end{aligned} \tag{4.3}$$

The proof is postponed to Appendix A.

4.4 Demodulation

The segment of the factor graph that corresponds to symbol estimation, often referred to as demodulation, involves the computation of specific messages. This part of the graph primarily handles the interaction between the data symbols and the observation constraint nodes, facilitating the accurate detection of transmitted symbols. Distributed inference of the n-th complex symbol for the u-th user is characterized by the exchange of messages, specifically $\mu_{d_n^{(u)} \to g_n}(d_n^{(u)})$ and $\mu_{g_n \to d_n^{(u)}}(d_n^{(u)})$. The message $\mu_{d_n^{(u)} \to g_n}(d_n^{(u)})$ represents the information passed from the data symbols $d_n^{(u)}$ to the observation constraint node g_n in Eq. (3.11), conveying the probabilistic belief about the value of the data symbol. Conversely, the message $\mu_{g_n \to d_n^{(u)}}(d_n^{(u)})$ represents the information transmitted from the observation constraint node g_n back to the data symbols $d_n^{(u)}$, updating the estimate based on the observed data and the likelihood of various symbol values.

These message exchanges are critical for iteratively refining the estimates of the data symbols, as each message contributes to the overall likelihood computation and symbol detection process. This iterative message-passing approach is fundamental to achieving robust demodulation, especially in complex scenarios with multiple users generating interference.

Message from g_n to $d_n^{(u)}$

Message from the receiver to the demodulator is computed using the projection operator defined earlier.

Applying the EP factor node rule Eq. (2.4) to g_n, the message $\mu_{g_n \to d_n^{(u)}}$ is computed as,

$$\mu_{g_n \to d_n^{(u)}}(d_n^{(u)}) = \frac{\text{proj}_{\Phi}\Big(k \cdot \mu_{d_n^{(u)} \to g_n}(d_n^u)\tilde{p}(\mathbf{y}_n|d_n^{(u)})\Big)}{\mu_{d_n^{(u)} \to g_n}(d_n^{(u)})}, \tag{4.4}$$

where $\tilde{p}(\mathbf{y}_n|d_n^{(u)})$ is given by the following equation,

$$\tilde{p}(\mathbf{y}_n|d_n^{(u)}) = \int p(\mathbf{y}_n|\mathbf{x}_n, \mathbf{d}_n, \boldsymbol{\theta}) \prod_{u' \neq u} \mu_{d_n^{(u')} \to g_n}(d_n^{(u')}) \prod_{u'=1}^{U} \mu_{\mathbf{x}_n^{(u')} \to g_n}(\mathbf{x}_n^{(u')}) \prod_{u'=1}^{U} \mu_{\theta^{(u')} \to g_n}(\theta^{(u')}) d \sim d_n^{(u)}, \tag{4.5}$$

and $d \sim d_n^{(u)}$ shows that the integration is performed with respect to all variables of the function $p(\mathbf{y}_n|\mathbf{x}_n, \mathbf{d}_n, \boldsymbol{\theta})$ except the variable $d_n^{(u)}$, so the Eq. (4.5) can also be written as,

$$\begin{aligned}\tilde{p}(\mathbf{y}_n|d_n^{(u)}) = \int\int\int p(\mathbf{y}_n|\mathbf{x}_n, \mathbf{d}_n, \boldsymbol{\theta}) \prod_{u' \neq u} \mu_{d_n^{(u')} \to g_n}(d_n^{(u')}) \prod_{u'=1}^{U} \mu_{\mathbf{x}_n^{(u')} \to g_n}(\mathbf{x}_n^{(u')}) \\ \times \prod_{u'=1}^{U} \mu_{\theta^{(u')} \to g_n}(\theta^{(u')}) \prod_{u' \neq u} dd_n^{(u')} \prod_{u'=1}^{U} d\mathbf{x}_n^{(u')} \prod_{u'=1}^{U} d\theta^{(u')}.\end{aligned} \tag{4.6}$$

For the sake of complexity reduction, as explained in Sec. 4.2, all the messages $\mu_{d_n^{u'} \to g_n}(d_n^{(u')})$, $\mu_{\mathbf{x}_n^{(u')} \to g_n}(d_n^{(u')})$, and $\mu_{\theta^{(u')} \to g_n}(\theta^{(u')})$ are to be projected to Gaussian distributions in the form,

$$\begin{aligned}\mu_{\mathbf{x}_n^{(u)} \to g_n}(\mathbf{x}_n^{(u)}) &= \mathcal{CN}(\mathbf{x}_n^{(u)}; \mathbf{m}_{\mathbf{x}_n^{(u)} \to g_n}, \Sigma_{\mathbf{x}_n^{(u)} \to g_n}) \\ \mu_{d_n^{(u)} \to g_n}(d_n^{(u)}) &= \mathcal{CN}(d_n^{(u)}; m_{d_n^{(u)} \to g_n}, \sigma^2_{d_n^{(u)} \to g_n}) \\ \mu_{\theta^{(u)} \to g_n}(\theta^{(u)}) &= \mathcal{CN}(\theta^{(u)}; m_{\theta^{(u)} \to g_n}, \sigma^2_{\theta^{(u)} \to g_n}).\end{aligned} \tag{4.7}$$

It follows that Eq. (4.6) can be written as a function of those messages,

$$\begin{aligned}\tilde{p}(\mathbf{y}_n|d_n^{(u)}) = \int\int\int \mathcal{CN}(\mathbf{y}_n; \sum_{u=1}^{U} \theta^{(u)} d_n^{(u)} \mathbf{x}_n^{(u)}, \mathbf{R}) \prod_{u' \neq u} \mathcal{CN}(d_n^{(u')}; m_{d_n^{(u')} \to g_n}, \sigma^2_{d_n^{(u')} \to g_n}) \\ \times \prod_{u'=1}^{U} \mathcal{CN}(\mathbf{x}_n^{(u')}; \mathbf{m}_{\mathbf{x}_n^{(u')} \to g_n}, \Sigma_{\mathbf{x}_n^{(u')} \to g_n}) \prod_{u'=1}^{U} \mathcal{CN}(\theta^{(u')}; m_{\theta^{(u')} \to g_n}, \sigma^2_{\theta^{(u')} \to g_n}) \\ \times \prod_{u' \neq u} dd_n^{(u')} \prod_{u'=1}^{U} d\mathbf{x}_n^{(u')} \prod_{u'=1}^{U} d\theta^{(u')}.\end{aligned} \tag{4.8}$$

By looking at the integrand, it can be observed that it is a Gaussian mixture of the argument $\mathbf{y}_n$, where the complex normal distributions $\mathcal{CN}(d_n^{(u')}; m_{d_n^{(u')} \to g_n}, \sigma^2_{d_n^{(u')} \to g_n})$, $\mathcal{CN}(\mathbf{x}_n^{(u')}; \mathbf{m}_{\mathbf{x}_n^{(u')} \to g_n}, \Sigma_{\mathbf{x}_n^{(u')} \to g_n})$, and $\mathcal{CN}(\theta^{(u')}; m_{\theta^{(u')} \to g_n}, \sigma^2_{\theta^{(u')} \to g_n})$ act as the mixture components.

Applying the projection operator in Eq. (4.2) letting $\mathbf{z} = \mathbf{y}_n$, $\boldsymbol{\pi} = d_n^{(u)}$ and $\boldsymbol{\theta} = [\{d_n^{(u')}\}_{u' \neq u}, \{\theta^{(u')}, \mathbf{x}_n^{(u')}\}_{u'=1}^{U}]$, the continuous Gaussian mixture Eq. (4.8) becomes a Gaussian distribution in $\mathbf{y}_n$ given $d_n^{(u)}$ of the form,

$$\tilde{p}(\mathbf{y}_n|d_n^{(u)}) = \mathcal{CN}(\mathbf{y}_n; \mathbf{m}_{\mathbf{y}_n|d_n^{(u)}}(d_n^{(u)}), \boldsymbol{\Sigma}_{\mathbf{y}_n|d_n^{(u)}}), \tag{4.9}$$

where the mean vector $\mathbf{m}_{\mathbf{y}_n|d_n^{(u)}}(d_n^{(u)})$ is given by,

$$\begin{aligned}\mathbf{m}_{\mathbf{y}_n|d_n^{(u)}}(d_n^{(u)}) &= \mathbf{h}_{d_n^{(u)}\to g_n} d_n^{(u)} + \mathbf{l}_{d_n^{(u)}\to g_n}, \quad \mathbf{h}_{d_n^{(u)}\to g_n} = m_{\theta^{(u)}\to g_n}\mathbf{m}_{\mathbf{x}_n^{(u)}\to g_n}, \\ \mathbf{l}_{d_n^{(u)}\to g_n} &= \sum_{u'\neq u} m_{\theta^{(u')}\to g_n} m_{d_n^{(u')}\to g_n} \mathbf{m}_{\mathbf{x}_n^{(u')}\to g_n},\end{aligned} \tag{4.10}$$

where the first term is the average u-th user's useful signal conditional on $d_n^{(u)}$, while the second term is the average multi-user interference affecting the u-th user.

Also the covariance matrix $\mathbf{\Sigma}_{\mathbf{y}_n|d_n^{(u)}}$ is given by

$$\begin{aligned}\mathbf{\Sigma}_{\mathbf{y}_n|d_n^{(u)}} &= \left[|m_{\theta^{(u)}\to g_n}|^2 \mathbf{\Sigma}_{\mathbf{x}_n^{(u)}\to g_n} + \sigma^2_{\theta^{(u)}\to g_n}(\mathbf{m}_{\mathbf{x}_n^{(u)}\to g_n}\mathbf{m}^H_{\mathbf{x}_n^{(u)}\to g_n} + \mathbf{\Sigma}_{\mathbf{x}_n^{(u)}\to g_n})\right] \\ &+ \sum_{u'\neq u} \sigma^2_{d_n^{(u')}\to g_n}(|m_{\theta^{(u')}\to g_n}|^2 + \sigma^2_{\theta^{(u')}\to g_n})(\mathbf{m}_{\mathbf{x}_n^{(u')}\to g_n}\mathbf{m}^H_{\mathbf{x}_n^{(u')}\to g_n} + \mathbf{\Sigma}_{\mathbf{x}_n^{(u')}\to g_n}) \\ &+ \sum_{u'\neq u} |m_{d_n^{(u')}\to g_n}|^2 \left[|m_{\theta^{(u')}\to g_n}|^2 \mathbf{\Sigma}_{\mathbf{x}_n^{(u')}\to g_n} + \sigma^2_{\theta^{(u')}\to g_n}(\mathbf{m}_{\mathbf{x}_n^{(u')}\to g_n}\mathbf{m}^H_{\mathbf{x}_n^{(u')}\to g_n} + \mathbf{\Sigma}_{\mathbf{x}_n^{(u')}\to g_n})\right] + \mathbf{R},\end{aligned} \tag{4.11}$$

where the first term accounts for the residual uncertainty of the u-th user's useful signal, while the other terms account for the residual interference plus noise affecting the u-th user.

The behavior of this covariance matrix $\mathbf{\Sigma}_{\mathbf{y}_n|d_n^{(u)}}$ as the message-passing converges can be analyzed as follows:

- The first term would become close to zero because $\mathbf{\Sigma}_{\mathbf{x}_n^{(u)}\to g_n}$ and $\sigma^2_{\theta^{(u)}\to g_n}$ would become close to zero
- The second term would also become close to zero because for active users $\sigma^2_{d_n^{(u')}\to g_n}$ would become close to zeros and for inactive users $|m_{\theta^{(u')}\to g_n}|^2$ and $\sigma^2_{\theta^{(u')}\to g_n}$ would become close to zeros
- The third term would also become close to zero because for active users $\mathbf{\Sigma}_{\mathbf{x}_n^{(u')}\to g_n}$ and $\sigma^2_{\theta^{(u')}\to g_n}$ become close to zero and for the inactive users $|m_{\theta^{(u')}\to g_n}|^2$ and $\sigma^2_{\theta^{(u')}\to g_n}$ become close to zeros.

so that the final covariance matrix should be close to the noise covariance **R**.

The numerator of the Eq. (4.4) after using the expression of $\tilde{p}(\mathbf{y}_n|d_n^{(u)})$ and using the expression of product of two Gaussian distributions from [87] has the following mean and variance,

$$\begin{aligned}\tilde{m}_{d_n^{(u)}|\mathbf{y}_n} &= m_{d_n^{(u)}\to g_n} \\ &+ \sigma^2_{d_n^{(u)}\to g_n}\mathbf{h}^H_{d_n^{(u)}\to g_n}(\sigma^2_{d_n^{(u)}\to g_n}\mathbf{h}_{d_n^{(u)}\to g_n}\mathbf{h}^H_{d_n^{(u)}\to g_n} + \mathbf{\Sigma}_{\mathbf{y}_n|d_n^{(u)}})^{-1}(\mathbf{y}_n - \mathbf{l}_{d_n^{(u)}\to g_n} - \mathbf{h}_{d_n^{(u)}\to g_n} m_{d_n^{(u)}\to g_n}),\end{aligned} \tag{4.12}$$

$$\tilde{\sigma}^2_{d_n^{(u)}|\mathbf{y}_n} = \sigma^2_{d_n^{(u)}\to g_n} - \sigma^4_{d_n^{(u)}\to g_n}\mathbf{h}^H_{d_n^{(u)}\to g_n}(\sigma^2_{d_n^{(u)}\to g_n}\mathbf{h}_{d_n^{(u)}\to g_n}\mathbf{h}^H_{d_n^{(u)}\to g_n} + \mathbf{\Sigma}_{\mathbf{y}_n|d_n^{(u)}})^{-1}\mathbf{h}_{d_n^{(u)}\to g_n}. \tag{4.13}$$

The mean and variance of the message $\mu_{g_n\to d_n^{(u)}}(d_n^{(u)})$ can now be computed using the relationship of Gaussian division given by the following equation,

$$\frac{\mathcal{N}(\mathbf{x}|\mathbf{m}_1,\mathbf{\Sigma}_1)}{\mathcal{N}(\mathbf{x}|\mathbf{m}_2,\mathbf{\Sigma}_2)} \propto \mathcal{N}(\mathbf{x}|\mathbf{m}_3,\mathbf{\Sigma}_3), \tag{4.14}$$

where the mean $\mathbf{m}_3$ and covariance matrix $\mathbf{\Sigma}_3$ are given by the following equation,

$$\mathbf{\Sigma}_3^{-1} = \mathbf{\Sigma}_1^{-1} - \mathbf{\Sigma}_2^{-1}, \qquad \mathbf{m}_3 = \mathbf{\Sigma}_3(\mathbf{\Sigma}_1^{-1}\mathbf{m}_1 - \mathbf{\Sigma}_2^{-1}\mathbf{m}_2), \tag{4.15}$$

so the message $\mu_{g_n \to d_n^{(u)}}(d_n^{(u)})$ is given by the following equation,

$$\mu_{g_n \to d_n^{(u)}}(d_n^{(u)}) = \frac{\mathcal{CN}(d_n^{(u)}; \tilde{m}_{d_n^{(u)}|\mathbf{y}_n}, \tilde{\sigma}^2_{d_n^{(u)}|\mathbf{y}_n})}{\mathcal{CN}(d_n^{(u)}; m_{d_n^{(u)} \to g_n}, \sigma^2_{d_n^{(u)} \to g_n})} \propto \mathcal{CN}(d_n^{(u)}; m_{g_n \to d_n^{(u)}}, \sigma^2_{g_n \to d_n^{(u)}}), \tag{4.16}$$

whose the mean and variance are also obtained via Eq. (4.15) as,

$$\begin{aligned} \sigma^{-2}_{g_n \to d_n} &= \tilde{\sigma}^{-2}_{d_n^{(u)}|\mathbf{y}_n} - \sigma^{-2}_{d_n^{(u)} \to g_n} \\ m_{g_n \to d_n^{(u)}} &= \sigma^2_{g_n \to d_n}(\tilde{\sigma}^{-2}_{d_n^{(u)}|\mathbf{y}_n} \tilde{m}_{d_n^{(u)}|\mathbf{y}_n} - \sigma^{-2}_{d_n^{(u)} \to g_n} m_{d_n^{(u)} \to g_n}). \end{aligned} \tag{4.17}$$

After substituting the expressions of mean $\tilde{m}_{d_n^{(u)}|\mathbf{y}_n}$ and variance $\tilde{\sigma}^2_{d_n^{(u)}|\mathbf{y}_n}$ from the Eqs. (4.12) and (4.13), we have the following equations from the mean and variance from g_n to $d_n^{(u)}$,

$$\begin{aligned} m_{g_n \to d_n^{(u)}} &= \frac{\mathbf{h}^H_{d_n^{(u)} \to g_n}\left(\sigma^2_{d_n^{(u)} \to g_n}\mathbf{h}_{d_n^{(u)} \to g_n}\mathbf{h}^H_{d_n^{(u)} \to g_n} + \mathbf{\Sigma}_{\mathbf{y}_n|d_n^{(u)}}\right)^{-1}\left(\mathbf{y}_n - \mathbf{l}_{d_n^{(u)} \to g_n}\right)}{\mathbf{h}^H_{d_n^{(u)} \to g_n}\left(\sigma^2_{d_n^{(u)} \to g_n}\mathbf{h}_{d_n^{(u)} \to g_n}\mathbf{h}^H_{d_n^{(u)} \to g_n} + \mathbf{\Sigma}_{\mathbf{y}_n|d_n^{(u)}}\right)^{-1}\mathbf{h}_{d_n^{(u)} \to g_n}} \\ \sigma^2_{g_n \to d_n^{(u)}} &= \frac{1}{\mathbf{h}^H_{d_n^{(u)} \to g_n}\left(\sigma^2_{d_n^{(u)} \to g_n}\mathbf{h}_{d_n^{(u)} \to g_n}\mathbf{h}^H_{d_n^{(u)} \to g_n} + \mathbf{\Sigma}_{\mathbf{y}_n|d_n^{(u)}}\right)^{-1}\mathbf{h}_{d_n^{(u)} \to g_n}} - \sigma^2_{d_n^{(u)} \to g_n}. \end{aligned} \tag{4.18}$$

Message from $d_n^{(u)}$ towards g_n

The message $\mu_{d_n^{(u)} \to g_n}(d_n^{(u)})$ is same as the message $\mu_{\chi_n^{(u)} \to d_n^{(u)}}(d_n^{(u)})$ - message from modulation constraint towards data symbol. Applying the EP factor node rule Eq. (2.4) to χ_n,

$$\begin{aligned} \mu_{d_n^{(u)} \to g_n}(d_n^{(u)}) &= \mu_{\chi_n^{(u)} \to d_n^{(u)}}(d_n^{(u)}) \\ &= \frac{\text{proj}_\Phi\left(k.\mu_{d_n^{(u)} \to \chi_n^{(u)}}(d_n^{(u)}) \sum_{\mathbf{c}_n^{(u)}} \delta(d_n^{(u)} - \chi_n^{(u)}[\mathbf{c}_n^{(u)}]) \prod_{q'=0}^{Q-1} \mu_{c_{n,q}^{(u)} \to \chi_n^{(u)}}(c_{n,q}^{(u)})\right)}{\mu_{g_n \to d_n^{(u)}}(d_n^{(u)})} \\ &= \frac{\text{proj}_\Phi\left(k.\mu_{d_n^{(u)} \to \chi_n^{(u)}}(d_n^{(u)}) \prod_{q'=0}^{Q-1} \mu_{c_{n,q}^{(u)} \to \chi_n^{(u)}}(c_{n,q}^{(u)})|_{c_{n,q}^{(u)} = \chi_q^{-1}[d_n^{(u)}]}\right)}{\mu_{g_n \to d_n^{(u)}}(d_n^{(u)})}. \end{aligned} \tag{4.19}$$

where $\chi_q^{-1}[d_n^{(u)}]$ denotes the q-th binary label associated to the complex symbol $d_n^{(u)}$ for $q \in \{0, \ldots, Q-1\}$.

The argument of the projection operator is the a posteriori probability mass function (pmf) of the data symbols $d_n^{(u)}$ expressed as,

$$p(d_n^{(u)}|\mathbf{y}_n) = k \cdot \mathcal{CN}(d_n^{(u)}; m_{g_n \to d_n^{(u)}}, \sigma^2_{g_n \to d_n^{(u)}}) \prod_{q'=0}^{Q-1} \mu_{c_{n,q}^{(u)} \to \chi_n^{(u)}}(c_{n,q}^{(u)})|_{c_{n,q}^{(u)} = \chi_q^{-1}[d_n^{(u)}]}, \tag{4.20}$$

This expression can in turn be expressed as a pdf using a mixture of dirac delta functions,

$$p(d_n^{(u)}|\mathbf{y}_n) = k\cdot\sum_{\beta\in\chi}\mathcal{CN}(\beta; m_{g_n\to d_n^{(u)}}, \sigma^2_{g_n\to d_n^{(u)}})\prod_{q'=0}^{Q-1}\mu_{c_{n,q}^{(u)}\to\chi_n^{(u)}}(c_{n,q}^{(u)})|_{c_{n,q}^{(u)}=\chi_q^{-1}[\beta]}\,\delta(d_n^{(u)}-\beta). \quad (4.21)$$

The mean and variance of this *a posteriori* distribution of $d_n^{(u)}$ are given by the following equation,

$$\begin{aligned} m_{d_n^{(u)}|\mathbf{y}_n} &= \frac{\sum_{\beta\in\chi}\beta e^{-\frac{\left|\beta-m_{g_n\to d_n^{(u)}}\right|^2}{\sigma^2_{g_n\to d_n(u)}}-\sum_{q=1}^{Q} l_{c_{n,q}^{(u)}\to\chi_n^{(u)}}\chi_q^{-1}(\beta)}}{\sum_{\beta\in\chi} e^{-\frac{\left|\beta-m_{g_n\to d_n^{(u)}}\right|^2}{\sigma^2_{g_n\to d_n(u)}}-\sum_{q=1}^{Q} l_{c_{n,q}^{(u)}\to\chi_n^{(u)}}\chi_q^{-1}(\beta)}} \\ \sigma^2_{d_n^{(u)}|\mathbf{y}_n} &= \frac{\sum_{\beta\in\chi}|\beta-m_{d_n^{(u)}|\mathbf{y}_n}|^2 e^{-\frac{\left|\beta-m_{g_n\to d_n^{(u)}}\right|^2}{\sigma^2_{g_n\to d_n(u)}}-\sum_{q=1}^{Q} l_{c_{n,q}^{(u)}\to\chi_n^{(u)}}\chi_q^{-1}(\beta)}}{\sum_{\beta\in\chi} e^{-\frac{\left|\beta-m_{g_n\to d_n^{(u)}}\right|^2}{\sigma^2_{g_n\to d_n^{(u)}}}-\sum_{q=1}^{Q} l_{c_{n,q}^{(u)}\to\chi_n^{(u)}}\chi_q^{-1}(\beta)}}. \end{aligned} \quad (4.22)$$

In the expressions for the *a posteriori* mean and variance provided above, $l_{c_{n,q}^{(u)}\to\chi}$ denotes the Log-Likelihood ratio (LLR) of the message from the decoder $c_{n,q}^{(u)}$ to the demodulator χ in Eq. (4.28). The parameter Q represents the order of the modulation scheme employed; for instance Q equals 2 for BPSK, 4 for QPSK, and 16 for 16QAM, among others.

Upon computing the mean and variance of the *a posteriori* distribution, moment matching can be utilized to project this *a posteriori* Probability Mass Function (PMF) onto a Gaussian distribution which is the application of standard minimization of KL-divergence.

$$\begin{aligned} \mu_{d_n^{(u)}\to g_n}(d_n^{(u)}) = \frac{\text{proj}_\Phi\left(p(d_n^{(u)}|\mathbf{y}_n)\right)}{\mu_{g_n\to d_n^{(u)}}(d_n^{(u)})} &= \frac{\mathcal{CN}(d_n^{(u)}; m_{d_n^{(u)}|\mathbf{y}_n}, \sigma^2_{d_n^{(u)}|\mathbf{y}_n})}{\mathcal{CN}(d_n^{(u)}; m_{g_n\to d_n^{(u)}}, \sigma^2_{g_n\to d_n^{(u)}})} \\ &= \mathcal{CN}(d_n^{(u)}; m_{d_n^{(u)}\to g_n}, \sigma^2_{d_n^{(u)}\to g_n}). \end{aligned} \quad (4.23)$$

Finally, the mean $m_{d_n^{(u)}\to g_n}$ and variance $\sigma^2_{d_n^{(u)}\to g_n}$ are given by using the formulas of mean and variance when dividing two Gaussian distributions,

$$\begin{aligned} \sigma^2_{d_n^{(u)}\to g_n} &= \sigma^{-2}_{d_n^{(u)}|\mathbf{y}_n} - \sigma^{-2}_{g_n\to d_n^{(u)}} \\ m_{d_n^{(u)}\to g_n} &= \sigma^2_{d_n^{(u)}\to g_n}\left(\sigma^{-2}_{d_n^{(u)}|\mathbf{y}_n} m_{d_n^{(u)}|\mathbf{y}_n} - \sigma^{-2}_{g_n\to d_n^{(u)}} m_{g_n\to d_n^{(u)}}\right). \end{aligned} \quad (4.24)$$

4.5 Decoding

The message from the demodulator to the decoder, specifically from the modulation constraint node χ to the encoded and interleaved bit $c_{n,q}^{(u)}$, is computed using the EP factor node rule. This computation involves projecting the probabilistic beliefs about the data symbols, derived from the received signal and the channel state information, onto the encoded bits associated with the modulation format. The EP factor node rule facilitates the approximation of complex, non-Gaussian distributions with simpler Gaussian distributions, enabling more tractable computations in iterative decoding algorithms.

The computation of this message is crucial for the iterative decoding process, as it directly impacts the accuracy of the subsequent decoding steps. The computed message is then passed to the decoder, which uses it to update the posterior probabilities of the encoded bits, thereby improving the overall reliability of the decoded information. Applying the EP factor node rule Eq. (2.4) to $\chi_n^{(u)}$ with a projection operator over the set of Bernoulli distributions Θ,

$$\begin{aligned}\mu_{\chi_n^{(u)}\to c_{n,q}^{(u)}}(c_{n,q}^{(u)}) = \Bigg[\text{proj}_\Theta\Bigg(k\cdot\mu_{c_{n,q}^{(u)}\to\chi_n^{(u)}}(c_{n,q}^{(u)})\sum_{c_{n,q'}^{(u)}|q'\neq q}\int\delta(d_n^{(u)}-\chi(\mathbf{c}_n^{(u)}))\\ \times\prod_{q'\neq q}\mu_{c_{n,q'}^{(u)}\to\chi_n^{(u)}}(c_{n,q'}^{(u)})\mu_{d_n^{(u)}\to\chi_n^{(u)}}(d_n^{(u)})dd_n^{(u)}\Bigg)\Bigg]\Bigg/\mu_{c_{n,q}^{(u)}\to\chi_n^{(u)}}(c_{n,q}^{(u)}).\end{aligned} \tag{4.25}$$

The message $\mu_{c_{n,q}^{(u)}\to\chi_n^{(u)}}(c_{n,q}^{(u)})$ is the message form the decoder to the demodulator. By computing the LLR of this message defined as,

$$l_{\chi_n^{(u)}\to c_{n,q}^{(u)}} = \ln\frac{\mu_{\chi_n^{(u)}\to c_{n,q}^{(u)}}(c_{n,q}^{(u)})_{|c_{n,q}^{(u)}=0}}{\mu_{\chi_n^{(u)}\to c_{n,q}^{(u)}}(c_{n,q}^{(u)})_{|c_{n,q}^{(u)}=1}}, \tag{4.26}$$

where $\ln$ stands for the natural logarithm. This LLR after simplications is given as,

$$l_{\chi_n^{(u)}\to c_{n,q}^{(u)}} = \ln\frac{\sum_{d_n^{(u)}|c_{n,q}^{(u)}=0} e^{-\sum_{q'=0}^{Q-1} l_{c_{n,q'}^{(u)}\to\chi_n^{(u)}}\chi_{q'}^{-1}(d_n^{(u)})-\frac{|d_n^{(u)}-m_{g_n\to d_n^{(u)}}|^2}{\sigma^2_{g_n\to d_n^{(u)}}}}}{\sum_{d_n^{(u)}|c_{n,q}^{(u)}=1} e^{-\sum_{q'=0}^{Q-1} l_{c_{n,q'}^{(u)}\to\chi_n^{(u)}}\chi_{q'}^{-1}(d_n^{(u)})-\frac{|d_n^{(u)}-m_{g_n\to d_n^{(u)}}|^2}{\sigma^2_{g_n\to d_n^{(u)}}}}} - l_{c_{n,q}^{(u)}\to\chi_n^{(u)}}, \tag{4.27}$$

where $l_{c_{n,q}^{(u)}\to\chi_n^{(u)}}$ is the LLR of the message from the decoder to the demodulator given by the following equation,

$$l_{c_{n,q}^{(u)}\to\chi_n^{(u)}} = \ln\frac{\mu_{c_{n,q}^{(u)}\to\chi_n^{(u)}}(c_{n,q}^{(u)})_{|c_{n,q}^{(u)}=0}}{\mu_{c_{n,q}^{(u)}\to\chi_n^{(u)}}(c_{n,q}^{(u)})_{|c_{n,q}^{(u)}=1}}. \tag{4.28}$$

obtained for graph-based FEC codes via standard methods described in [76].

4.6 Channel estimation

In the channel estimation subgraph, the estimation process involves applying Kalman filtering in both the forward and backward directions to accurately estimate the channel state. This dual pass is crucial for refining the channel estimates by leveraging the temporal or frequency correlation of the channel across REs. Additionally, the subgraph includes the computation of the messages from the factor node g_n to the channel variable $\mathbf{x}_n^{(u)}$, which represents the channel state for the n-th RE and the u-th user.

The process begins with the detailed computation of the message from g_n to $\mathbf{x}_n^{(u)}$. This message encapsulates the information propagated from observation constraint node g_n to the channel variable node $\mathbf{x}_n^{(u)}$, which is critical for refining the channel estimates. Following this, the Kalman filter is employed to perform forward and backward passes over CE subgraph in Fig. 3.4. The forward pass of the Kalman filter predicts the state of the channel based on past observations, while the backward pass refines these predictions by incorporating future observations, thereby reducing estimation error and enhancing the overall accuracy of the channel estimation process. This iterative approach is essential for obtaining reliable channel estimates, which are foundational for subsequent data detection and decoding stages in the communication system.

Message from g_n to $\mathbf{x}_n^{(u)}$

EP message $\mu_{g_n \to \mathbf{x}_n^{(u)}}(\mathbf{x}_n^{(u)})$ is computed using the EP factor node rule similar to the message $\mu_{g_n \to d_n^{(u)}}(d_n^{(u)})$. The message $\mu_{g_n \to \mathbf{x}_n^{(u)}}(\mathbf{x}_n^{(u)})$ is given by the following EP factor node rule in Eq. (2.4),

$$\mu_{g_n \to \mathbf{x}_n^{(u)}}(\mathbf{x}_n^{(u)}) = \frac{\text{proj}_{\Phi}\left(\mu_{\mathbf{x}_n^{(u)} \to g_n}(\mathbf{x}_n^{(u)}) \tilde{p}(\mathbf{y}_n | \mathbf{x}_n^{(u)})\right)}{\mu_{\mathbf{x}_n^{(u)} \to g_n}(\mathbf{x}_n^{(u)})}, \tag{4.29}$$

where the conditional pdf $\tilde{p}(\mathbf{y}_n | \mathbf{x}_n^{(u)})$ is given by the following equation,

$$\tilde{p}(\mathbf{y}_n | \mathbf{x}_n^{(u)}) = \int p(\mathbf{y}_n | \mathbf{x}_n, \mathbf{d}_n, \boldsymbol{\theta}) \prod_{u' \neq u} \mu_{\mathbf{x}_n^{(u')} \to g_n}(\mathbf{x}_n^{(u')}) \prod_{u'=1}^{U} \mu_{d_n^{(u')} \to g_n}(d_n^{(u')}) \prod_{u'=1}^{U} \mu_{\theta^{(u')} \to g_n}(\theta^{(u')}) d \sim \mathbf{x}_n^{(u)}. \tag{4.30}$$

Eq. (4.30) after subtituting the expressions of the messages $\mu_{\mathbf{x}_n^{(u')} \to g_n}(\mathbf{x}_n^{(u')})$, $\mu_{d_n^{(u')} \to g_n}(d_n^{(u')})$ and $\mu_{\theta^{(u')} \to g_n}(\theta^{(u')})$ becomes

$$\begin{aligned} \tilde{p}(\mathbf{y}_n | \mathbf{x}_n^{(u)}) = \int \int \int \mathcal{CN}(\mathbf{y}_n; \sum_{u=1}^{U} \theta^{(u)} d_n^{(u)} \mathbf{x}_n^{(u)}, \mathbf{R}) \prod_{u' \neq u} \mathcal{CN}(\mathbf{x}_n^{(u')}; \mathbf{m}_{\mathbf{x}_n^{(u')} \to g_n}, \boldsymbol{\Sigma}_{\mathbf{x}_n^{(u')} \to g_n}) \\ \times \prod_{u'=1}^{U} \mathcal{CN}(d_n^{(u')}; m_{d_n^{(u')} \to g_n}, \sigma^2_{d_n^{(u')} \to g_n}) \prod_{u'=1}^{U} \mathcal{CN}(\theta^{(u')}; m_{\theta^{(u')} \to g_n}, \sigma^2_{\theta^{(u')} \to g_n}) \\ \times \prod_{u' \neq u} d\mathbf{x}_n^{(u')} \prod_{u'=1}^{U} d d_n^{(u')} \prod_{u'=1}^{U} d\theta^{(u')}. \end{aligned} \tag{4.31}$$

Applying the projection operator in Eq. (4.2) letting $\mathbf{z} = \mathbf{y}_n$, $\boldsymbol{\pi} = \mathbf{x}_n^{(u)}$ and $\boldsymbol{\theta} = [\{\mathbf{x}_n^{(u')}\}_{u' \neq u}, \{\theta^{(u')}, d_n^{(u')}\}_{u'=1}^{U}]$, this continuous Gaussian mixture becomes a Gaussian distribution in $\mathbf{y}_n$ given $\mathbf{x}_n^{(u)}$ of the form,

$$\tilde{p}(\mathbf{y}_n | \mathbf{x}_n^{(u)}) = \mathcal{CN}(\mathbf{y}_n; \mathbf{m}_{\mathbf{y}_n | \mathbf{x}_n^{(u)}}(\mathbf{x}_n^{(u)}), \boldsymbol{\Sigma}_{\mathbf{y}_n | \mathbf{x}_n^{(u)}}), \tag{4.32}$$

where the mean vector $\mathbf{m}_{\mathbf{y}_n | \mathbf{x}_n^{(u)}}(\mathbf{x}_n^{(u)})$, and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{y}_n | \mathbf{x}_n^{(u)}}(\mathbf{x}_n^{(u)})$ are given by the following equations,

$$\mathbf{m}_{\mathbf{y}_n | \mathbf{x}_n^{(u)}}(\mathbf{x}_n^{(u)}) = h_{\mathbf{x}_n^{(u)} \to g_n} \mathbf{x}_n^{(u)} + \mathbf{I}_{d_n^{(u)} \to g_n}, \quad h_{\mathbf{x}_n^{(u)} \to g_n} = m_{\theta^{(u)} \to g_n} m_{d_n^{(u)} \to g_n}, \tag{4.33}$$

$$\begin{aligned} \boldsymbol{\Sigma}_{\mathbf{y}_n | \mathbf{x}_n^{(u)}} = {} & E_s^{(u)} \boldsymbol{\Gamma}^{(u)} \left[|m_{\theta^{(u)} \to g_n}|^2 \sigma^2_{d_n^{(u)} \to g_n} + \sigma^2_{\theta^{(u)} \to g_n} (|m_{d_n^{(u)} \to g_n}|^2 + \sigma^2_{d_n^{(u)} \to g_n}) \right] \\ & + \sum_{u' \neq u} \boldsymbol{\Sigma}_{\mathbf{x}_n^{(u')} \to g_n} (|m_{\theta^{(u')} \to g_n}|^2 + \sigma^2_{\theta^{(u')} \to g_n})(|m_{d_n^{(u')} \to g_n}|^2 + \sigma^2_{d_n^{(u')} \to g_n}) \\ & + \sum_{u' \neq u} \left[|m_{\theta^{(u')} \to g_n}|^2 \sigma^2_{d_n^{(u')} \to g_n} + \sigma^2_{\theta^{(u')} \to g_n} (|m_{d_n^{(u')} \to g_n}|^2 + \sigma^2_{d_n^{(u')} \to g_n}) \right] \mathbf{m}_{\mathbf{x}_n^{(u')} \to g_n} \mathbf{m}^H_{\mathbf{x}_n^{(u')} \to g_n} \\ & + \mathbf{R}, \end{aligned} \tag{4.34}$$

where $\boldsymbol{\Gamma}^{(u)}$ is the antenna correlation matrix. Using similar arguments as for the covariance matrix $\boldsymbol{\Sigma}_{\mathbf{y}_n | d_n^{(u)}}$ in Eq. (4.11), $\boldsymbol{\Sigma}_{\mathbf{y}_n | \mathbf{x}_n^{(u)}}$ approaches the noise covariance matrix $\mathbf{R}$ at message-passing convergence.

After using the same approach as used in the computation of the message $\mu_{g_n \to d_n^{(u)}}$, the message $\mu_{g_n \to \mathbf{x}_n^{(u)}}(\mathbf{x}_n^{(u)})$ in Eq. (4.29) has the mean $\mathbf{m}_{g_n \to \mathbf{x}_n^{(u)}}$ and the variance $\mathbf{\Sigma}_{g_n \to \mathbf{x}_n^{(u)}}$ given by the following equations,

$$\begin{aligned} \mathbf{m}_{g_n \to \mathbf{x}_n^{(u)}} &= \frac{\mathbf{y}_n - \mathbf{I}_{\mathbf{d}_n^{(u)} \to g_n}}{h_{\mathbf{x}_n^{(u)} \to g_n}} \\ \mathbf{\Sigma}_{g_n \to \mathbf{x}_n^{(u)}} &= \frac{\mathbf{\Sigma}_{\mathbf{y}_n | \mathbf{x}_n^{(u)}}}{|h_{\mathbf{x}_n^{(u)} \to g_n}|^2}. \end{aligned} \tag{4.35}$$

Considering an active user, as the message-passing converges, the mean vector $\mathbf{m}_{g_n \to \mathbf{x}_n^{(u)}}$ is expected to closely approximate the true value of the channel $\mathbf{x}_n^{(u)}$. Concurrently, the covariance matrix $\mathbf{\Sigma}_{g_n \to \mathbf{x}_n^{(u)}}$ is anticipated to converge towards $\frac{\mathbf{R}}{|m_{d_n^{(u)} \to g_n}|^2}$, where $\mathbf{R}$ represents the noise covariance matrix, and $m_{d_n^{(u)} \to g_n}$ is the message passed from the data symbol node to the factor node g_n. This convergence signifies the algorithm's ability to refine its estimates, resulting in more accurate representations of both the channel and the associated uncertainty, as influenced by the noise in the system.

Messages inside the CE subgraph

Forward and backward passes inside the CE subgraph for EP are same as BP given by the following equations,

$$\begin{aligned} \mu_{f_n^{(u)} \to \mathbf{x}_n^{(u)}}(\mathbf{x}_n^{(u)}) &\propto \mathcal{CN}(\mathbf{x}_n^{(u)} : \mathbf{m}_{f_n^{(u)} \to \mathbf{x}_n^{(u)}}, \mathbf{\Sigma}_{f_n^{(u)} \to \mathbf{x}_n^{(u)}}) \\ \mu_{f_{n+1}^{(u)} \to \mathbf{x}_n^{(u)}}(\mathbf{x}_n^{(u)}) &\propto \mathcal{CN}(\mathbf{x}_n^{(u)} : \mathbf{m}_{f_{n+1}^{(u)} \to \mathbf{x}_n^{(u)}}, \mathbf{\Sigma}_{f_{n+1}^{(u)} \to \mathbf{x}_n^{(u)}}), \end{aligned} \tag{4.36}$$

whose mean and covariance are updated similarly to Kalman filtering and smoothing (we omit the details here, the interested reader is referred to [76, Fig. 15], [86]).

After the forward and backward passes of the CE subgraph, the extrinsic smoothing pass computes the message $\mu_{\mathbf{x}_n^{(u)} \to g_n}(\mathbf{x}_n^{(u)})$, obtained by applying the EP variable node rule in Eq. (2.5)

$$\mu_{\mathbf{x}_n^{(u)} \to g_n}(\mathbf{x}_n^{(u)}) = \mu_{f_n^{(u)} \to \mathbf{x}_n^{(u)}}(\mathbf{x}_n^{(u)}) \mu_{f_{n+1}^{(u)} \to \mathbf{x}_n^{(u)}}(\mathbf{x}_n^{(u)}) \propto \mathcal{CN}(\mathbf{x}_n^{(u)}; \mathbf{m}_{\mathbf{x}_n^{(u)} \to g_n}, \mathbf{\Sigma}_{\mathbf{x}_n^{(u)} \to g_n}), \tag{4.37}$$

whose mean and covariance are computed as

$$\begin{aligned} \mathbf{\Sigma}_{\mathbf{x}_n^{(u)} \to g_n}{}^{-1} &= \mathbf{\Sigma}_{f_n^{(u)} \to \mathbf{x}_n^{(u)}}{}^{-1} + \mathbf{\Sigma}_{f_{n+1}^{(u)} \to \mathbf{x}_n^{(u)}}{}^{-1} \\ \mathbf{\Sigma}_{\mathbf{x}_n^{(u)} \to g_n}{}^{-1} \mathbf{m}_{\mathbf{x}_n^{(u)} \to g_n} &= \mathbf{\Sigma}_{f_n^{(u)} \to \mathbf{x}_n^{(u)}}{}^{-1} \mathbf{m}_{f_n^{(u)} \to \mathbf{x}_n^{(u)}} + \mathbf{\Sigma}_{f_{n+1}^{(u)} \to \mathbf{x}_n^{(u)}}{}^{-1} \mathbf{m}_{f_{n+1}^{(u)} \to \mathbf{x}_n^{(u)}}. \end{aligned} \tag{4.38}$$

.

4.7 User activity detection

Let us now turn our attention to UAD, whose task is to estimate user activity variables $\theta^{(u)}$. Initial attempts to apply naive EP for UAD were found to be highly ineffective. Specifically, using the same EP-based approach for UAD as that used for symbol detection and channel estimation—where the projection operator onto the set of Gaussian distributions is the one introduced in Sec. 4.3—resulted in a significantly high probability of false alarms. This ineffectiveness arises because the messages $\mu_{g_n \to \theta^{(u)}}(\theta^{(u)})$ become uninformative under this

method, failing to accurately convey the necessary information about user activity.

This challenge provided the motivation to hybridize EP with BP, followed by a Gaussian approximation to keep the same complexity. This hybrid approach, known as Gaussian Belief Propagation (GaBP) [30], is then employed for UAD. GaBP offers a more effective strategy for user activity detection by combining the strengths of BP with the Gaussian approximation, thereby enhancing the informativeness and reliability of the messages involved in the detection process. Through this method, the probability of false alarms is significantly reduced, resulting in a more robust and accurate UAD.

Message from g_n to $\theta^{(u)}$

The message $\mu_{g_n \to \theta^{(u)}}(\theta^{(u)})$ can be computed in a manner similar to the computation of messages $\mu_{g_n \to d_n^{(u)}}(d_n^{(u)})$ and $\mu_{g_n \to \mathbf{x}_n^{(u)}}(\mathbf{x}_n^{(u)})$ in Sec. 4.4 and Sec. 4.6, respectively. However, there are notable differences in the approach for $\mu_{g_n \to \theta^{(u)}}(\theta^{(u)})$. Specifically, instead of employing the operator introduced in Sec. 4.3 whose aim is to remove the contribution of $\theta^{(u)}$ inside the covariance of the projected Gaussian, the standard KL-divergence that keeps this dependence is utilized instead. This subtle change enables to discriminate inactive users from active ones.

Furthermore, instead of using the EP factor node rule typically applied in other message computations, the BP factor node rule is used. The computation of the message $\mu_{g_n \to \theta^{(u)}}(\theta^{(u)})$ is formalized by the BP factor node rule in Eq. (2.2) to g_n,

$$\mu_{g_n \to \theta^{(u)}}(\theta^{(u)}) = k \cdot \tilde{p}(\mathbf{y}_n|\theta^{(u)}), \tag{4.39}$$

where the pdf $\tilde{p}(\mathbf{y}_n|\theta^{(u)})$ is given by the following equation,

$$\tilde{p}(\mathbf{y}_n|\theta^{(u)}) = \int p(\mathbf{y}_n|\mathbf{x}_n, \mathbf{d}_n, \boldsymbol{\theta}) \prod_{u' \neq u} \mu_{\theta^{(u')} \to g_n}(\theta^{(u')}) \prod_{u'=1}^{U} \mu_{\mathbf{x}_n^{(u')} \to g_n}(\mathbf{x}_n^{(u')}) \prod_{u'=1}^{U} \mu_{d_n^{(u')} \to g_n}(d_n^{(u')}) d \sim \theta^{(u)}. \tag{4.40}$$

This conditional pdf after subtituting the expressions of the messages $\mu_{\mathbf{x}_n^{(u')} \to g_n}(\mathbf{x}_n^{(u')})$, $\mu_{d_n^{(u')} \to g_n}(d_n^{(u')})$ and $\mu_{\theta^{(u')} \to g_n}(\theta^{(u')})$ becomes,

$$\begin{aligned} \tilde{p}(\mathbf{y}_n|\theta^{(u)}) = \int \int \int \mathcal{CN}(\mathbf{y}_n; \sum_{u=1}^{U} \theta^{(u)} d_n^{(u)} \mathbf{x}_n^{(u)}, \mathbf{R}) \prod_{u' \neq u} \mathcal{CN}(\theta^{(u')}; m_{\theta^{(u')} \to g_n}, \sigma^2_{\theta^{(u')} \to g_n}) \\ \times \prod_{u'=1}^{U} \mathcal{CN}(\mathbf{x}_n^{(u')}; \mathbf{m}_{\mathbf{x}_n^{(u')} \to g_n}, \boldsymbol{\Sigma}_{\mathbf{x}_n^{(u')} \to g_n}) \prod_{u'=1}^{U} \mathcal{CN}(d_n^{(u')}; m_{d_n^{(u')} \to g_n}, \sigma^2_{d_n^{(u')} \to g_n}) \\ \times \prod_{u' \neq u} d\theta^{(u')} \prod_{u'=1}^{U} d\mathbf{x}_n^{(u')} \prod_{u'=1}^{U} d d_n^{(u')} \end{aligned} \tag{4.41}$$

This continuous Gaussian mixture can be resolved into a single Gaussian distribution of the form,

$$\tilde{q}(\mathbf{y}_n|\theta^{(u)}) = \mathcal{CN}(\mathbf{y}_n; \mathbf{m}_{\mathbf{y}_n|\theta^{(u)}}(\theta^{(u)}), \boldsymbol{\Sigma}_{\mathbf{y}_n|\theta^{(u)}}(\theta^{(u)})) \tag{4.42}$$

by minimizing $KL(\tilde{p}||\tilde{q})$.

It follows that, the mean $\mathbf{m}_{\mathbf{y}_n|\theta^{(u)}}(\theta^{(u)})$ and covariance $\boldsymbol{\Sigma}_{\mathbf{y}_n|\theta^{(u)}}(\theta^{(u)})$ are given by the following equations,

$$\begin{aligned}
&\mathbf{m}_{\mathbf{y}_n|\theta^{(u)}}(\theta^{(u)}) = \mathbf{h}_{\theta^{(u)}\to g_n}\theta^{(u)} + \mathbf{I}_{d_n^{(u)}\to g_n} \\
&\mathbf{h}_{\theta^{(u)}\to g_n} = m_{d_n^{(u)}\to g_n}\mathbf{m}_{\mathbf{x}_n^{(u)}\to g_n}, \qquad \mathbf{I}_{d_n^{(u)}\to g_n} = \sum_{u'\neq u} m_{d_n^{(u')}\to g_n}\mathbf{m}_{\mathbf{x}_n^{(u')}\to g_n} m_{\theta^{(u')}\to g_n},
\end{aligned} \tag{4.43}$$

$$\begin{aligned}
\mathbf{\Sigma}_{\mathbf{y}_n|\theta^{(u)}}(\theta^{(u)}) = &\left[|m_{d_n^{(u)}\to g_n}|^2\mathbf{\Sigma}_{\mathbf{x}_n^{(u)}\to g_n} + \sigma^2_{d_n^{(u)}\to g_n}(\mathbf{m}_{\mathbf{x}_n^{(u)}\to g_n}\mathbf{m}^H_{\mathbf{x}_n^{(u)}\to g_n} + \mathbf{\Sigma}_{\mathbf{x}_n^{(u)}\to g_n})\right]|\theta^{(u)}|^2 \\
&+ \sum_{u'\neq u}\sigma^2_{\theta^{(u')}\to g_n}(|m_{d_n^{(u')}\to g_n}|^2 + \sigma^2_{d_n^{(u')}\to g_n})(\mathbf{m}_{\mathbf{x}_n^{(u')}\to g_n}\mathbf{m}^H_{\mathbf{x}_n^{(u')}\to g_n} + \mathbf{\Sigma}_{\mathbf{x}_n^{(u')}\to g_n}) \\
+\sum_{u'\neq u}|m_{\theta^{(u')}\to g_n}|^2&\left[|m_{d_n^{(u')}\to g_n}|^2\mathbf{\Sigma}_{\mathbf{x}_n^{(u')}\to g_n} + \sigma^2_{d_n^{(u')}\to g_n}(\mathbf{m}_{\mathbf{x}_n^{(u')}\to g_n}\mathbf{m}^H_{\mathbf{x}_n^{(u')}\to g_n} + \mathbf{\Sigma}_{\mathbf{x}_n^{(u')}\to g_n})\right] + \mathbf{R}.
\end{aligned} \tag{4.44}$$

As the message $\mu_{g_n\to\theta^{(u)}}(\theta^{(u)})$ is a PMF of $\theta^{(u)}$, with support $\theta^{(u)} \in \{0, 1\}$, we can compute the LLR of this message defined as,

$$l_{g_n\to\theta^{(u)}} = \ln\frac{\mu_{g_n\to\theta^{(u)}}(\theta^{(u)})_{|\theta^{(u)}=0}}{\mu_{g_n\to\theta^{(u)}}(\theta^{(u)})_{|\theta^{(u)}=1}}. \tag{4.45}$$

After subsituting the values of the message $\mu_{g_n\to\theta^{(u)}}(\theta^{(u)})$ at $\theta^{(u)} = 0$ and $\theta^{(u)} = 1$, the LLR in Eq. (4.45) is given by the following equation,

$$\begin{aligned}
l_{g_n\to\theta^{(u)}} = &(\mathbf{y}_n - \mathbf{m}_{\mathbf{y}_n|\theta^{(u)}}(1))^H\mathbf{\Sigma}_{\mathbf{y}_n|\theta^{(u)}}(1)^{-1}(\mathbf{y}_n - \mathbf{m}_{\mathbf{y}_n|\theta^{(u)}}(1)) + \ln\det\mathbf{\Sigma}_{\mathbf{y}_n|\theta^{(u)}}(1) \\
&-\Big((\mathbf{y}_n - \mathbf{m}_{\mathbf{y}_n|\theta^{(u)}}(0))^H\mathbf{\Sigma}_{\mathbf{y}_n|\theta^{(u)}}(0)^{-1}(\mathbf{y}_n - \mathbf{m}_{\mathbf{y}_n|\theta^{(u)}}(0)) + \ln\det\mathbf{\Sigma}_{\mathbf{y}_n|\theta^{(u)}}(0)\Big).
\end{aligned} \tag{4.46}$$

Interpretation: These LLRs play a crucial role in the subsequent steps of the inference process. Specifically, they are utilized for the computation of the messages $\mu_{\theta^{(u)}\to g_n}(\theta^{(u)})$, which represent the information passed from the user activity variable $\theta^{(u)}$ to the factor node g_n within the factor graph.

The LLRs are also instrumental in deriving the *a posteriori* PMF of the user activity variables $\theta^{(u)}$. This *a posteriori* PMF provides a probabilistic estimate of the activity status of each user, taking into account the observed data and the prior information. By accurately computing these messages and the corresponding *a posteriori* PMF, the system can effectively perform UAD.

The $l_{g_n\to\theta^{(u)}}$, must be interpreted differently for active and inactive users. To do this, it is necessary to understand how the mean $\mathbf{m}_{\mathbf{y}_n|\theta^{(u)}}(\theta^{(u)})$ and variance $\mathbf{\Sigma}_{\mathbf{y}_n|\theta^{(u)}}(\theta^{(u)})$ behave as the algorithm converges.

For an active user, the estimate $\mathbf{h}_{\theta^{(u)}\to g_n}$ approaches the true transmitted signal $d_n^{(u)}\mathbf{x}_n^{(u)}$, while the interference term $\mathbf{I}_{d_n^{(u)}\to g_n}$ converges toward the actual Multi-access Interference (MAI). This implies that as the algorithm converges, the residual between the received signal and the predicted signal, when $\theta^{(u)}$ is hypothesized as 1, approaches the noise term:

$$\mathbf{y}_n - \mathbf{m}_{\mathbf{y}_n|\theta^{(u)}}(1) \approx \mathbf{w}_n, \tag{4.47}$$

where $\mathbf{w}_n$ represents the noise. Similarly, when $\theta^{(u)}$ is hypothesized as 0, we have:

$$\mathbf{y}_n - \mathbf{m}_{\mathbf{y}_n|\theta^{(u)}}(0) \approx d_n^{(u)}\mathbf{x}_n^{(u)} + \mathbf{w}_n. \tag{4.48}$$

Both variances, $\mathbf{\Sigma}_{\mathbf{y}_n|\theta^{(u)}}(1)$ and $\mathbf{\Sigma}_{\mathbf{y}_n|\theta^{(u)}}(0)$, converge to the noise covariance $\mathbf{R} = N_0\mathbf{I}_{N_R}$. Consequently, at high Signal-to-noise Ratio (SNR), the LLR $l_{g_n\to\theta^{(u)}}$ becomes negative, as expressed in the equation:

$$l_{g_n\to\theta^{(u)}} \approx \frac{||\mathbf{w}_n||_2^2}{N_0} - \frac{||d_n^{(u)}\mathbf{x}_n^{(u)} + \mathbf{w}_n||_2^2}{N_0}. \tag{4.49}$$

Here, the term $||d_n^{(u)}\mathbf{x}_n^{(u)} + \mathbf{w}_n||_2^2$ exceeds $||\mathbf{w}_n||_2^2$ in the positive SNR range, as the signal component $d_n^{(u)}\mathbf{x}_n^{(u)}$ dominates the noise.

For an inactive user, irrespective of the iteration index $m_{d_n^{(u)}\to g_n}, \mathbf{m}_{\mathbf{x}_n^{(u)}\to g_n}, \sigma^2_{\theta^{(u')}\to g_n}$ and $\mathbf{\Sigma}_{\mathbf{x}_n^{(u)}\to g_n}$ stay close to their initializations at their *a priori* values. Consequently, $\mathbf{h}_{\theta^{(u)}\to g_n}$ remains close to zero. For the same reason the first line in Eq. (4.44) remains close to $E_s^{(u)}\mathbf{\Gamma}^{(u)}|\theta^{(u)}|^2$.

As a result, $\mathbf{m}_{\mathbf{y}_n|\theta^{(u)}}(\theta^{(u)})$ approaches $\mathbf{I}_{d_n^{(u)}\to g_n}$ both when $\theta^{(u)}$ is hypothesized as 0 or 1. At convergence, the interference term $\mathbf{I}_{d_n^{(u)}\to g_n}$ approaches the true MAI, meaning the residual between the received and predicted signals, $\mathbf{y}_n - \mathbf{m}_{\mathbf{y}_n|\theta^{(u)}}(\theta^{(u)})$, approximates $\mathbf{w}_n$ for both hypotheses.

For simplicity consider the particular case where $\mathbf{\Gamma}^{(u)} = \mathbf{I}_{N_R}$, then the covariance $\mathbf{\Sigma}_{\mathbf{y}_n|\theta^{(u)}}(1)$ converges to $(E_s^{(u)} + N_0)\mathbf{I}_{N_R}$, where $E_s^{(u)}$ represents the signal power, while $\mathbf{\Sigma}_{\mathbf{y}_n|\theta^{(u)}}(0)$ approaches $N_0\mathbf{I}_{N_R}$. Therefore, the LLR $l_{g_n\to\theta^{(u)}}$ for the inactive user becomes:

$$l_{g_n\to\theta^{(u)}} \approx \frac{||\mathbf{w}_n||_2^2}{E_s^{(u)} + N_0} + N_R\ln(E_s^{(u)} + N_0) - \left(\frac{||\mathbf{w}_n||_2^2}{N_0} + N_R\ln(N_0)\right). \tag{4.50}$$

For the user to be classified as inactive, the LLR must be positive:

$$\ln\left(1 + \frac{E_s^{(u)}}{N_0}\right)) > \frac{||\mathbf{w}_n||_2^2}{N_0N_R}\left(1 - \frac{1}{1 + \frac{E_s^{(u)}}{N_0}}\right). \tag{4.51}$$

This inequality is generally valid for moderate to high SNR values, under the assumption that $\mathbf{w}_n$ is zero-mean white Gaussian noise (WGN) with covariance $\mathbf{R} = N_0\mathbf{I}_{N_R}$.

Message from $\theta^{(u)}$ to g_n

The computation of the message $\mu_{\theta^{(u)}\to g_n}$ is done using the BP variable node rule in (2.3) to $\theta^{(u)}$, so this message is given by the following equation,

$$\mu_{\theta^{(u)}\to g_n}(\theta^{(u)}) = k\cdot p(\theta^{(u)}) \prod_{n'=0|n'\neq n}^{N-1} \mu_{g_{n'}\to\theta^{(u)}}(\theta^{(u)}), \tag{4.52}$$

where $p(\theta^{(u)})$ is the prior PMF of $\theta^{(u)}$, k is the normalization constant, and the product is over all REs except the current RE n.

Using the LLR form of the messages $\mu_{g'_n\to\theta^{(u)}}(\theta^{(u)})$, defined in Eq. (4.45), we obtain

$$\mu_{\theta^{(u)}\to g_n}(\theta^{(u)}) = k\cdot p(\theta^{(u)}) \prod_{n'=0|n'\neq n}^{N-1} e^{-l_{g_{n'},\to\theta^{(u)}}\theta^{(u)}} \tag{4.53}$$

$$k = \frac{1}{\sum_{\theta^{(u)}=0}^{1} p(\theta^{(u)})e^{-\sum_{n'\neq n} l_{g_{n'}\to\theta^{(u)}}\theta^{(u)}}}. \tag{4.54}$$

Again this PMF can be projected to a Gaussian distribution of the form,

$$\mathcal{CN}(\theta^{(u)}; m_{\theta^{(u)}\to g_n}, \sigma^2_{\theta^{(u)}\to g_{n'}}) \tag{4.55}$$

Consequently by minimizing KL divergence between Eqs. (4.53) and (4.55), the expressions of the mean $m_{\theta^{(u)}\to g_n}$ and variance $\sigma^2_{\theta^{(u)}\to g_n}$ are given by the following equations,

$$\begin{aligned} m_{\theta^{(u)}\to g_n} &= \frac{p_a^{(u)} e^{-\sum_{n'\neq n} l_{g_n\to\theta^{(u)}}\theta^{(u)}}}{1-p_a^{(u)}+p_a^{(u)} e^{-\sum_{n'\neq n} l_{g_{n'}\to\theta^{(u)}}\theta^{(u)}}} \\ \sigma^2_{\theta^{(u)}\to g_{n'}} &= m_{\theta^{(u)}\to g_n}(1-m_{\theta^{(u)}\to g_n}), \end{aligned} \tag{4.56}$$

where $p_a^{(u)}$ is the prior probability of user u being active.

The LLR of aposteriori PMF of $\theta^{(u)}$ can be computed using the LLRs $l_{g_n\to\theta^{(u)}}$ using the following equation,

$$l_{\theta^{(u)}|\{\mathbf{y}_n\}_{n=0}^{N-1}} = \ln\frac{p(\theta^{(u)}|\{\mathbf{y}_n\}_{n=0}^{N-1})_{|\theta^{(u)}=0}}{p(\theta^{(u)}|\{\mathbf{y}_n\}_{n=0}^{N-1})_{|\theta^{(u)}=1}} = \ln(1-p_a^{(u)}) - \ln(p_a^{(u)}) + \sum_{n=0}^{N-1} l_{g_n\to\theta^{(u)}} \tag{4.57}$$

It follows that hard UAD is obtained by performing hypothesis testing

$$\hat{\theta}^{(u)} = \begin{cases} 1 & \text{if } l_{\theta^{(u)}|\{\mathbf{y}_n\}_{n=0}^{N-1}} < 0 \\ 0 & \text{otherwise.} \end{cases} \tag{4.58}$$

4.8 Hyperparameters estimation

In the system model, the parameters $E_s^{(u)}$ (signal energy of user u) and $\rho^{(u)}$ (inter-antenna correlation of user u) are typically unknown and must be estimated before running the message-passing algorithm. For the sake of simplicity, this section will consider pilot-only parameter estimation for the orthogonal pilot setting defined by Fig. 3.2.

We consider two methods for this estimation:

The first method is Maximum-Likelihood (ML) estimation, which identifies the parameter values that that make the observations most probable. This method is accurate but can be computationally intensive.

The second method is heuristic estimation based on sample means. This approach estimates parameters by aligning sample moments, like mean and variance, with the theoretical moments of the distribution. While simpler and faster than ML estimation, it remains effective.

The EP algorithm uses heuristic hyperparameter estimation because it is computationally efficient and performs similarly to ML estimation, especially at high SNRs, where the difference in accuracy is negligible.

The hyperparameters based on the heuristic estimation are computed using the following equations,

$$\hat{E}_s^{(u)} = \frac{1}{|\mathcal{P}^{(u)}|N_R} \sum_{n\in\mathcal{P}^{(u)}} \sum_{r=1}^{N_R} \frac{|y_{n,r}|^2 - N_0}{|d_n^{(u)}|^2} \tag{4.59}$$

$$\hat{\rho}^{(u)} = \frac{1}{\hat{E}_s^{(u)} |\mathcal{P}^{(u)}| (N_R - 1)} \sum_{n \in \mathcal{P}^{(u)}} \sum_{r=1}^{N_R - 1} \frac{y_{n,r} y_{n,r+1}^*}{|d_n^{(u)}|^2}, \tag{4.60}$$

where $\mathcal{P}^{(u)}$ is the set of pilot subcarriers of u-th user.

Alternatively, ML estimation of the parameters of the u-th user boils down to

$$(\hat{E}_s^{(u),ML}, \hat{\rho}^{(u),ML}) = \arg \min_{(E_s^{(u)}, \rho^{(u)})} \sum_{n \in \mathcal{P}^{(u)}} \left(\mathbf{y}_n^H (|d_n^{(u)}|^2 E_s^{(u)} \mathbf{\Gamma}^{(u)} + \mathbf{R})^{-1} \mathbf{y}_n + \log \det(|d_n^{(u)}|^2 E_s^{(u)} \mathbf{\Gamma}^{(u)} + \mathbf{R}) \right). \tag{4.61}$$

A constrained optimization algorithm can be used to solve this minimization problem. In the simulation results, BLEIC [97] algorithm has been used.

At the first iteration of the proposed receiver, if for the u-th user $\hat{E}_s^{(u)} < N_0$ (i.e. below the noise floor) the u-th user is pre-classified as inactive and we let

$$\begin{cases} \hat{E}_s^{(u)} = \max\limits_{u'=1,\dots,U} \hat{E}_s^{(u')} \\ \hat{\rho}^{(u)} = \max\limits_{u'=1,\dots,U} \hat{\rho}^{(u')}. \end{cases} \tag{4.62}$$

so as to maximize the uncertainty measured by the aforementioned hyperparameter-dependent covariance matrices. During subsequent iterations, we apply the same procedure, except that the u-th user is pre-classified as inactive, when at the previous iteration $\hat{\theta}^{(u)} = 0$ in Eq. (4.58).

4.9 Receiver implementation

In this section, we first present the detailed initialization of the messages required to start the algorithm followed by the scheduling of the hybrid EP-BP algorithm. The scheduling outlines the step-by-step execution of the algorithm, describing how the EP and BP components are integrated and interact with each other throughout the iterative process. This includes the order in which the messages are passed, the conditions under which different parts of the algorithm are updated, and how convergence is determined.

Initilization

Before commencing the first iteration of the hybrid EP-BP algorithm, it is essential to properly initialize the messages. This initialization is crucial for ensuring that the iterative process begins with reasonable estimates, which can significantly influence the convergence behavior and accuracy of the algorithm.

First, we initialize the messages $\mu_{c_{n,q}^{(u)} \to \chi_n^{(u)}}(c_{n,q}^{(u)})$, which represent the messages passed from the encoded and interleaved bit $c_{n,q}^{(u)}$ to the modulation constraint $\chi_n^{(u)}$. This message is initialized to a uniform PMF across all coded digits and for all users. The uniform initialization reflects an equal likelihood for each bit value, acknowledging that no prior information is available before the iterations begin.

For the pilot and information REs, the initialization of the messages is handled differently. On a pilot REs, the message $\mu_{d_n^{(u)} \to g_n}(d_n^{(u)})$, which is passed from the data symbol $d_n^{(u)}$ to the function node g_n, is initialized as a Gaussian distribution with its mean equal to the pilot symbol and with zero variance. This initialization ensures that the pilot symbols, which are known *a priori*, are accurately represented at the start of the iterations. On information REs,

where the data symbols are not known in advance, the message $\mu_{d_n^{(u)} \to g_n}(d_n^{(u)})$ is initialized to a Gaussian distribution with zero mean and unit variance. This reflects an initial state of uncertainty about the data symbols.

Similarly, the messages $\mu_{\mathbf{x}_n^{(u)} \to g_n}(\mathbf{x}_n^{(u)})$, which are passed from the channel $\mathbf{x}_n^{(u)}$ to g_n, are initialized as a complex Gaussian distribution with zero mean and a covariance matrix $E_s^{(u)}\mathbf{\Gamma}^{(u)}$. Here, $E_s^{(u)}$ is the energy per symbol of user u, and $\mathbf{\Gamma}^{(u)}$ represents the antenna correlation matrix. This initialization takes into account the energy and correlation properties of the channel before the iterations begin.

Finally, for the user activity variable $\theta^{(u)}$, the messages $\mu_{\theta^{(u)} \to g_n}(\theta^{(u)})$ are initialized with a complex Gaussian distribution $\mathcal{CN}(\theta^{(u)}; 1, 0)$. This corresponds to assuming that all users are active (i.e., $\theta^{(u)} = 1$) before any processing takes place, with zero variance, meaning complete certainty in this assumption at the outset. This particular initialization was found to yield the best performance for the proposed method, as it provides a strong initial hypothesis that can be refined through subsequent iterations.

Overall, these initialization steps are vital for setting up the algorithm with informed starting points, thereby enhancing the efficiency and effectiveness of the message-passing process in the hybrid EP-BP algorithm.

Message-passing schedule

Over a loopy graph such as the one considered in Fig. 3.4, a wide variety of message-passing schedules are possible. We choose to process all U user subgraphs one at a time with the following serial schedule:

1. For the current user index u, we reset $\mu_{\theta^{(u)} \to g_n}(\theta^{(u)}) = \mathcal{CN}(\theta^{(u)}; 1, 0)$, which improves the Probability of Missed Detection (P_{md}) and the Probability of False Detection (P_{fa}) considerably
2. CE subgraph processing executing the steps in Sec. 4.6 in that order
3. DEM subgraph processing executing the steps in Sec. 4.4 in that order
4. DEC subgraph processing (see Sec. 4.5)
5. UAD subgraph processing executing the steps in Sec. 4.7 in that order.

The rationale for initiating the process with CE lies in the critical importance of accurate channel estimates for the performance of the entire system. All subsequent processing stages, such as symbol detection and user activity detection, rely heavily on the quality of these channel estimates. If the channel estimates are inaccurate, it could lead to significant errors in the following stages, thereby degrading the overall system performance.

Moreover, the decision to perform UAD after DEC rather than after DEM is based on the fact that DEC enhances the reliability of the symbol estimates. In other words, the decoding process improves the confidence in the detected symbols, which makes it more effective to carry out UAD at this stage. If UAD were performed immediately after DEM, it might lead to a higher probability of false detections due to less reliable symbol estimates.

Finally, given that the underlying factor graph used in the message-passing algorithm contains loops (i.e., the graph is loopy), the proposed schedule of operations—starting with CE, followed by DEM and DEC, and concluding with UAD—must be iterated multiple times, as illustrated in the receiver flowchart in Fig. 4.1. Specifically, the schedule needs to be repeated N_{it} times until the algorithm converges to a stable solution. The iteration process allows the algorithm to refine its estimates progressively, ultimately leading to improved accuracy and reliability of the system's outputs.

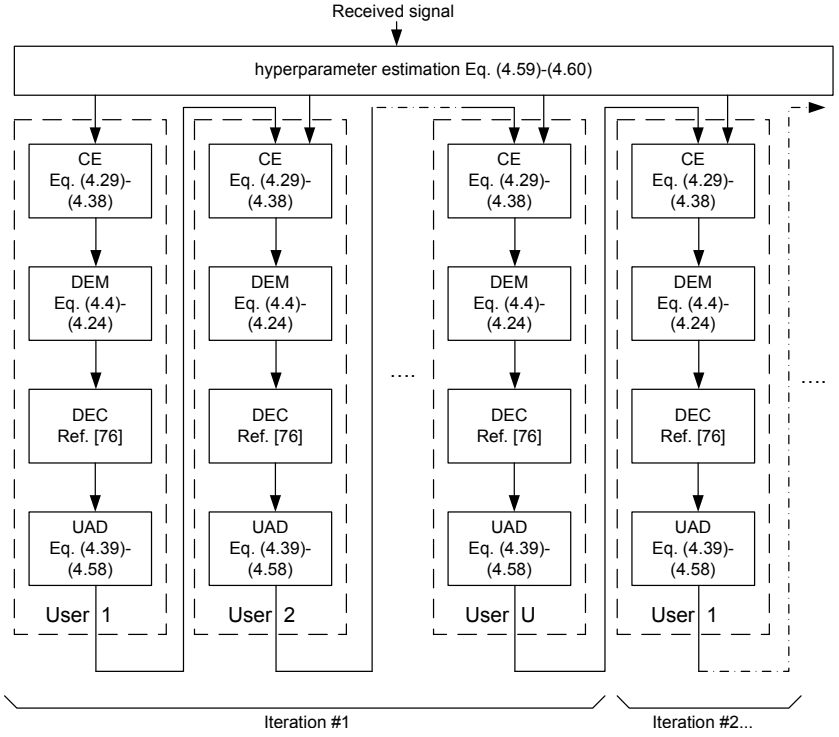

Figure 4.1: Receiver flow chart.

Algorithm 1 Hybrid EP/BP

Require: Observations $\mathbf{y}_n$ for $n = 1, 2, \ldots, N$; Pilot symbols $d_{n_1}^{(u)}, d_{n_2}^{(u)}, \ldots, d_{n_P}^{(u)}$; Pilot RE indices, $n_1^{(u)}, n_2^{(u)}, \ldots, n_P^{(u)}, E_s^{(u)}, \rho^{(u)}, \forall u \in \{1, \ldots, U\}$

Ensure: Estimated signal parameters $\hat{d}_n^{(u)}, \hat{\mathbf{x}}_n^{(u)}$ and $\hat{\theta}^{(u)}$

Initialize the messages $\mu_{d_n^{(u)} \to g_n}(d_n^{(u)}) = \mathcal{CN}(d_n^{(u)}; 0, 1)$, $\mu_{\mathbf{x}_n^{(u)} \to g_n}(\mathbf{x}_n^{(u)}) = \mathcal{CN}(\mathbf{x}_n^{(u)}; \mathbf{0}, E_s^{(u)} \mathbf{\Gamma}^{(u)})$ and $\mu_{\theta^{(u)} \to g_n}(\theta^{(u)}) = \mathcal{CN}(\theta^{(u)}; 1, 0)$

for $iteration = 1$ **to** N_{it} **do**

  Re-initialize $\mu_{\theta^{(u)} \to g_n}(\theta^{(u)}) = \mathcal{CN}(\theta^{(u)}; 1, 0)$

  for $u = 1$ **to** U **do**

    CE:

    Compute the message $\mu_{g_n \to \mathbf{x}_n^{(u)}}(\mathbf{x}_n^{(u)}), \forall n \in \{0, ..., N-1\}$ using the Eq. (4.35), where $h_{\mathbf{x}_n^{(u)} \to g_n}$ is given by the Eq. (4.33) and $\mathbf{\Sigma}_{\mathbf{y}_n | \mathbf{x}_n^{(u)}}$ is given by the Eq. (4.34)

    Compute the forward and backward passes of the CE subgraph, $\mu_{f_n^{(u)} \to \mathbf{x}_n^{(u)}}(\mathbf{x}_n^{(u)})$ and $\mu_{f_{n+1}^{(u)} \to \mathbf{x}_n^{(u)}}(\mathbf{x}_n^{(u)})$ using Eq. (4.36), and then the message $\mu_{\mathbf{x}_n^{(u)} \to g_n}(\mathbf{x}_n^{(u)}), \forall n \in \{0, ..., N-1\}$ using the Eq. (4.38)

    DEM (1):

    Compute the message $\mu_{g_n \to d_n^{(u)}}(d_n^{(u)}), \forall n \in \{0, ..., N-1\}$ using the Eq. (4.18), where $\mathbf{h}_{d_n^{(u)} \to g_n}, \mathbf{I}_{d_n^{(u)} \to g_n}$ is given by the Eq. (4.10) and $\mathbf{\Sigma}_{\mathbf{y}_n | d_n^{(u)}}$ is given by the Eq. (4.11)

    DEC:

    Compute the LLR $l_{\chi_n^{(u)} \to c_{n,q}^{(u)}}$ using the Eq. (4.27)

    Rest of the decoder LLRs are computed using the standard BP

    DEM (2):

    Compute the message $\mu_{d_n^{(u)} \to g_n}(d_n^{(u)}), \forall n \in \{0, ..., N-1\}$ using the Eq. (4.24), where $m_{d_n^{(u)} | \mathbf{y}_n}$ and $\sigma^2_{d_n^{(u)} | \mathbf{y}_n}$ are given by the Eq. (4.22)

    UAD:

    Compute the message $\mu_{g_n \to \theta^{(u)}}(\theta^{(u)}), \forall n \in \{0, ..., N-1\}$ using the Eq. (4.46)

    Compute the message $\mu_{\theta^{(u)} \to g_n}(\theta^{(u)}), \forall n \in \{0, ..., N-1\}$ using the Eq. (4.56)

  end for

end for

return $BER^{(u)}, CE\ MSE^{(u)}, P_{md}^{(u)}, P_{fa}^{(u)}$

4.9.1 Numerical stability

In this algorithm, numerical stability is maintained by addressing potential overflow or underflow issues when dealing with bit or symbol probabilities. Specifically, these issues are mitigated by working in the log domain and using LLRs. LLRs are instrumental in avoiding extreme values during the iterative processes involved in the algorithm.

Additionally, the Gaussian message-passing process in EP involves dividing two Gaussian densities. This operation is mathematically valid only when the variance or covariance matrix involved is semi positive-definite. In rare cases where this condition does not hold, we follow the solution proposed in the literature by Senst and Ascheid (2011) [88], where the division of Gaussian distributions is replaced by considering only the numerator of the Gaussian which is the belief of the hidden variable, effectively bypassing the problematic division.

Beyond these general numerical challenges, the proposed algorithm encounters a few specific numerical issues that are worth addressing.

In cases where the expression $\|\mathbf{h}_{d_n^{(u)} \to g_n}\|_2^2 / N_R$ falls below 10^{-15}, which typically occurs when the u-th user is inactive, the probability $\tilde{p}(\mathbf{y}_n | d_n^{(u)})$ can be approximated as a constant. This scenario simplifies the derivation in Eq. (4.4) to a constant. To ensure that this uninformative message can still be represented as a Gaussian distribution, it is modeled as a zero-mean complex Gaussian with a variance of 10^{12} or another large value.

Similarly, when $|h_{\mathbf{x}_n^{(u)} \to g_n}|^2$ is less than or equal to 10^{-15} or another small value, which typically occurs when the u-th user is inactive or when the RE index n corresponds to a known zero-valued pilot data symbol for the u-th user, the probability $\tilde{p}(\mathbf{y}_n | \mathbf{x}_n^{(u)})$ can also be approximated as a constant. This reduction simplifies the derivation in Eq. (4.29) to a constant. To represent this uninformative message as a Gaussian, it is modeled as a zero-mean complex Gaussian with a variance of $10^{12}\mathbf{I}_{N_R}$, where $\mathbf{I}_{N_R}$ is the identity matrix of size N_R.

4.9.2 Complexity analysis

The computational complexity per-user, per-RE and per-iteration of

- each mean vector evaluation in Eqs. (4.10), (4.33) and (4.43) is $\mathcal{O}(N_R)$;
- each covariance matrix evaluation in Eqs. (4.11), (4.34) and (4.44) is $\mathcal{O}(N_R^2)$;
- CE subgraph processing in Sec. 4.6 is $\mathcal{O}(N_R^3)$; due to matrix inversion in Eq. (4.38)
- DEM subgraph processing in Sec. 4.4 is $\mathcal{O}(N_R^3)$ due to matrix inversions in Eqs. (4.12) and (4.13);
- UAD subgraph processing in Sec. 4.7 is $\mathcal{O}(2N_R^3)$ due to matrix inversion in evaluating Eq. (4.46) for each value of the existence variable in $\{0, 1\}$.

As a result, the proposed hybrid GaBP/EP message-passing receiver exhibits the advantageous characteristic that its computational complexity per RE and per iteration increases linearly with U, the maximum number of users. Additionally, when compared to standard BP as described in [21], all processing steps—except for the CE and UAD subgraph—experience a complexity reduction by a factor of Q, which is particularly beneficial for higher-order modulation schemes. This reduction is achieved because EP is used to represent all messages associated with data symbol variables as Gaussian densities rather than discrete probability mass functions, thereby eliminating the need for Q-fold discrete summations during the marginalization of data symbols.

4.9.3 Benchmark algorithm

We propose a reduced-complexity benchmark algorithm that is inspired by the approaches outlined in [89]-[90]. This benchmark algorithm is designed to closely resemble the proposed method in almost all aspects; however, it differs in its approach to UAD. Unlike the proposed method, which employs *a posteriori* PMF-based soft UAD, the benchmark algorithm utilizes a hard UAD approach.

In the hard UAD approach, the decision on user activity is made in a binary fashion—either active or inactive—without considering the soft probabilistic nature of user activity. The key idea behind this method is to base the detection of user activity on the normalized correlation between the received signal that has been cleansed of the estimated MAI and the re-estimated useful signal corresponding to the u-th user. This correlation serves as an indicator of user presence or absence, allowing for a computationally simpler yet potentially less accurate detection method. The mathematical formulation for this normalized correlation is expressed by the following equation,

$$R^{(u)} = \frac{|\sum_{n=0}^{N-1}(\hat{d}_n^{(u)}\mathbf{m}_{\mathbf{x}_n^{(u)}\to g_n})^H\mathbf{z}_n|}{\sqrt{\sum_{n=0}^{N-1}\mathbf{z}_n^H\mathbf{z}_n}}, \tag{4.63}$$

where $\mathbf{z}_n$ is the received signal after cancelling the estimated MAI

$$\mathbf{z}_n = \mathbf{y}_n - \sum_{u'\neq u}\hat{\theta}^{(u')}\hat{d}_n^{(u')}\mathbf{m}_{\mathbf{x}_n^{(u')}\to g_n} \tag{4.64}$$

and $\hat{d}_n^{(u)}$ is the hard decision on the u-th user's symbol over RE n reconstructed from the DEC output. Here hard UAD is obtained by performing hypothesis testing

$$\hat{\theta}^{(u)} = \begin{cases} 1 & \text{if } R^{(u)} > \lambda_t \\ 0 & \text{otherwise,} \end{cases} \tag{4.65}$$

where λ_t is the threshold needed to obtain a constant $P_{fa} \approx 10^{-5}$.

Parameter	Value
Channel coding	$(5/7)_8$ recursive convolutional code
Spreader	rate-$1/4$ repetition code
Modulation	16-QAM
Total number of users (U)	16
Number of active users	12
OFDM subcarriers (N)	1024
Pilot spacing	24
CP size	$N/8$ samples
Channel power delay profile	exponential decay constant=3 taps
Tx antennas	1
Rx antennas (N_R)	4
Rx modeling error parameter (ζ)	15

Table 4.1: System model parameters.

user index	$\theta^{(u)}$	$\rho^{(u)}$	$E_s^{(u)}/N_0$ (dB)
$u \in \{1, 2, 3, 4\}$	0	0	E_s/N_0 (dB)
$u \in \{5, \ldots, 16\}$	1	0	E_s/N_0 (dB)

Table 4.2: Equal receive energy scenario with 12 active users out of $U = 16$.

user index	$\theta^{(u)}$	$\rho^{(u)}$	$E_s^{(u)}/N_0$ (dB)
$u \in \{1, 2, 3, 4\}$	0	0.4	E_s/N_0 (dB)
$u \in \{7, 8, 10, 11, 13, 15\}$	1	0.4	E_s/N_0 (dB)
$u \in \{5, 6, 9, 12, 14, 16\}$	1	0.4	E_s/N_0 (dB) - 6 dB

Table 4.3: Unequal receive energy scenario with 12 active users out of $U = 16$.

4.10 Simulation results

We use the running example of grant-free OFDM-IDMA in Sec. 3.3 using the orthogonal pilots sequences described in Fig. 3.2 to evaluate the performance of the proposed method. The simulation settings in Sec. 4.10.1 include details on the system parameters, such as the number of users, the SNR levels, encoding scheme, modulation scheme, user power levels, and any other relevant configurations that are used to replicate realistic communication scenarios. The choice of these parameters is critical, as they directly influence the validity and generalizability of the results.

Finally, we present the results obtained from the simulations. The results section includes a comprehensive analysis of the algorithm's performance, highlighting key metrics such as accuracy, convergence speed, computational complexity, and robustness under various conditions. The effectiveness of the hybrid EP-BP algorithm is demonstrated through comparisons with other existing methods, and insights are provided on its potential advantages and limitations in practical applications. This section ultimately serves to validate the proposed approach and provides evidence of its efficacy in addressing the challenges posed by multi-user communication systems.

4.10.1 Setup

The parameters of the system model, as detailed in Sec. 3.3, are summarized in Tab. 4.1. These parameters define key aspects of the transmitter configuration, channel characteristics, and other relevant variables essential to the system's operation.

Two different scenarios are considered regarding the signal-to-noise ratio (SNR) conditions for the users. In the first scenario, all users have the same reference SNR, denoted by E_s/N_0 in decibels (dB), which is detailed in Tab. 4.2. This setup assumes uniform energy distribution across all active users, providing a baseline for performance evaluation under equal energy conditions.

In the second scenario, as described in Tab. 4.3, an unequal energy distribution is introduced. Specifically, half of the active users experience a 6 dB SNR penalty relative to the reference SNR E_s/N_0 (dB). This setup is designed to simulate more realistic conditions where not all users have equal transmission power, leading to a mixed SNR environment that can affect system performance and the effectiveness of user activity detection and symbol decoding. The inclusion of both equal and unequal energy scenarios allows for a comprehensive assessment of the system's robustness and adaptability to varying user conditions.

4.10.2 Evolution with respect to the iteration index

Fig. 4.2 and 4.3 illustrate the performance of the proposed hybrid EP/BP algorithm with known hyperparameters as a function of the iteration index at $E_s/N_0 = 4\,$dB. The results indicate that the algorithm converges by the 6th iteration, beyond which any additional improvement is negligible. Consequently, all subsequent simulation results are reported for the 6th iteration, ensuring computational efficiency without sacrificing accuracy.

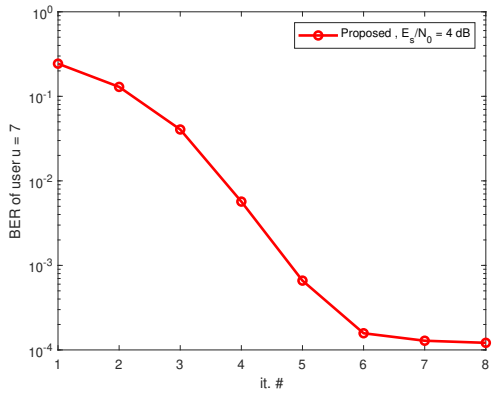

Figure 4.2: BER evolution w.r.t. iteration index at E_s/N_0 = 4 dB.

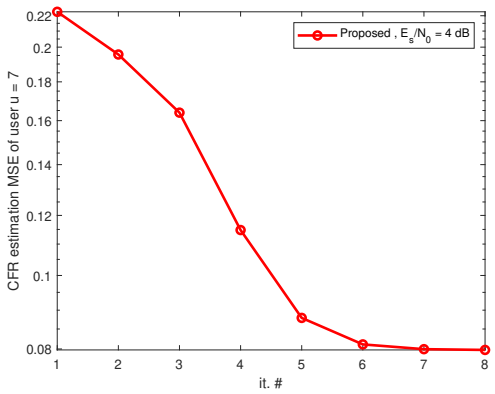

Figure 4.3: BER evolution w.r.t. iteration index at E_s/N_0 = 4 dB.

4.10.3 Comparison with existing methods

We evaluate the proposed method against existing alternatives such as the BP method in [21], state-of-the-art hybrid EP/BP based on scalar auxiliary variables [23, 24] and the correlation-based benchmark in Sec. 4.9.3, under standard conditions of equal energy reception (see Tab. 4.2) and known hyperparameters.

Fig. 4.4 and 4.5 display the Bit-error Rate (BER) and CFR estimation Mean-squared Error (MSE), respectively, for an active user indexed by $u = 7$.

After six iterations, the proposed method outperforms both the less computationally intensive benchmark algorithm and the more complex BP [21]. This indicates that the proposed hybrid GaBP/EP approach achieves a compelling balance between performance and complexity.

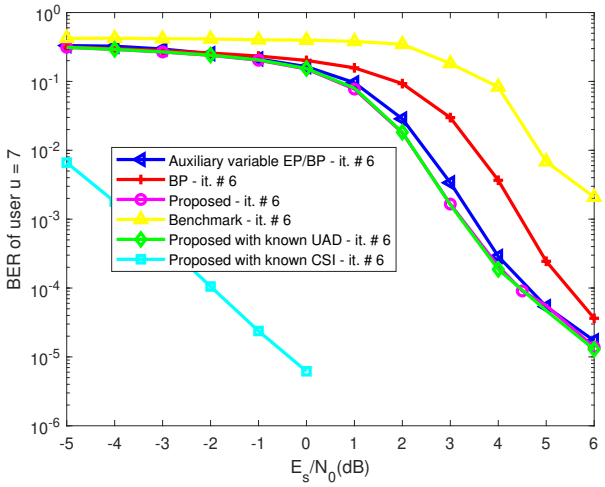

Figure 4.4: BER under equal energy and known hyperparameters.

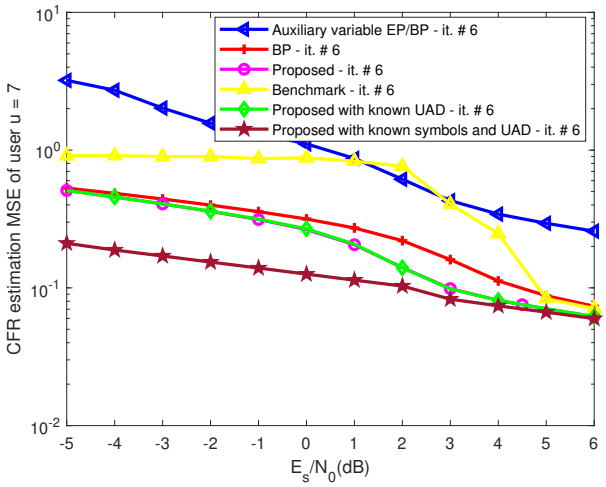

Figure 4.5: CFR estimation MSE under equal energy and known hyperparameters.

Notably, the proposed method maintains performance on par with the known UAD lower bound, without any degradation. Results, which are omitted here for the sake of readability of the figures, suggest that the single-user lower bound with known CSI matches the performance of the proposed method with known CSI. Thus, the 6 dB performance loss relative to known CSI can be seen as a trade-off for fixed-complexity message update rules using Gaussian approximations in higher-order 16-QAM modulation.

It is worth noting that although the auxiliary variable hybrid BP/EP shows a significant gap in channel parameter estimation MSE compared to the proposed method, this only results in a minor BER difference, consistent with previous observations in [91].

UAD performance is assessed in terms of the P_{md} and P_{fa} for active and inactive users, respectively, as illustrated in Fig. 4.6 and Fig. 4.7. We find that the benchmark algorithm achieves the target $P_{fa} \sim 10^{-5}$ but at the cost of high P_{md} across the SNR range. Conversely, while BP shows a modest 1 dB power efficiency improvement in terms of P_{md} over the proposed method, its P_{fa} is consistently worse throughout the SNR range. The auxiliary variable hybrid BP/EP has negligible P_{fa} across the SNR range but suffers from an error floor in P_{md}, which is disadvantageous in a grant-free setting where non-cooperative users cannot be asked to

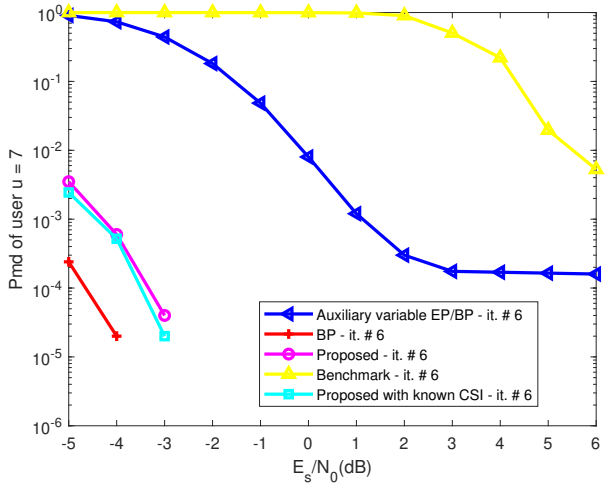

Figure 4.6: P_{md} under equal energy and known hyperparameters.

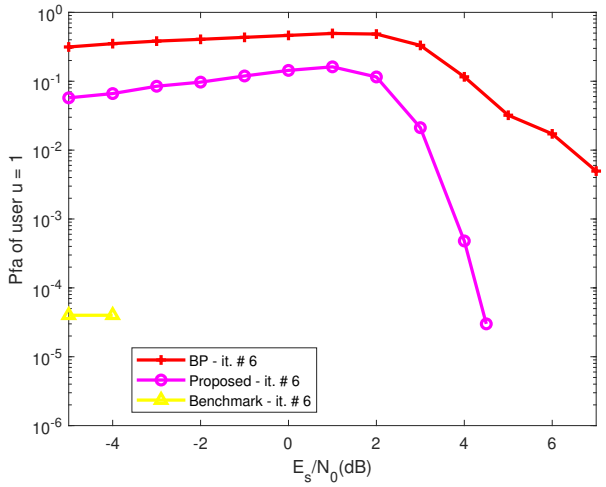

Figure 4.7: P_{fa} under equal energy and known hyperparameters.

retransmit lost packets.

In conclusion, the proposed hybrid GaBP/EP approach offers a superior trade-off between P_{md} and P_{fa} compared to BP, the benchmark algorithm, and the auxiliary variable hybrid BP/EP methods.

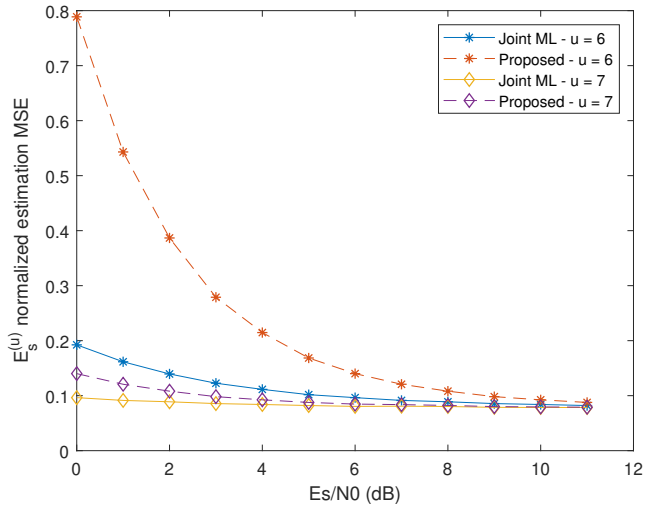

Figure 4.8: Normalized estimation MSE of the symbol energy under unequal energy and unknown hyperparameters - low energy user $u = 6$ and reference energy user $u = 7$.

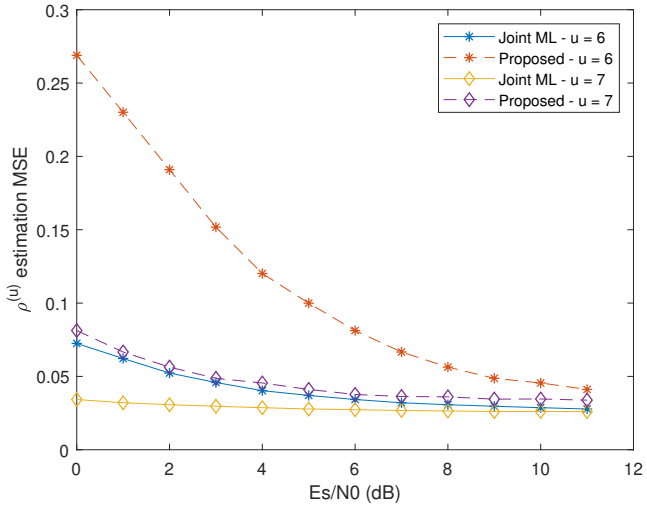

Figure 4.9: Antenna correlation estimation MSE under unequal energy and unknown hyperparameters - low energy user $u = 6$ and reference energy user $u = 7$.

4.10.4 Robustness w.r.t. unknown hyperparameters

Let us now consider the scenario of unequal energy reception (refer to Tab. 4.3) and evaluate the capability of the proposed receiver to estimate unknown hyperparameters dynamically, as user activity may vary from one OFDM block to the next. This is achieved using the approach detailed in Sec. 4.8.

Fig. 4.8 (resp. Fig. 4.9) presents the normalized estimation MSE for symbol energy (resp. for the antenna correlation coefficient). It is important to note that the 6 dB reduction in transmit power for the low-energy user compared to the reference energy user is reflected in the performance curves of the proposed hyperparameter estimation method.

At high SNR, for energy estimation (resp. antenna correlation estimation), the performance degradation is negligible (resp. mild) when compared to joint maximum likelihood (ML) hy-

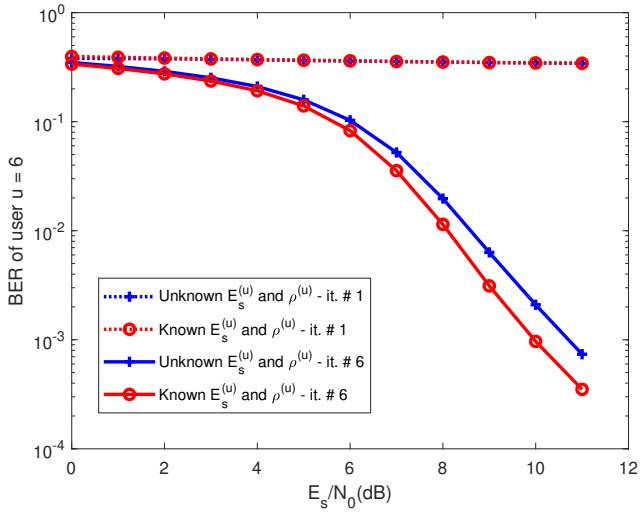

Figure 4.10: BER under unknown hyperparameters - low energy user $u = 6$.

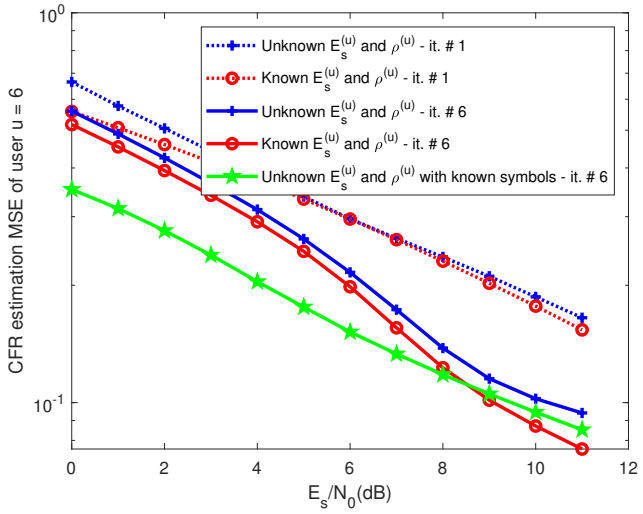

Figure 4.11: CFR estimation MSE under unequal energy and unknown hyperparameters - low energy user $u = 6$.

perparameter estimation, which involves a more complex optimization process, particularly with respect to boundary conditions on the antenna correlation coefficient—implemented using BLEIC [97].

Figs. 4.10–4.12 (resp. Figs. 4.14–4.16) display the performance metrics for a low-energy user (resp. a reference energy user), as well as P_{fa} for an inactive user in Fig. 4.13, for the proposed receiver with hyperparameter estimation.

When compared to perfect hyperparameter knowledge, the BER (resp. P_{md} and P_{fa}) experiences a moderate penalty of less than 1 dB (resp. less than 2 dB, respectively).

This reception scenario, characterized by unequal and unknown hyperparameters, is common in grant-free access systems, where the receiver and users are non-cooperative. Our results clearly demonstrate that the proposed method is robust in such a challenging yet realistic environment, with only a minor complexity increase (due to the simple hyperparameter

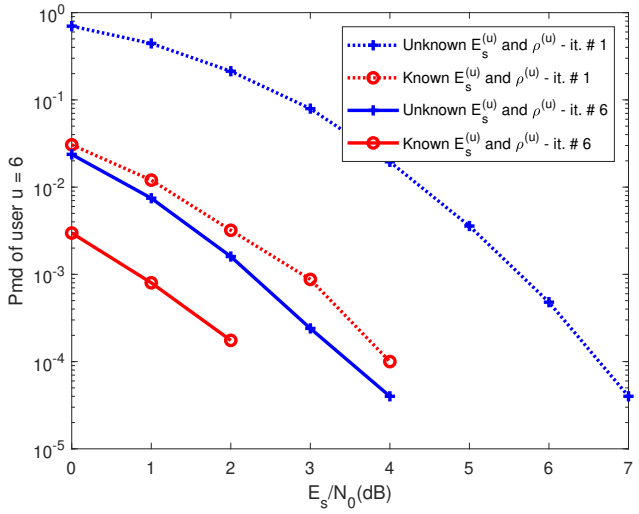

Figure 4.12: P_{md} under unequal energy and unknown hyperparameters - low energy user $u = 6$.

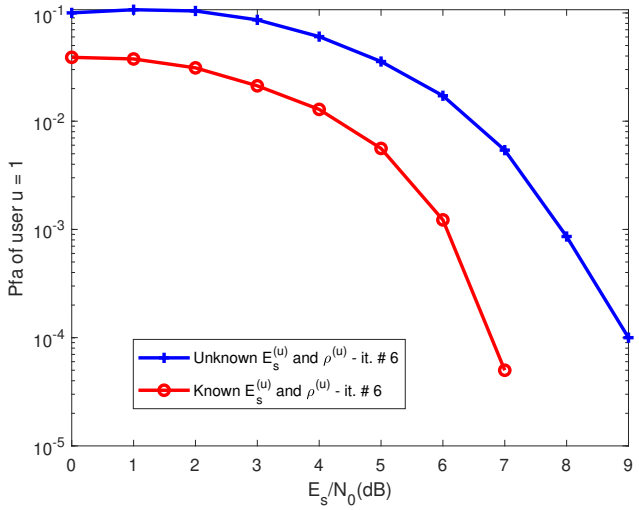

Figure 4.13: P_{fa} under unequal energy and unknown hyperparameters - zero-energy user $u = 1$.

estimation technique in Sec. 4.8) and slight performance losses.

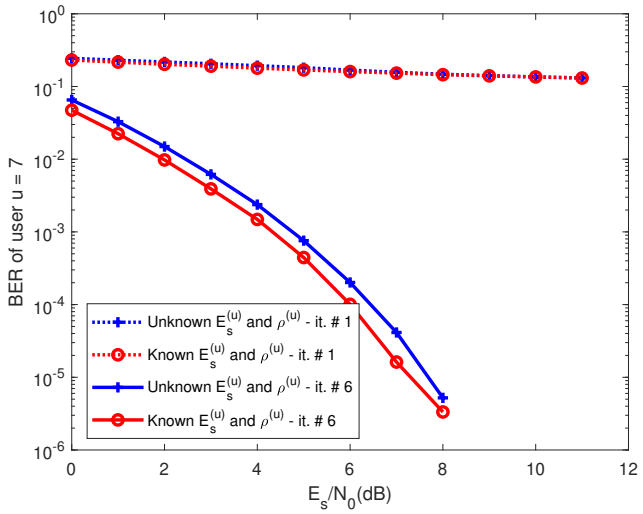

Figure 4.14: BER under unequal energy and unknown hyperparameters - reference energy user $u = 7$.

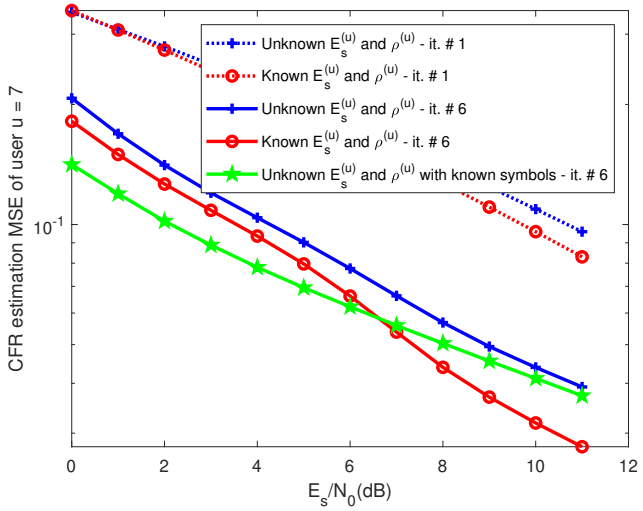

Figure 4.15: CFR estimation MSE under unequal energy and unknown hyperparameters - reference energy user $u = 7$.

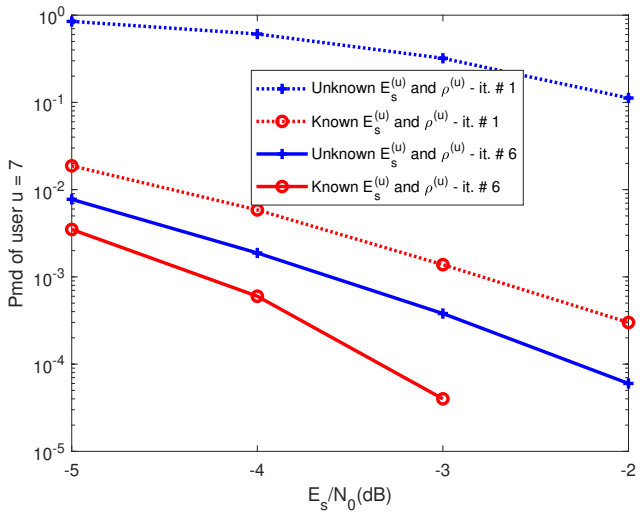

Figure 4.16: P_{md} under unequal energy and unknown hyperparameters - reference energy user $u = 7$.

Chapter 5

Hybrid EP-BP grant-free receiver ignoring antenna correlation

In this chapter, we pursue our efforts towards low-complexity grant-free receiver designs.

A grant-free receiver design based on Gaussian message-passing applicable over the factor graph in Fig. 3.5, suitable for the scalar channel model in Eq. (3.6) and the scalar observation model in Eq. (3.8).

We show that by voluntarily ignoring inter-antenna correlation at the receiver, the methodology of chapter 4 leads to message-passing with univariate Gaussians only, thus leading to substantial complexity gains but at the expense of performance loss.

Unless otherwise specified, in the sequel we will refer to messages exchanged over the graphical model for grant-free NOMA access in Fig. 3.5, derived under the false assumption of zero inter-antenna correlation in Sec. 3.2.2.

5.1 Related work

Many low-complexity receivers operate under the assumption that the channel is spatially uncorrelated at the receiver, despite substantial evidence suggesting that channel correlation is inevitable in 5G networks [26]. In this context, we examine the validity of this assumption and explore the challenges of CE for resource element (RE)-varying multi-antenna channels. This study is of importance for joint CE, MUD, DEM, and DEC operations.

CE using message-passing techniques typically factors the joint distribution of channel coefficients into marginal distributions. This simplification often ignores antenna correlation to reduce complexity (see [28] for an alternative approach). This approach is widely used in multi-antenna NOMA for CE, integrated as part of iterative code-aided receivers.

In [29] and [21], CE is applied within IDMA and Space-division Multiple-access (SDMA) contexts, respectively. These techniques handle quasi-static or frequency-varying channels under known or unknown user activity. Both rely on GaBP [30] to perform CE.

The only other related work addressing time-varying grant-free NOMA is [31]. However, it is limited to single-antenna reception and requires an additional ad-hoc EM procedure to coordinate the CE, MUD, and UAD stages.

5.2 Main contributions

The main contributions in this chapter are summarized as follows:

1. An adaptation of the hybrid EP-BP message-passing algorithm in chapter 4 to a different factor graph corresponding to assumed zero inter-antenna correlation
2. A receiver exchanging only univariate Gaussian messages, thus circumventing the complexity bottleneck of chapter 4 due to the $\mathcal{O}(N_R^3)$ complexity order due to inversions matrices of dimension N_R
3. A robustness study of the new receiver against the false hypothesis of zero inter-antenna correlation is conducted.

Using the projection operator introduced in Sec. 4.3, restricted to projections over the set of univariate circularly symmetric Gaussian densities, we rederive the messages over the DEM, the CE and the UAD subgraphs following a similar approach to that described in the previous chapter.

Additionally, since the DEC subgraph is unchanged, the corresponding EP messages over the set of binomial distributions in Sec. 4.5, remain unchanged.

5.3 Demodulation

Demodulation refers to the process of estimating the data symbols $d_n^{(u)}$ for all users. In this process, two messages play a role: $\mu_{g_{n,r}\to d_n^{(u)}}(d_n^{(u)})$, which represents the message from the observation constraint nodes $g_{n,r}$ to the data symbols $d_n^{(u)}$, and $\mu_{d_n^{(u)}\to g_{n,r}}(d_n^{(u)})$, which is the message from the data symbol $d_n^{(u)}$ to the observation constraint node $g_{n,r}$.

Message from $g_{n,r}$ to $d_n^{(u)}$

The message from observation constraint nodes $g_{n,r}$ to the data symbol $d_n^{(u)}$ is computed using the EP factor node rule Eq. (2.4) to $g_{n,r}$,

$$\mu_{g_{n,r}\to d_n^{(u)}}(d_n^{(u)}) = \frac{\text{proj}_{\Phi}\left(\mu_{d_n^{(u)}\to g_{n,r}}(d_n^{(u)})\tilde{p}(y_{n,r}|d_n^{(u)})\right)}{\mu_{d_n^{(u)}\to g_{n,r}}(d_n^{(u)})}, \tag{5.1}$$

where $\tilde{p}(y_{n,r}|d_n^{(u)})$ is given by the following equation,

$$\begin{aligned}\tilde{p}(y_{n,r}|d_n^{(u)}) = &\int p(y_{n,r}|\{x_{n,r}^{(u)}\}_{u=1}^U, \{d_n^{(u)}\}_{u=1}^U, \{\theta^{(u)}\}_{u=1}^U) \\ &\times \prod_{u'\neq u} \mu_{d_n^{(u')}\to g_{n,r}}(d_n^{(u')}) \prod_{u'=1}^U \mu_{x_{n,r}^{(u')}\to g_{n,r}}(x_{n,r}^{(u')}) \prod_{u'=1}^U \mu_{\theta^{(u')}\to g_{n,r}}(\theta^{(u')}) d \sim d_n^{(u)}.\end{aligned} \tag{5.2}$$

Applying the projection operator in Eq. (4.2) letting $\mathbf{z} = y_{n,r}$, $\boldsymbol{\pi} = d_n^{(u)}$ and $\boldsymbol{\theta} = [\{d_n^{(u')}\}_{u'\neq u}, \{\theta^{(u')}, x_{n,r}^{(u')}\}_{u'=1}^U]$, the continuous Gaussian mixture Eq. (5.2) becomes a Gaussian distribution in $y_{n,r}$ given $d_n^{(u)}$ of the form,

$$\tilde{p}(y_{n,r}|d_n^{(u)}) = \mathcal{CN}(y_{n,r}; m_{y_{n,r}|d_n^{(u)}}(d_n^{(u)}), \sigma^2_{y_{n,r}|d_n^{(u)}}), \tag{5.3}$$

where the mean $m_{y_{n,r}|d_n^{(u)}}(d_n^{(u)})$ and variance $\sigma^2_{y_{n,r}|d_n^{(u)}}$ are given by the following equations,

$$
\begin{aligned}
&m_{y_{n,r}|d_n^{(u)}}(d_n^{(u)}) = h_{d_n^{(u)}\to g_{n,r}} d_n^{(u)} + I_{d_n^{(u)}\to g_{n,r}} \\
&I_{d_n^{(u)}\to g_{n,r}} = \sum_{u'\neq u} m_{\theta^{(u')}\to g_{n,r}} m_{d_n^{(u')}\to g_{n,r}} m_{x_{n,r}^{(u')}\to g_{n,r}},\ h_{d_n^{(u)}\to g_{n,r}} = m_{\theta^{(u)}\to g_{n,r}} m_{x_{n,r}^{(u)}\to g_{n,r}},
\end{aligned} \tag{5.4}
$$

$$
\begin{aligned}
\sigma^2_{y_{n,r}|d_n^{(u)}} &= \left[|m_{\theta^{(u)}\to g_{n,r}}|^2 \sigma^2_{x_{n,r}^{(u)}\to g_{n,r}} + \sigma^2_{\theta^{(u)}\to g_{n,r}}(|m_{x_{n,r}^{(u)}\to g_{n,r}}|^2 + \sigma^2_{x_{n,r}^{(u)}\to g_{n,r}})\right] \\
&+ \sum_{u'\neq u} \sigma^2_{d_n^{(u')}\to g_{n,r}}(|m_{\theta^{(u')}\to g_{n,r}}|^2 + \sigma^2_{\theta^{(u')}\to g_{n,r}})(|m_{x_{n,r}^{(u')}\to g_{n,r}}|^2 + \sigma^2_{x_{n,r}^{(u')}\to g_{n,r}}) \\
&+ \sum_{u'\neq u} |m_{d_n^{(u')}\to g_{n,r}}|^2 \left[|m_{\theta^{(u')}\to g_{n,r}}|^2 \sigma^2_{x_{n,r}^{(u')}\to g_{n,r}} + \sigma^2_{\theta^{(u')}\to g_{n,r}}(|m_{x_{n,r}^{(u')}\to g_{n,r}}|^2 + \sigma^2_{x_{n,r}^{(u')}\to g_{n,r}})\right] + N_0.
\end{aligned} \tag{5.5}
$$

Now, in the Eq. (5.3), if the coeffcient of the hidden variable $d_n^{(u)}$ is factored out, then the message $\mu_{g_{n,r}\to d_n^{(u)}}(d_n^{(u)})$ can be written as a Gaussian distribution of $d_n^{(u)}$ with mean $m_{g_{n,r}\to d_n^{(u)}}$ and variance $\sigma^2_{g_{n,r}\to d_n^{(u)}}$,

$$
\mu_{g_{n,r}\to d_n^{(u)}}(d_n^{(u)}) = \mathcal{CN}(d_n^{(u)}; m_{g_{n,r}\to d_n^{(u)}}, \sigma^2_{g_{n,r}\to d_n^{(u)}}), \tag{5.6}
$$

where the mean $m_{g_{n,r}\to d_n^{(u)}}$ and variance $\sigma^2_{g_{n,r}\to d_n^{(u)}}$ are given by the following equations,

$$
m_{g_{n,r}\to d_n^{(u)}} = \frac{y_{n,r} - I_{d_n^{(u)}\to g_{n,r}}}{h_{d_n^{(u)}\to g_{n,r}}}, \tag{5.7}
$$

$$
\sigma^2_{g_{n,r}\to d_n^{(u)}} = \frac{\sigma^2_{y_{n,r}|d_n^{(u)}}}{|h_{d_n^{(u)}\to g_{n,r}}|^2}. \tag{5.8}
$$

These equations can be interpreted to suggest that, at convergence for an active user, the estimated interference term $I_{d_n^{(u)}\to g_{n,r}}$ in Eq. (5.7) becomes progressively closer to the actual interference present in the system. At the same time, the value of $h_{d_n^{(u)}\to g_{n,r}}$ approaches $x_{n,r}^{(u)}$, where $x_{n,r}^{(u)}$ represents the channel response at the r-th receive antenna. This behavior indicates that the interference and channel parameters are being accurately learned over time. Consequently as the message-passing algorithm converges, $m_{g_{n,r}\to d_n^{(u)}}$ will gradually become a good estimate of $d_n^{(u)}$, while $\sigma^2_{g_{n,r}\to d_n^{(u)}}$ will reflect the reliability of that estimate.

Message from $d_n^{(u)}$ to $g_{n,r}$

The message $\mu_{d_n^{(u)}\to g_{n,r}}(d_n^{(u)})$ is the message from the data symbol to the observation constraint node. The message is computed using the EP variable node rule Eq. (2.5) to the variable node $d_n^{(u)}$,

$$
\mu_{d_n^{(u)}\to g_{n,r}}(d_n^{(u)}) = \frac{\text{proj}_\Phi\left(k \cdot \mu_{\chi_n^{(u)}\to d_n^{(u)}}(d_n^{(u)}) \prod_{r'=1}^{N_R} \mu_{g_{n,r'}\to d_n^{(u)}}(d_n^{(u)})\right)}{\mu_{g_{n,r}\to d_n^{(u)}}(d_n^{(u)})}. \tag{5.9}
$$

Now, $\prod_{r'=1}^{N_R} \mu_{g_{n,r'}\to d_n^{(u)}}(d_n^{(u)})$ inside the projection operator being a product of Gaussian densities, it can be written as $\mathcal{CN}(d_n^{(u)}; m_{g_n\to d_n^{(u)}}, \sigma^2_{g_n\to d_n^{(u)}})$ where the mean $m_{g_n\to d_n^{(u)}}(d_n^{(u)})$ and variance $\sigma^2_{g_n\to d_n^{(u)}}$ are given by the following equations,

$$\begin{aligned}\sigma^{-2}_{g_n\to d_n^{(u)}} &= \sum_{r=1}^{N_R}\sigma^{-2}_{g_{n,r}\to d_n^{(u)}}\\ m_{g_n\to d_n^{(u)}} &= \sigma^2_{g_n\to d_n^{(u)}}\sum_{r=1}^{N_R}\sigma^{-2}_{g_{n,r}\to d_n^{(u)}} m_{g_{n,r}\to d_n^{(u)}}.\end{aligned} \tag{5.10}$$

After rewriting $\mu_{\chi_n^{(u)}\to d_n^{(u)}}(d_n^{(u)})$ in the LLR form at the decoder output (see Eq. (4.28)

$$\begin{aligned}\mu_{d_n^{(u)}\to g_{n,r}}(d_n^{(u)}) &= \frac{\text{proj}_\Phi\left(k\cdot\mathcal{CN}(d_n^{(u)};m_{g_n\to d_n^{(u)}},\sigma^2_{g_n\to d_n^{(u)}})e^{-\sum_{q=1}^{Q} l_{c_{n,q}^{(u)}\to\chi_n^{(u)}}\chi_q^{-1}[d_n^{(u)}]}\right)}{\mu_{g_{n,r}\to d_n^{(u)}}(d_n^{(u)})}\\ &= \frac{\text{proj}_\Phi\left(k\cdot e^{-\frac{|d_n^{(u)}-m_{g_n\to d_n^{(u)}}|^2}{\sigma^2_{g_n\to d_n^{(u)}}}-\sum_{q=1}^{Q} l_{c_{n,q}^{(u)}\to\chi_n^{(u)}}\chi_q^{-1}[d_n^{(u)}]}\right)}{\mu_{g_{n,r}\to d_n^{(u)}}(d_n^{(u)})},\end{aligned} \tag{5.11}$$

In the Eq. (5.11), the numerator of the fraction is a PMF of $d_n^{(u)}$, that can be projected to a Gaussian distribution of the form $\mathcal{CN}(d_n^{(u)};m_{d_n^{(u)}|\mathbf{y}_n},\sigma^2_{d_n^{(u)}|\mathbf{y}_n})$, whose mean and variance are computed in the same way as in Eq. (4.22). Thus,

$$\begin{aligned}\mu_{d_n^{(u)}\to g_{n,r}}(d_n^{(u)}) &= \frac{\mathcal{CN}(d_n^{(u)};m_{d_n^{(u)}|\mathbf{y}_n},\sigma^2_{d_n^{(u)}|\mathbf{y}_n})}{\mathcal{CN}(d_n^{(u)};m_{g_{n,r}\to d_n^{(u)}},\sigma^2_{g_{n,r}\to d_n^{(u)}})}\\ &= \mathcal{CN}(d_n^{(u)};m_{d_n^{(u)}\to g_{n,r}},\sigma^2_{d_n^{(u)}\to g_{n,r}}).\end{aligned} \tag{5.12}$$

Finally, the mean $m_{d_n^{(u)}\to g_{n,r}}$ and variance $\sigma^2_{d_n^{(u)}\to g_{n,r}}$ are given by,

$$\begin{aligned}\sigma^2_{d_n^{(u)}\to g_{n,r}} &= \sigma^{-2}_{d_n^{(u)}|\mathbf{y}_n}-\sigma^{-2}_{g_{n,r}\to d_n^{(u)}}\\ m_{d_n^{(u)}\to g_{n,r}} &= \sigma^2_{d_n^{(u)}\to g_{n,r}}\left(\sigma^{-2}_{d_n^{(u)}|\mathbf{y}_n}m_{d_n^{(u)}|\mathbf{y}_n}-\sigma^{-2}_{g_{n,r}\to d_n^{(u)}}m_{g_{n,r}\to d_n^{(u)}}\right).\end{aligned} \tag{5.13}$$

5.4 Channel estimation

Channel estimation involves estimating the complex channel coefficents $x_{n,r}^{(u)}$ for all users. In the corresponding CE subgraph indexed by receive antenna r in Fig. 3.5, this process entails several key steps: the forward pass, backward pass, and smoothing pass of the Kalman filter. Each of these steps refines the estimation of the complex channel coefficents by leveraging both past and future observations. In addition to these Kalman filtering steps, the process also includes computing the message $\mu_{g_{n,r}\to x_{n,r}^{(u)}}(x_{n,r}^{(u)})$, which propagates information between the the observation constraint node $g_{n,r}$ and the variable node $x_{n,r}^{(u)}$ in the factor graph.

First, the method for computing the message $\mu_{g_{n,r}\to x_{n,r}^{(u)}}(x_{n,r}^{(u)})$ will be described in detail. This message plays a crucial role in updating the estimates of the channel coefficent $x_{n,r}^{(u)}$ based on the most recent observation. Once this message is derived, the operation of the entire CE subgraph will be outlined, integrating the Kalman filtering.

Message from $g_{n,r}$ to $x_{n,r}^{(u)}$

The computation of the message $\mu_{g_{n,r}\to x_{n,r}^{(u)}}(x_{n,r}^{(u)})$ is similar to the computation of the message $\mu_{g_{n,r}\to d_n^{(u)}}(d_n^{(u)})$ computed using the EP factor node rule,

$$\mu_{g_{n,r}\to x_{n,r}^{(u)}}(x_{n,r}^{(u)}) = \frac{\text{proj}_{\Phi}\left(k\cdot \mu_{x_{n,r}^{(u)}\to g_{n,r}}(x_{n,r}^{(u)})\tilde{p}(y_{n,r}|x_{n,r}^{(u)})\right)}{\mu_{x_{n,r}^{(u)}\to g_{n,r}}(x_{n,r}^{(u)})}, \tag{5.14}$$

where $\tilde{p}(y_{n,r}|x_{n,r}^{(u)}$ is given by the following equation,

$$\begin{aligned}\tilde{p}(y_{n,r}|x_{n,r}^{(u)}) &= \int p(y_{n,r}|\{x_{n,r}^{(u)}\}_{u=1}^{U},\{d_n^{(u)}\}_{u=1}^{U},\{\theta^{(u)}\}_{u=1}^{U}) \\ &\times \prod_{u'\neq u}\mu_{x_{n,r}^{(u')}\to g_{n,r}}(x_{n,r}^{(u')})\prod_{u'=1}^{U}\mu_{d_n^{(u')}\to g_{n,r}}(d_n^{(u')})\prod_{u'=1}^{U}\mu_{\theta^{(u')}\to g_{n,r}}(\theta^{(u')})d \sim x_{n,r}^{(u)}.\end{aligned} \tag{5.15}$$

Applying the projection operator in Eq. (4.2) letting $\mathbf{z} = y_{n,r}$, $\boldsymbol{\pi} = x_{n,r}^{(u)}$ and $\boldsymbol{\theta} = [\{x_{n,r}^{(u')}\}_{u'\neq u},\{\theta^{(u')},d_n^{(u')}\}_{u'=1}^{U}]$, this continuous Gaussian mixture becomes a Gaussian distribution in $y_{n,r}$ given $x_{n,r}^{(u)}$ of the form,

$$\tilde{p}(y_{n,r}|x_{n,r}^{(u)}) = \mathcal{CN}(y_{n,r};m_{y_{n,r}|x_{n,r}^{(u)}}(x_{n,r}^{(u)}),\sigma^2_{y_{n,r}|x_{n,r}^{(u)}}), \tag{5.16}$$

where the mean $m_{y_{n,r}|x_{n,r}^{(u)}}(x_{n,r}^{(u)})$ and variance $\sigma^2_{y_{n,r}|x_{n,r}^{(u)}}$ are given by the following equations,

$$m_{y_{n,r}|x_{n,r}^{(u)}}(x_{n,r}^{(u)}) = h_{x_{n,r}^{(u)}\to g_{n,r}}x_{n,r}^{(u)} + I_{d_n^{(u)}\to g_{n,r}}, \quad h_{x_{n,r}^{(u)}\to g_{n,r}} = m_{\theta^{(u)}\to g_n}m_{d_n^{(u)}\to g_n}, \tag{5.17}$$

$$\begin{aligned}\sigma^2_{y_{n,r}|x_{n,r}^{(u)}} &= E_s^{(u)}\left[|m_{\theta^{(u)}\to g_{n,r}}|^2\sigma^2_{d_n^{(u)}\to g_{n,r}} + \sigma^2_{\theta^{(u)}\to g_{n,r}}(|m_{d_n^{(u)}\to g_{n,r}}|^2+\sigma^2_{d_n^{(u)}\to g_{n,r}})\right] \\ &+ \sum_{u'\neq u}\sigma^2_{x_{n,r}^{(u')}\to g_{n,r}}(|m_{\theta^{(u')}\to g_{n,r}}|^2+\sigma^2_{\theta^{(u')}\to g_{n,r}})(|m_{d_n^{(u')}\to g_{n,r}}|^2+\sigma^2_{d_n^{(u')}\to g_{n,r}}) \\ &+ \sum_{u'\neq u}\left[|m_{\theta^{(u')}\to g_{n,r}}|^2\sigma^2_{d_n^{(u')}\to g_{n,r}} + \sigma^2_{\theta^{(u')}\to g_{n,r}}(|m_{d_n^{(u')}\to g_{n,r}}|^2+\sigma^2_{d_n^{(u')}\to g_{n,r}})\right]|m_{x_{n,r}^{(u')}\to g_{n,r}}|^2 + N_0.\end{aligned} \tag{5.18}$$

By factoring out $h_{x_{n,r}^{(u)}\to g_{n,r}}$ in Eq. (5.16), the Eq. (5.14) becomes a Gaussian distribution in $x_{n,r}^{(u)}$ with mean $m_{g_{n,r}\to x_{n,r}^{(u)}}$ and variance $\sigma^2_{g_{n,r}\to x_{n,r}^{(u)}}$,

$$\mu_{g_{n,r}\to x_{n,r}^{(u)}} = \mathcal{CN}(x_{n,r}^{(u)};m_{g_{n,r}\to x_{n,r}^{(u)}},\sigma^2_{g_{n,r}\to x_{n,r}^{(u)}}), \tag{5.19}$$

where the mean $m_{g_{n,r}\to x_{n,r}^{(u)}}$ and variance $\sigma^2_{g_{n,r}\to x_{n,r}^{(u)}}$ are given by the following equations,

$$m_{g_{n,r}\to x_{n,r}^{(u)}} = \frac{y_{n,r} - I_{d_n^{(u)}\to g_{n,r}}}{h_{x_{n,r}^{(u)}\to g_{n,r}}}, \tag{5.20}$$

$$\sigma^2_{g_{n,r}\to x_{n,r}^{(u)}} = \frac{\sigma^2_{y_{n,r}|x_{n,r}^{(u)}}}{|h_{x_{n,r}^{(u)}\to g_{n,r}}|^2}. \tag{5.21}$$

These equations can be interpreted to suggest that, at convergence for an active user, the estimated interference term $I_{d_n^{(u)}\to g_{n,r}}$ in Eq. (5.20) becomes progressively closer to the actual interference present in the system. At the same time, the value of $h_{x_{n,r}^{(u)}\to g_{n,r}}$ approaches $d_n^{(u)}$, where $d_n^{(u)}$ represents the data symbol. This behavior indicates that the interference and

channel parameters are being accurately learned over time. Consequently as the message-passing algorithm converges, $m_{g_{n,r}\to x_{n,r}^{(u)}}$ will gradually become a good estimate of $x_{n,r}^{(u)}$, while $\sigma^2_{g_{n,r}\to x_{n,r}^{(u)}}$ will reflect the reliability of that estimate.

Messages inside the CE subgraph

The forward and backward passes in the CE subgraph for EP follow the same procedure as in BP, as described by the following equations:

$$\begin{aligned}\mu_{f_{n,r}^{(u)}\to x_{n,r}^{(u)}}(x_{n,r}^{(u)}) &\propto \mathcal{CN}(x_{n,r}^{(u)} : m_{f_{n,r}^{(u)}\to x_{n,r}^{(u)}}, \sigma^2_{f_{n,r}^{(u)}\to x_{n,r}^{(u)}}),\\ \mu_{f_{n+1,r}^{(u)}\to x_{n,r}^{(u)}}(x_{n,r}^{(u)}) &\propto \mathcal{CN}(x_{n,r}^{(u)} : m_{f_{n+1,r}^{(u)}\to x_{n,r}^{(u)}}, \sigma^2_{f_{n+1,r}^{(u)}\to x_{n,r}^{(u)}}),\end{aligned} \tag{5.22}$$

where the mean and variance are updated similarly to Kalman filtering and smoothing (details are omitted here, but interested readers can refer to [76, Fig. 15], [86]).

Once the forward and backward passes of the CE subgraph are completed, the extrinsic smoothing pass computes the message $\mu_{x_{n,r}^{(u)}\to g_n}(x_{n,r}^{(u)})$, obtained by applying the EP variable node rule in Eq. (2.5)

$$\mu_{x_{n,r}^{(u)}\to g_{n,r}}(x_{n,r}^{(u)}) = \mu_{f_{n,r}^{(u)}\to x_{n,r}^{(u)}}(x_{n,r}^{(u)})\mu_{f_{n+1,r}^{(u)}\to x_{n,r}^{(u)}}(x_{n,r}^{(u)}) \propto \mathcal{CN}(x_{n,r}^{(u)}; m_{x_{n,r}^{(u)}\to g_{n,r}}, \sigma^2_{x_{n,r}^{(u)}\to g_{n,r}}), \tag{5.23}$$

with the mean and variance calculated as:

$$\begin{aligned}{\sigma^2_{x_{n,r}^{(u)}\to g_{n,r}}}^{-1} &= {\sigma^2_{f_{n,r}^{(u)}\to x_{n,r}^{(u)}}}^{-1} + {\sigma^2_{f_{n+1,r}^{(u)}\to x_{n,r}^{(u)}}}^{-1},\\ {\sigma^2_{x_{n,r}^{(u)}\to g_{n,r}}}^{-1} m_{x_{n,r}^{(u)}\to g_{n,r}} &= {\sigma^2_{f_{n,r}^{(u)}\to x_{n,r}^{(u)}}}^{-1} m_{f_{n,r}^{(u)}\to x_{n,r}^{(u)}} + {\sigma^2_{f_{n+1,r}^{(u)}\to x_{n,r}^{(u)}}}^{-1} m_{f_{n+1,r}^{(u)}\to x_{n,r}^{(u)}}.\end{aligned} \tag{5.24}$$

5.5 User activity detection

UAD for the scalarized CFR model is similar to the vectorized CFR model employing BP instead of EP. The reason of using BP for UAD is same as explained in the previous chapter. There are two types of messages for UAD, message from an observation constraint node to a user activity variable $\mu_{g_{n,r}\to\theta^{(u)}}(\theta^{(u)})$ and in reverse direction $\mu_{\theta^{(u)}\to_{g_{n,r}}}(\theta^{(u)})$.

Message from $g_{n,r}$ to $\theta^{(u)}$

The message $\mu_{g_{n,r}\to\theta^{(u)}}(\theta^{(u)})$ is computed using the BP factor node rule in Eq. (2.2) to $g_{n,r}$,

$$\mu_{g_{n,r}\to\theta^{(u)}}(\theta^{(u)}) = k\cdot\tilde{p}(y_{n,r}|\theta^{(u)}), \tag{5.25}$$

where the conditional pdf $\tilde{p}(y_{n,r}|\theta^{(u)})$ is given by the following equation,

$$\begin{aligned}\tilde{p}(y_{n,r}|\theta^{(u)}) = \int p(y_{n,r}|\{x_{n,r}^{(u)}\}_{u=1}^{U}, \{d_n^{(u)}\}_{u=1}^{U}, \{\theta^{(u)}\}_{u=1}^{U}) \prod_{u'\neq u}\mu_{\theta^{(u')}\to g_n}(\theta_{u'})\\ \times\prod_{u'=1}^{U}\mu_{x_{n,r}^{(u')}\to g_n}(x_{n,r}^{(u')})\prod_{u'=1}^{U}\mu_{d_n^{(u')}\to g_n}(d_n^{(u')})d \sim \theta^{(u)}.\end{aligned} \tag{5.26}$$

This continuous Gaussian mixture can be resolved to a single Gaussian distribution of $y_{n,r}$,

$$\tilde{q}(y_{n,r}|\theta^{(u)}) = \mathcal{CN}(y_{n,r}; m_{y_{n,r}|\theta^{(u)}}(\theta^{(u)}), \sigma^2_{y_{n,r}|\theta^{(u)}}(\theta^{(u)})) \tag{5.27}$$

by minimizing $KL(\tilde{p}||\tilde{q})$.

It follows that the mean $m_{y_{n,r}|\theta^{(u)}}(\theta^{(u)})$ and covariance $\sigma^2_{y_{n,r}|\theta^{(u)}}(\theta^{(u)})$ are given by the following equations,

$$\begin{aligned} m_{y_{n,r}|\theta^{(u)}}(\theta^{(u)}) &= h_{\theta^{(u)}\to g_{n,r}}\theta^{(u)} + I_{d_n^{(u)}\to g_{n,r}} \\ h_{\theta^{(u)}\to g_{n,r}} &= m_{d_n^{(u)}\to g_{n,r}} m_{x_{n,r}^{(u)}\to g_{n,r}}, \quad I_{d_n^{(u)}\to g_{n,r}} = \sum_{u'\neq u} m_{d_n^{(u')}\to g_{n,r}} m_{x_{n,r}^{(u')}\to g_{n,r}} m_{\theta^{(u')}\to g_{n,r}}, \end{aligned} \tag{5.28}$$

$$\begin{aligned} \sigma^2_{y_{n,r}|\theta^{(u)}}(\theta^{(u)}) &= \left[|m_{d_n^{(u)}\to g_{n,r}}|^2\sigma^2_{x_{n,r}^{(u)}\to g_{n,r}} + \sigma^2_{d_n^{(u)}\to g_{n,r}}(|m_{x_{n,r}^{(u)}\to g_{n,r}}|^2 + \sigma^2_{x_{n,r}^{(u)}\to g_{n,r}})\right]|\theta^{(u)}|^2 \\ &+ \sum_{u'\neq u} \sigma^2_{\theta^{(u')}\to g_{n,r}}(|m_{d_n^{(u')}\to g_{n,r}}|^2 + \sigma^2_{d_n^{(u')}\to g_{n,r}})(|m_{x_{n,r}^{(u')}\to g_{n,r}}|^2 + \sigma^2_{x_{n,r}^{(u')}\to g_{n,r}}) \\ &+ \sum_{u'\neq u} |m_{\theta^{(u')}\to g_{n,r}}|^2\left[|m_{d_n^{(u')}\to g_{n,r}}|^2\sigma^2_{x_{n,r}^{(u')}\to g_{n,r}} + \sigma^2_{d_n^{(u')}\to g_{n,r}}(|m_{x_{n,r}^{(u')}\to g_{n,r}}|^2 + \sigma^2_{x_{n,r}^{(u')}\to g_{n,r}})\right] + N_0. \end{aligned} \tag{5.29}$$

As the message $\mu_{g_{n,r}\to\theta^{(u)}}(\theta^{(u)})$ is a PMF of $\theta^{(u)}$, with support $\theta^{(u)} \in \{0, 1\}$, we can compute the LLR of this message,

$$l_{g_{n,r}\to\theta^{(u)}} = \ln \frac{\mu_{g_{n,r}\to\theta^{(u)}}(\theta^{(u)})_{|\theta^{(u)}=0}}{\mu_{g_{n,r}\to\theta^{(u)}}(\theta^{(u)})_{|\theta^{(u)}=1}}. \tag{5.30}$$

After subsituting the values of the message $\mu_{g_{n,r}\to\theta^{(u)}}(\theta^{(u)})$ at $\theta^{(u)} = 0$ and $\theta^{(u)} = 1$, the LLR in Eq. (5.30) is given by the following equation,

$$l_{g_{n,r}\to\theta^{(u)}} = \frac{|y_{n,r} - m_{y_{n,r}|\theta^{(u)}}(1)|^2}{\sigma^2_{y_{n,r}|\theta^{(u)}}(1)} + \ln\sigma^2_{y_{n,r}|\theta^{(u)}}(1) - \left(\frac{|y_{n,r} - m_{y_{n,r}|\theta^{(u)}}(0)|^2}{\sigma^2_{y_{n,r}|\theta^{(u)}}(0)} + \ln\sigma^2_{y_{n,r}|\theta^{(u)}}(0)\right). \tag{5.31}$$

Interpretation: The *a posteriori* LLR of $\theta^{(u)}$, denoted as $l_{g_{n,r}\to\theta^{(u)}}$, needs to be interpreted differently for both active and inactive users. To achieve this, we must analyze what happens to the mean $m_{y_{n,r}|\theta^{(u)}}(\theta^{(u)})$ and variance $\sigma^2_{y_{n,r}|\theta^{(u)}}(\theta^{(u)})$ as the algorithm converges.

For an active user, the term $h_{\theta^{(u)}\to g_{n,r}}$ approaches the true value of the transmitted signal $d_n^{(u)}x_{n,r}^{(u)}$, while the interference term $I_{d_n^{(u)}\to g_{n,r}}$ converges towards the actual MAI. This implies that, as the algorithm converges, the residual between the received signal and the expected signal when $\theta^{(u)}$ is hypothesized as 1 approximates the noise term, i.e.,

$$y_{n,r} - m_{y_{n,r}|\theta^{(u)}}(1) \approx w_{n,r}, \tag{5.32}$$

where $w_{n,r}$ represents the noise. Similarly, when when $\theta^{(u)}$ is hypothesized as 0, we have:

$$y_{n,r} - m_{y_{n,r}|\theta^{(u)}}(0) \approx d_n^{(u)}x_{n,r}^{(u)} + w_{n,r}. \tag{5.33}$$

Both variances, $\sigma^2_{y_{n,r}|\theta^{(u)}}(1)$ and $\sigma^2_{y_{n,r}|\theta^{(u)}}(0)$, converge to the noise variance N_0. Consequently, at high SNR, the LLR $l_{g_{n,r}\to\theta^{(u)}}$ becomes negative, as expressed in the following equation:

$$l_{g_{n,r}\to\theta^{(u)}} \approx \frac{|w_{n,r}|^2}{N_0} - \frac{|d_n^{(u)}x_{n,r}^{(u)} + w_{n,r}|^2}{N_0}. \tag{5.34}$$

Here, the term $|d_n^{(u)}x_{n,r}^{(u)} + w_{n,r}|^2$ is greater than $|w_{n,r}|^2$ in the positive SNR range because the signal component $d_n^{(u)}x_{n,r}^{(u)}$ dominates over the noise.

For an inactive user, irrespective of the iteration index $m_{d_n^{(u)}\to g_{n,r}} m_{x_{n,r}^{(u)}\to g_{n,r}}, \sigma^2_{d_n^{(u)}\to g_{n,r}}$ and $\sigma^2_{x_{n,r}^{(u)}\to g_{n,r}}$ stay close to their initializations at their *a priori* values. Consequently, $h_{\theta^{(u)}\to g_{n,r}}$ remains close to zero. For the same reason the first line in Eq. (5.29) remains close to $E_s^{(u)}|\theta^{(u)}|^2$.

As a result, $m_{y_{n,r}|\theta^{(u)}}(\theta^{(u)})$ approaches $I_{d_n^{(u)}\to g_{n,r}}$ both when $\theta^{(u)}$ is hypothesized as 0 or 1. At convergence, the interference term $I_{d_n^{(u)}\to g_{n,r}}$ approximates the true MAI, making the residual between the received signal and the expected signal, $y_{n,r} - m_{y_{n,r}|\theta^{(u)}}(\theta^{(u)})$, approximately equal to $w_{n,r}$ for both cases.

Meanwhile, the variance $\sigma^2_{y_{n,r}|\theta^{(u)}}(1)$ approaches $E_s^{(u)} + N_0$, where $E_s^{(u)}$ is the signal power, and $\sigma^2_{y_{n,r}|\theta^{(u)}}(0)$ approaches N_0. Therefore, the LLR $l_{g_{n,r}\to\theta^{(u)}}$ for the inactive user becomes:

$$l_{g_{n,r}\to\theta^{(u)}} \approx \frac{|w_{n,r}|^2}{E_s^{(u)} + N_0} + \ln(E_s^{(u)} + N_0) - \left(\frac{|w_{n,r}|^2}{N_0} + \ln(N_0)\right). \tag{5.35}$$

For the user to be estimated as inactive, the LLR must be positive.

$$\ln\left(1 + \frac{E_s^{(u)}}{N_0}\right)) > \frac{|w_{n,r}|^2}{N_0}\left(1 - \frac{1}{1 + \frac{E_s^{(u)}}{N_0}}\right). \tag{5.36}$$

This inequality holds primarily for moderate and high SNR values.

Message from $\theta^{(u)}$ to $g_{n,r}$

The computation of the message $\mu_{\theta^{(u)}\to g_{n,r}}$ follows the BP variable node rule (2.3) applied to $\theta^{(u)}$ and can be expressed as:

$$\mu_{\theta^{(u)}\to g_{n,r}}(\theta^{(u)}) = k \cdot p(\theta^{(u)}) \prod_{n'=0,r'=1|(n',r')\neq(n,r)}^{N-1,N_R} \mu_{g_{n',r'}\to\theta^{(u)}}(\theta^{(u)}), \tag{5.37}$$

where $p(\theta^{(u)})$ represents the PMF of $\theta^{(u)}$, and k is a normalization constant. The product spans all REs (resp. antenna indices) except the current one, n (resp., r). Since the messages $\mu_{g'_{n,r}\to\theta^{(u)}}(\theta^{(u)})$ are proportional to $e^{-l_{g_{n,r}\to\theta^{(u)}}\theta^{(u)}}$, the message equation can be simplified as:

$$\mu_{\theta^{(u)}\to g_{n,r}}(\theta^{(u)}) = k \cdot p(\theta^{(u)}) \prod_{n'=0,r'=1|(n',r')\neq(n,r)}^{N-1,N_R} e^{-l_{g_{n',r'}\to\theta^{(u)}}\theta^{(u)}}, \tag{5.38}$$

where $l_{g_{n,r}\to\theta^{(u)}}$ is the LLR of the message from $g_{n,r}$ to $\theta^{(u)}$. The normalization constant k is given by:

$$k = \frac{1}{\sum_{\theta^{(u)}=0}^{1} p(\theta^{(u)}) e^{-\sum_{n'=0,r'=1|(n',r')\neq(n,r)}^{N-1,N_R} l_{g_{n',r'}\to\theta^{(u)}}\theta^{(u)}}}. \tag{5.39}$$

The expressions for the mean $m_{\theta^{(u)}\to g_{n,r}}$ and variance $\sigma^2_{\theta^{(u)}\to g_{n,r}}$ are as follows:

$$m_{\theta^{(u)}\to g_{n,r}} = \frac{p_a^{(u)} e^{-\sum_{n'=0,r'=1|(n',r')\neq(n,r)}^{N-1,N_R} l_{g_{n',r'}\to\theta^{(u)}}}}{1 - p_a^{(u)} + p_a^{(u)} e^{-\sum_{n'=0,r'=1|(n',r')\neq(n,r)}^{N-1,N_R-1} l_{g_{n',r'}\to\theta^{(u)}}}}, \tag{5.40}$$

$$\sigma^2_{\theta^{(u)}\to g_{n',r'}} = m_{\theta^{(u)}\to g_{n',r'}}(1 - m_{\theta^{(u)}\to g_{n',r'}}), \tag{5.41}$$

where $p_a^{(u)}$ denotes the prior probability of user u being active. The a posteriori PMF of $\theta^{(u)}$ can be computed from the LLRs $l_{g_{n,r}\to\theta^{(u)}}$ using the following equation:

$$l_{\theta^{(u)}|\{y_{n,r}\}_{n=0,r=1}^{N-1,N_R}} = \ln \frac{p(\theta^{(u)}|\{y_{n,r}\}_{n=0,r=1}^{N-1,N_R})_{|\theta^{(u)}=0}}{p(\theta^{(u)}|\{y_{n,r}\}_{n=0,r=1}^{N-1,N_R})_{|\theta^{(u)}=1}} = \ln(1-p_a^{(u)}) - \ln(p_a^{(u)}) + \sum_{n=0}^{N-1}\sum_{r=1}^{N_R} l_{g_{n,r}\to\theta^{(u)}}. \tag{5.42}$$

It follows that hard UAD is obtained by performing hypothesis testing

$$\hat{\theta}^{(u)} = \begin{cases} 1 & \text{if } l_{\theta^{(u)}|\{y_{n,r}\}_{n=0,r=1}^{N-1,N_R}} < 0 \\ 0 & \text{otherwise.} \end{cases} \tag{5.43}$$

5.6 Receiver implementation

In this section, we present the detailed implementation of the receiver. The implementation process is broken down into several key components to ensure clarity and a comprehensive understanding of the steps involved.

We begin by discussing the initialization of the messages, which plays a critical role in the message-passing algorithm. This section outlines how the initial beliefs and parameters are set.

Next, we outline the message-passing schedule, which governs how messages are exchanged in the factor graph. This ensures efficient information propagation, improving convergence speed and reducing computational costs.

Finally, a complexity analysis shows the potential of computational speed-up by ignoring antenna correlation at the receiver wrt the previous chapter.

5.6.1 Initialization

To initiate the scalarized hybrid EP/BP, it is crucial to initialize certain messages properly to ensure smooth convergence. Proper initialization not only facilitates faster convergence but also enhances the stability of the algorithm during iterations.

In the DEM subgraph, for a data (resp. a pilot) RE the message $\mu_{d_n^{(u)}\to g_{n,r}}(d_n^{(u)})$ is initialized as complex Gaussian distribution with mean zero (resp. equal to the known pilot symbol) and unit (resp. zero) variance. Similarly, the LLRs at the decoder output $l_{c_{n,q}^{(u)}\to\chi_n^{(u)}}(c_{n,q}^{(u)})$ are initialized to zero. This ensures a uniform probability distribution over the values of $c_{n,q}^{(u)}$, implying equal likelihoods for $c_{n,q}^{(u)} = 0$ or $c_{n,q}^{(u)} = 1$.

In the CE subgraph, the messages $\mu_{x_{n,r}^{(u)}\to g_{n,r}}(x_{n,r}^{(u)})$ are initialized with zero mean and variance $E_s^{(u)}$. This is denoted as $\mu_{x_{n,r}^{(u)}\to g_{n,r}}(x_{n,r}^{(u)}) \sim \mathcal{CN}(x_{n,r}^{(u)}; 0, E_s^{(u)})$.

For the UAD subgraph, the messages $\mu_{\theta^{(u)}\to g_{n,r}}(\theta^{(u)})$ are initialized with a unit mean and zero variance, expressed as $\mu_{\theta^{(u)}\to g_{n,r}}(\theta^{(u)}) \sim \mathcal{CN}(\theta^{(u)}; 1, 0)$, which is tantamount to forcing the existence of all users before the start of message-passing.

5.6.2 Message-passing schedule

We process all U user subgraphs sequentially using the following serial schedule:

1. For the current user u, we reset the message $\mu_{\theta^{(u)} \to g_{n,r}}(\theta^{(u)}) = \mathcal{CN}(\theta^{(u)}; 1, 0)$. This step significantly improves the probability of missed detection (P_{md}) and false alarm (P_{fa}).
2. Process the CE subgraph by following the steps described in Sec. 5.4.
3. Process the DEM subgraph by following the steps described in Sec. 5.3.
4. Perform DEC subgraph processing, which is the same as in Sec. 4.5.
5. Process the UAD subgraph by following the steps described in Sec. 5.5.

The reasons for choosing this message-passing schedule are same as described in the previous chapter.

5.6.3 Complexity analysis

In this section, we compare the proposed method, based on the scalarized hybrid EP/BP, with the vectorized hybrid EP/BP presented in the previous chapter. For ease of reference in the rest of this chapter, the hybrid EP/BP method exploiting (resp. ignoring) antenna correlation in the previous (resp. current) chapter will be referred to as the vectorized (resp. scalarized) algorithm or method.

The vectorized method performs joint estimation of the channel vector $\mathbf{x}_n^{(u)}$ across all receive antennas. This approach accounts for antenna correlation, which can improve accuracy in environments where such correlation exists. However, this comes at the cost of increased computational complexity. Specifically, the joint estimation requires matrix operations that take into consideration the covariance across the antennas. These operations involve matrix inversions, which are computationally expensive and have a complexity of $\mathcal{O}(N_R^3)$, where N_R is the number of receive antennas.

In contrast, the scalarized method simplifies the estimation process by treating each component of the channel vector $\mathbf{x}_n^{(u)}$ independently for each receive antenna. Instead of performing joint estimation for all antennas, the channel gain for each antenna is estimated separately, ignoring any potential antenna correlation. This reduces the computational load, as the method no longer requires matrix inversion operations, which are the primary contributors to the high complexity of the vectorized method.

However, it is important to note that when the correlation between antennas is high, the approximate low-complexity scalarized method proposed in this chapter is in general suboptimal.

The comparative complexity analysis is summarized in Tab. 5.1. It can be seen that the key advantage of the approach advocated in this chapter is its significant reduction in computational complexity. By scalarizing the messages, the scalarized method computes messages for each antenna individually, lowering the complexity from $\mathcal{O}(N_R^3)$ in the vectorized method to $\mathcal{O}(N_R)$ in the scalarized method. This simplification is especially beneficial in scenarios where the number of receive antennas is large, as it results in a substantial decrease in the overall computational cost.

Table 5.1: Per iteration and per user complexity order of hybrid EP/BP.

Task	CE	DEM	UAD
Scalarized method in this chapter	$\mathcal{O}(N_R)$	$\mathcal{O}(N_R Q)$	$\mathcal{O}(N_R)$
Vectorized method in chapter 4	$\mathcal{O}(N_R^3)$	$\mathcal{O}(N_R^3 Q)$	$\mathcal{O}(N_R^3)$

5.7 Simulation results

We use the running example of grant-free OFDM-IDMA in Sec. 3.3 using the orthogonal pilots sequences described in Fig. 3.2 to evaluate the performance of the proposed method.

For the sake of fair comparison with the method accounting for antenna correlation in the previous chapter, the system model parameters are identical (see Tab. 4.1) under standard conditions of equal energy reception (see Tab. 4.2, except that $\rho^{(u)}$ can be non-zero) and known hyperparameters. In this setup, recall that there are $U = 16$ users, with the last 12 users active and the first 4 users inactive. All other model parameters, including channel coding, the number of transmit and receive antennas, 16-QAM modulation, and pilot spacing, remain unchanged.

5.7.1 Performance evolution as a function of SNR at fixed $\rho^{(u)}$

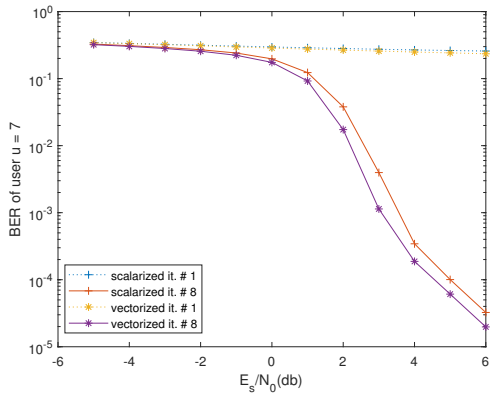

Figure 5.1: BER comparison of the scalarized and vectorized algorithm for $\rho^{(u)} = 0.4$.

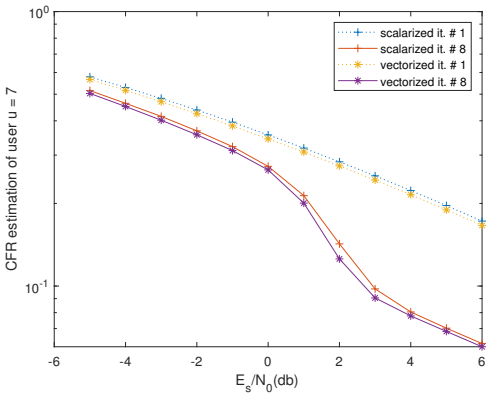

Figure 5.2: CFR MSE comparison of the scalarized and vectorized algorithm for $\rho^{(u)} = 0.4$.

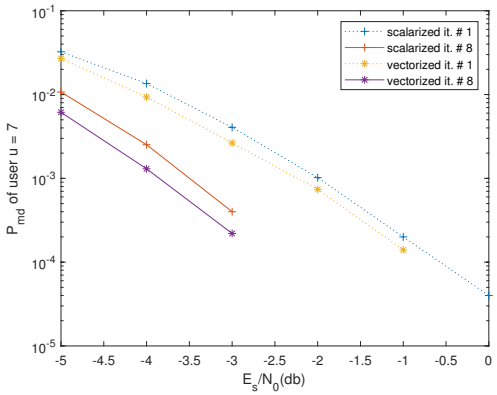

Figure 5.3: P_{md} comparison of the scalarized and vectorized algorithm for $\rho^{(u)} = 0.4$.

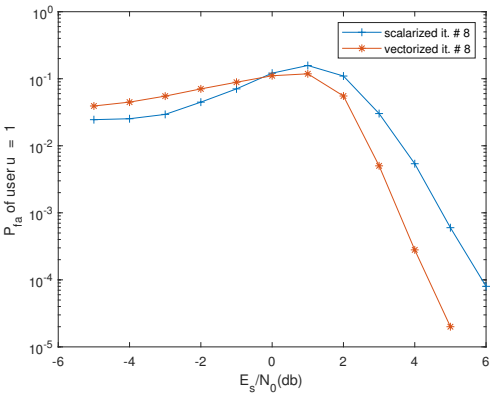

Figure 5.4: P_{fa} of the scalarized and vectorized algorithm for $\rho^{(u)} = 0.4$.

Let us begin with the example where all users have equal antenna correlation $\rho^{(u)} = \rho = 0.4$, for $u = 1, \ldots, U$ as originally suggested in Tab. 4.1. Fig. 5.1 and 5.2 show that the BER

and CFR estimation MSE for both the vectorized and scalarized methods are nearly identical. Regarding the performance comparison in terms of UAD, both methods exhibit comparable P_{md} in Fig. 5.3. However, Fig 5.4 shows that P_{fa} of the scalarized suffers from a 1 dB power efficiency loss wrt the vectorized algorithm at high SNR.

Overall, the results show that the scalarized method performs nearly as well as the more complex vectorized method, for this particular low antenna correlation scenario. Even with a moderate suboptimality in terms of P_{fa}, the scalarized method remains a practical alternative, balancing accuracy and computational efficiency.

5.7.2 Performance evolution as a function of $\rho^{(u)}$ at fixed SNR

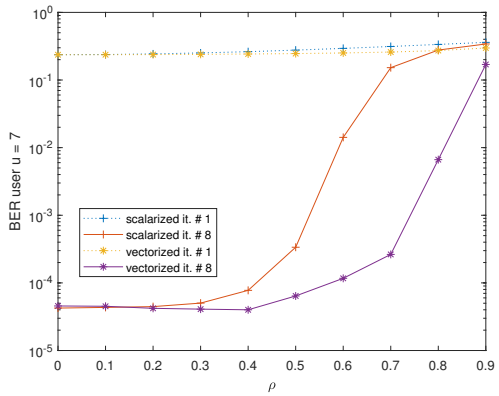

Figure 5.5: BER comparison of the scalarized and vectorized algorithm at $E_s/N_0 = 5$ dB.

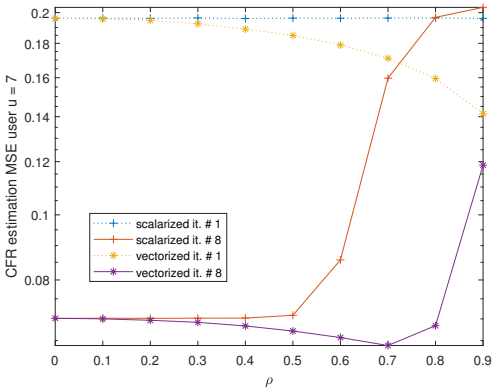

Figure 5.6: CFR MSE comparison of the scalarized and vectorized algorithm at $E_s/N_0 = 5$ dB.

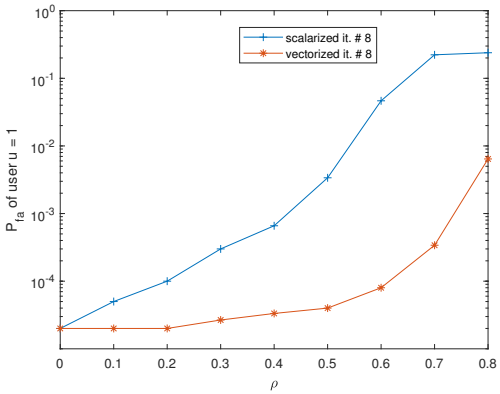

Figure 5.7: P_{fa} comparison of the scalarized and vectorized algorithm at $E_s/N_0 = 5$ dB.

We wish to exemplify the behavior of the considered performance metrics with increasing values of $\rho^{(u)} = \rho$ (assumed identical for users indexed by $u = 1, \ldots, U$ for simplicity). For this purpose, we select a fixed but high enough SNR so that the BER is typically lower than 10^{-4} when $\rho = 0$, i.e. $E_s/N_0 = 5$ dB. Therefore, we plot the performances curves for the BER, the CFR estimation MSE and P_{fa} in Fig. 5.5 to 5.7. (Note that P_{md} need not be plotted, as it is typically vanishing, i.e. lower than 10^{-4}, at $E_s/N_0 = 5$ dB).

Note that all performance metrics are identical at $\rho = 0$ after the 8-th iteration. Indeed, from a theoretic point of view all independence assumptions underlying the factorizations of the

posterior distribution of all hidden variables in Sec. 3.2.2 leading to the factor graph in Fig. 3.5, become exact.
Consequently, hybrid EP/BP performed by the vectorized method over the factor graph in Fig. 3.4 coincides with hybrid EP/BP performed by the scalarized method over the factor graph in Fig. 3.5.

Also, the vectorized and the scalarized method suffer from performance degradation wrt all metrics with increasing ρ due to progressive loss of receive antenna diversity.
Moreover, when $\rho > 0$, the scalarized method works under the false assumption of zero inter-antenna correlation, which make it suboptimal wrt the vectorized algorithm.
While the suboptimality is not very pronounced as long as $\rho < 0.3$ to 0.4, a threshold effect occurs around $\rho = 0.4$, meaning that for higher values of antenna correlation there is a surge in the performance gap between both methods.

5.7.3 Performance evolution as a function of $(E_s/N_0, \rho^{(u)})$

In order to validate the finding of Sec. 5.7.2, extensive simulations are performed accross the SNR range $-5 \leq E_s/N_0\ (dB) \leq 6$, with the antenna correlation factor $\rho^{(u)} = \rho$ varied between 0.0 and 0.7 to observe its impact on performance. The corresponding surface plots comparing the performances for all metrics under consideration are presented in Fig. 5.8 to 5.11.

As expected, the performances of the vectorized method accounting for antenna correlation are uniformly better than the performances of the scalarized algorithm for all metrics over the entire range of SNR and ρ. Again, this fact can be explained by the fact that the false assumption of zero antenna correlation in the scalarized method induces suboptimality whenever $\rho \neq 0$.

In summary, the scalarized hybrid EP/BP receiver is a viable solution with a good performance vs. complexity tradeoff for environments with low to moderate antenna correlation (i.e. $\rho < 0.4$). However, in highly correlated scenarios, the vectorized hybrid EP/BP method is essential to ensure satisfactory performance.

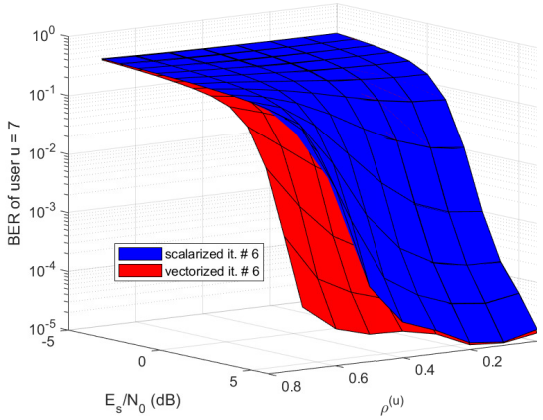

Figure 5.8: BER of the scalarized and vectorized algorithm.

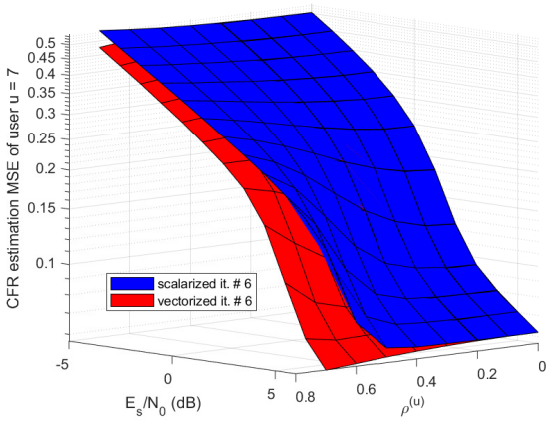

Figure 5.9: CFR estimation MSE of the scalarized and vectorized algorithm.

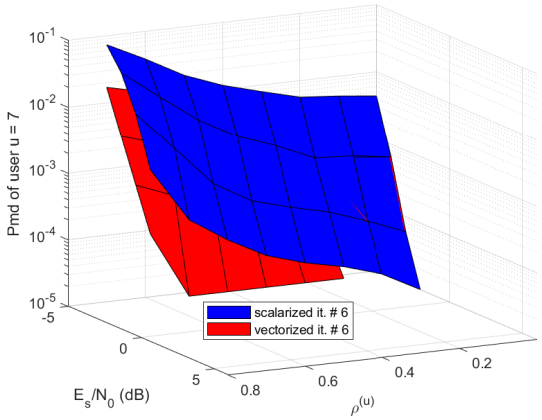

Figure 5.10: P_{md} of the scalarized and vectorized algorithm.

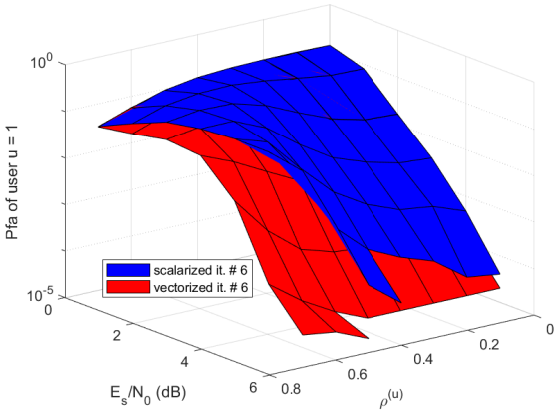

Figure 5.11: P_{fa} of the scalarized and vectorized algorithm.

Chapter 6

EP grant-free receiver based on a Wirtinger calculus Taylor series approximation

This chapter introduces a novel grant-free receiver design based on EP message-passing applicable over the factor graph in Fig. 3.4 suitable for the vector channel model in Eq. (3.5). This design is motivated by the fact that EP is the proper mathematical tool for assumed-density message passing. While this was made possible using the projection operator introduced in chapter 4 for demodulation and channel estimation, this method completely failed to produce a consistent EP rule for user activity detection. The present chapter introduces a remedy to this issue.

To be specific, we propose a novel approach for UAD by employing a Taylor series-based approximation of the infinite Gaussian mixture that emerges during the application of EP.

A key aspect of this approximation is the use of Wirtinger calculus [92], a tool specifically designed for handling functions of complex variables. Wirtinger calculus provides a systematic way to differentiate and manipulate complex-valued functions, which is particularly beneficial in communication systems where signals and variables are often complex in nature. Leveraging Wirtinger calculus to obtain a Taylor series approximation of functions of the complex variable that are not differentiable in the usual sense, we are able to approximate reliably a continuous mixture of complex Gaussians as a single complex Gaussian distribution.

Unless otherwise stated, the EP steps introduced in chapter 4 for the sake of for DEM in Sec. 4.4, CE in Sec. 4.6 and DEC in Sec. 4.5 remain the same. Therefore in a sense, the receiver designed in this chapter as compared to the the one in chapter 4 can be seen

- from a theoretical perspective as a algorithm unified under the umbrella of the EP framework (i.e. dispensing with hybridization with BP, which was considered previously in chapter 4 for UAD)
- as an improved version with better performances.

Also, in this chapter we will consider the messages exchanged over the graphical model for grant-free NOMA access in Fig. 3.4.

6.1 Related work

While a similar concept was presented in [93] within the context of massive MIMO OFDM-based systems, our derivation is broader and more versatile. It does not impose limitations on

scalar observables or rely on the implicit assumption of i.i.d. hidden variables. Additionally, our approach is more straightforward, as it avoids dependence on two extra intermediate variables beyond the primary latent variable of interest.

6.2 Main contributions

The main contributions in this chapter are summarized as follows:

1. Leveraging a Wirtinger calculus-based second-order expansion, a new approximation to continuous mixtures of complex Gaussians as a single complex Gaussian distribution is proposed.
2. The new approximation is used for the sake of EP message-passing implementing UAD, giving rise to a complete receiver that never leaves the EP framework
3. The new receiver is shown to reach better performances than competing state-of-the-art ccounterparts for OFDM-IDMA with inter-antenna correlation.

6.3 General results on Wirtinger calculus

Wirtinger calculus [94] is a mathematical tool used for handling functions of complex variables, especially when dealing with functions that are not holomorphic (i.e., not complex differentiable in the usual sense). It simplifies the differentiation process by treating the real and imaginary parts of complex variables independently, which is particularly useful in optimization problems and signal processing applications.

6.3.1 Wirtinger derivatives

In Wirtinger calculus, a complex variable $z = x + jy$ (where x is the real part and y is the imaginary part) is treated using two independent variables: z and its conjugate $z^* = x - jy$. This approach allows the differentiation of a function $f(z)$ with respect to z and z^* separately, using partial derivatives $\frac{\partial}{\partial z}$ and $\frac{\partial}{\partial z^*}$.

The key formulas in Wirtinger calculus are:

- The Wirtinger derivative with respect to z:

$$\frac{\partial}{\partial z} = \frac{1}{2}\left(\frac{\partial}{\partial x} - j\frac{\partial}{\partial y}\right)$$

- The Wirtinger derivative with respect to $\overline{z}$:

$$\frac{\partial}{\partial z^*} = \frac{1}{2}\left(\frac{\partial}{\partial x} + j\frac{\partial}{\partial y}\right).$$

These derivatives allow for efficient manipulation and analysis of complex-valued functions, particularly in scenarios where traditional complex differentiation does not apply. Wirtinger calculus is widely used in fields like machine learning, communications, and control theory, where complex variables play a significant role.

6.3.2 Wirtinger calculus based Taylor series

The Taylor series is a mathematical tool used to approximate a function around a particular point by expanding it into an infinite series. For a real-valued function $f(x)$, where x is a real variable, the Taylor series expansion about a point x_0 is given as:

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n, \tag{6.1}$$

where $f^{(n)}(x_0)$ represents the n-th derivative of the function evaluated at x_0. This series is particularly useful for approximating functions locally around x_0.

Similarly, this concept extends to functions of a complex variable. Consider a function $f(z) = u(x, y) + jv(x, y)$ of the complex variable $z = x + jy$. If u and v have continuous first partial derivatives and satisfy the Cauchy-Riemann equations

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \quad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}, \tag{6.2}$$

then is f is holomorphic [92] and a Taylor series around a point z_0 can be written as:

$$f(z) = \sum_{n=0}^{\infty} \frac{f^{(n)}(z_0)}{n!}(z - z_0)^n. \tag{6.3}$$

For the general case of a non-holomorphic function, a Taylor series around a point z_0 can still be obtained as [92, Eq. (99)]:

$$\begin{aligned} f(z) \approx f(z_0) + \Re\Bigg(& 2\frac{\partial f(z)}{\partial z}\Big|_{z=z_0}(z - z_0) + \frac{\partial}{\partial z}\left(\frac{\partial f(z)}{\partial z}\right)^*\Big|_{z=z_0}|z - z_0|^2 \\ & + \frac{\partial}{\partial z^*}\left(\frac{\partial f(z)}{\partial z}\right)^*\Big|_{z=z_0}(z^* - z_0^*)^2\Bigg) + \text{h.o.t.} \end{aligned} \tag{6.4}$$

Ignoring higher order terms (h.o.t) in in Eq. (6.4), the resulting second-order Taylor series approximation will be instrumental for the sake of projecting a function of a complex latent variable to a circularly symmetric Gaussian as is done in EP message-passing.

6.4 Novel EP-based user activity detection

For simplicity, since as already mentioned DEM and CE perform EP message-passing in the set Φ corresponding to the family of circularly symmetric complex Gaussian densities, we wish do to the same for UAD in order to perform integration, multiplication and division when applying the EP factor node rule Eq. (2.4) and the EP variable node rule Eq. (2.5) without leaving Φ. Alternatively, one could for instance attempt to take advantage of the fact that user activity variables are real-valued (resp. defined over $[0, 1]$), thus alternating projections over real Gaussian densities (resp. beta distributions) and complex Gaussian densities when performing UAD (resp. CE and DEM), but we leave this a a subject of future research since it would complicate the derivations.

6.4.1 EP Message from g_n to $\theta^{(u)}$

The message $\mu_{g_n \to \theta^{(u)}}$ is computed by applying the EP factor node rule (2.4) to g_n and is given by the following equation,

$$\mu_{g_n \to \theta^{(u)}}(\theta^{(u)}) = \frac{\text{proj}_{\Phi}\left(k \cdot \mu_{\theta^{(u)} \to g_n}(\theta^{(u)})\tilde{p}(\mathbf{y}_n|\theta^{(u)})\right)}{\mu_{\theta^{(u)} \to g_n}(\theta^{(u)})}, \tag{6.5}$$

where $\tilde{p}(\mathbf{y}_n|\theta^{(u)})$ is the continuous Gaussian mixture already defined in (4.40) and projected to the complex Gaussian,

$$\tilde{q}(\mathbf{y}_n|\theta^{(u)}) = \mathcal{CN}(\mathbf{y}_n; \mathbf{m}_{\mathbf{y}_n|\theta^{(u)}}(\theta^{(u)}), \mathbf{\Sigma}_{\mathbf{y}_n|\theta^{(u)}}(\theta^{(u)})), \tag{6.6}$$

where the mean vector $\mathbf{m}_{\mathbf{y}_n|\theta^{(u)}}(\theta^{(u)})$ in Eq. (4.43) and covariance matrix $\mathbf{\Sigma}_{\mathbf{y}_n|\theta^{(u)}}(\theta^{(u)})$ in Eq. (4.44) can be rewritten using shorthand notations,

$$\mathbf{m}_{\mathbf{y}_n|\theta^{(u)}}(\theta^{(u)}) = \mathbf{h}_{\theta^{(u)}\to g_n}\theta^{(u)} + \mathbf{I}_{d_n^{(u)}\to g_n} \tag{6.7}$$

$$\mathbf{\Sigma}_{\mathbf{y}_n|\theta^{(u)}}(\theta^{(u)}) = |\theta^{(u)}|^2\mathbf{A}_n^{(u)} + \mathbf{B}_n^{(u)}. \tag{6.8}$$

In Eqs. (6.7) and (6.8), the parameters are given by the following equations,

$$\begin{aligned}
&\mathbf{h}_{\theta^{(u)}\to g_n} = m_{d_n^{(u)}\to g_n}\mathbf{m}_{\mathbf{x}_n^{(u)}\to g_n},\ \mathbf{I}_{d_n^{(u)}\to g_n} = \sum_{u'\neq u} m_{\theta^{(u')}\to g_n} m_{d_n^{(u')}\to g_n}\mathbf{m}_{\mathbf{x}_n^{(u')}\to g_n}\\
&\mathbf{A}_n^{(u)} = |m_{d_n^{(u)}\to g_n}|^2\mathbf{\Sigma}_{\mathbf{x}_n^{(u)}\to g_n} + \sigma^2_{d_n^{(u)}\to g_n}(\mathbf{m}_{\mathbf{x}_n^{(u)}\to g_n}\mathbf{m}^H_{\mathbf{x}_n^{(u)}\to g_n} + \mathbf{\Sigma}_{\mathbf{x}_n^{(u)}\to g_n})\\
&\mathbf{B}_n^{(u)} = \sum_{u'\neq u}\sigma^2_{\theta^{(u')}\to g_n}(|m_{d_n^{(u')}\to g_n}|^2 + \sigma^2_{d_n^{(u')}\to g_n})(\mathbf{m}_{\mathbf{x}_n^{(u')}\to g_n}\mathbf{m}^H_{\mathbf{x}_n^{(u')}\to g_n} + \mathbf{\Sigma}_{\mathbf{x}_n^{(u')}\to g_n})\\
&+ \sum_{u'\neq u}|m_{\theta^{(u')}\to g_n}|^2\left[|m_{d_n^{(u')}\to g_n}|^2\mathbf{\Sigma}_{\mathbf{x}_n^{(u')}\to g_n} + \sigma^2_{d_n^{(u')}\to g_n}(\mathbf{m}_{\mathbf{x}_n^{(u')}\to g_n}\mathbf{m}^H_{\mathbf{x}_n^{(u')}\to g_n} + \mathbf{\Sigma}_{\mathbf{x}_n^{(u')}\to g_n})\right]\\
&\quad + \mathbf{R}.
\end{aligned} \tag{6.9}$$

The first term in Eq. (6.7) gives the expected signal for the u-th user, conditioned on $\theta^{(u)}$ while the second term accounts for the expected inter-user interference affecting the u-th user.

The first term in Eq. (6.8) captures the uncertainty associated with the hidden variables of the u-th user, given $\theta^{(u)}$, while the second term account for the uncertainty introduced by both inter-user interference and noise.

Applying the Wirtinger calculus-based second-order Taylor series approximation in Eq. (6.4) to the logarithm of the argument of the projection operator in Eq. (6.5) about a point θ_0 that plays the role of a design parameter, the corresponding message takes the form $\mu_{g_n\to\theta^{(u)}}(\theta^{(u)}) = \mathcal{CN}(\theta^{(u)}; m_{g_n\to\theta^{(u)}}, \sigma^2_{g_n\to\theta^{(u)}})$, whose mean and variance are expressed as follows:

$$
\begin{aligned}
\frac{1}{\sigma^2_{g_n\to\theta^{(u)}}} &= \operatorname{trace}\left\{(|\theta_0|^2\mathbf{A}_n^{(u)}+\mathbf{B}_n^{(u)})^{-1}\mathbf{A}_n^{(u)}(|\theta_0|^2\mathbf{A}_n^{(u)}+\mathbf{B}_n^{(u)})^{-1}\mathbf{B}_n^{(u)}\right\}\\
&\quad+\mathbf{h}^H_{\theta^{(u)}\to g_n}(|\theta_0|^2\mathbf{A}_n^{(u)}+\mathbf{B}_n^{(u)})^{-1}\mathbf{B}_n^{(u)}(|\theta_0|^2\mathbf{A}_n^{(u)}+\mathbf{B}_n^{(u)})^{-1}\mathbf{h}_{\theta^{(u)}\to g_n}\\
&\quad+(\mathbf{y}_n-\mathbf{I}_{d_n^{(u)}\to g_n})^H(|\theta_0|^2\mathbf{A}_n^{(u)}+\mathbf{B}_n^{(u)})^{-1}\mathbf{A}_n^{(u)}(|\theta_0|^2\mathbf{A}_n^{(u)}+\mathbf{B}_n^{(u)})^{-1}(\mathbf{y}_n-\mathbf{I}_{d_n^{(u)}\to g_n})\\
&\quad-2(\mathbf{y}_n-\mathbf{I}_{d_n^{(u)}\to g_n}-\mathbf{h}_{\theta^{(u)}\to g_n}\theta_0)^H(|\theta_0|^2\mathbf{A}_n^{(u)}+\mathbf{B}_n^{(u)})^{-1}\mathbf{A}_n^{(u)}(|\theta_0|^2\mathbf{A}_n^{(u)}+\mathbf{B}_n^{(u)})^{-1}\\
&\quad\times\mathbf{B}_n^{(u)}(|\theta_0|^2\mathbf{A}_n^{(u)}+\mathbf{B}_n^{(u)})^{-1}(\mathbf{y}_n-\mathbf{I}_{d_n^{(u)}\to g_n}-\mathbf{h}_{\theta^{(u)}\to g_n}\theta_0)\\
\frac{m_{g_n\to\theta^{(u)}}}{\sigma^2_{g_n\to\theta^{(u)}}} &= \mathbf{h}^H_{\theta^{(u)}\to g_n}(|\theta_0|^2\mathbf{A}_n^{(u)}+\mathbf{B}_n^{(u)})^{-1}(\mathbf{y}_n-\mathbf{I}_{d_n^{(u)}\to g_n}-\mathbf{h}_{\theta^{(u)}\to g_n}\theta_0)\\
&\quad+\theta_0\Bigg(\frac{1}{\sigma^2_{g_n\to\theta^{(u)}}}-\operatorname{trace}\left\{\mathbf{A}_n^{(u)}(|\theta_0|^2\mathbf{A}_n^{(u)}+\mathbf{B}_n^{(u)})^{-1}\right\}\\
&\qquad+(\mathbf{y}_n-\mathbf{I}_{d_n^{(u)}\to g_n}-\mathbf{h}_{\theta^{(u)}\to g_n}\theta_0)^H(|\theta_0|^2\mathbf{A}_n^{(u)}+\mathbf{B}_n^{(u)})^{-1}\\
&\qquad\times\mathbf{A}_n^{(u)}(|\theta_0|^2\mathbf{A}_n^{(u)}+\mathbf{B}_n^{(u)})^{-1}(\mathbf{y}_n-\mathbf{I}_{d_n^{(u)}\to g_n}-\mathbf{h}_{\theta^{(u)}\to g_n}\theta_0)\Bigg).
\end{aligned}
\tag{6.10}
$$

The proof is postponed to Appendix B.

Although there is a variety of possible choices for the design parameter θ_0, the discussion below indicates that selecting 1 at initialization and the *a posteriori* mean of $\theta^{(u)}$, $m_{\theta^{(u)}|\mathbf{y}}$ from the previous iteration later on, is a sensible choice.

Interpretation: The mean $m_{g_n\to\theta^{(u)}}$ and variance $\sigma^2_{g_n\to\theta^{(u)}}$ in Eq. (6.10) has to be interpreted for both active and inactive users.

For an active user, at convergence $\mathbf{h}_{\theta^{(u)}\to g_n}$ approaches true value of the signal $d_n^{(u)}\mathbf{x}_n^{(u)}$ and the estimated MAI $\mathbf{I}_{d_n^{(u)}\to g_n}$ approaches the true MAI. Moreover, $\mathbf{A}_n^{(u)}$ approaches the all-zero matrix and $\mathbf{B}_n^{(u)}$ approaches $\mathbf{R}=N_0\mathbf{I}_{N_R}$. Thus the variance in Eq. (6.10) approaches,

$$
\sigma^2_{g_n\to\theta^{(u)}}\approx\frac{N_0}{|d_n^{(u)}|^2||\mathbf{x}_n^{(u)}||_2^2},
\tag{6.11}
$$

which decreases with the SNR, $E_s^{(u)}/N_0$ (dB) as $\mathbf{x}_n^{(u)}$ depends on the SNR.

Assuming a sensible choice for the design parameter θ_0 is the *a posteriori* mean at previous iteration, so in case of an active user it would typically be $\theta_0=1$. Consequently, the mean $m_{g_n\to\theta^{(u)}}$ would typically approach

$$
m_{g_n\to\theta^{(u)}}\approx\theta_0+\underbrace{\frac{\mathbf{x}_n^{(u)H}\mathbf{w}_n}{d_n^{(u)}\mathbf{x}_n^{(u)H}\mathbf{x}_n^{(u)}}}_{noise\,term}.
\tag{6.12}
$$

i.e. 1 plus an noise term decreasing with the SNR.

For an inactive user, at convergence $\mathbf{h}_{\theta^{(u)}\to g_n}\approx 0$, $\mathbf{I}_{d_n^{(u)}\to g_n}$ approaches true MAI, $\mathbf{A}_n^{(u)}\approx E_s^{(u)}\mathbf{\Gamma}^{(u)}$ and $\mathbf{B}_n^{(u)}\approx N_0\mathbf{I}_{N_R}$.

Again, selecting θ_0 as the *a posteriori* mean at previous iteration, in case of an inactive user typically $\theta_0=0$ at convergence. Also for simplicity consider the particular case where $\mathbf{\Gamma}^{(u)}=\mathbf{I}_{N_R}$, then at convergence

$$\sigma^2_{g_n\to\theta^{(u)}} \approx \frac{1}{N_R\frac{E_s^{(u)}}{N_0}(1-\frac{||\mathbf{w}_n||_2^2}{N_R N_0})}, \tag{6.13}$$

which decreases with the SNR.

On the other hand, the mean $m_{g_n\to\theta^{(u)}}$ approaches zero implying that the belief that the u-th user is inactive increases.

6.4.2 Message form $\theta^{(u)}$ to g_n

Applying the EP variable node rule Eq. (2.5) to $\theta^{(u)}$

$$\mu_{\theta^{(u)}\to g_n}(\theta^{(u)}) = \frac{\text{proj}_\Phi\Big(p(\theta^{(u)}|\mathbf{y})\Big)}{\mu_{g_n\to\theta^{(u)}}(\theta^{(u)})}. \tag{6.14}$$

The argument of the projection operator is the *a posteriori* pdf of $\theta^{(u)}$ so that

$$p(\theta^{(u)}|\mathbf{y}) = k\cdot p(\theta^{(u)})\prod_{n'=0}^{N-1}\mu_{g_{n'}\to\theta^{(u)}}(\theta^{(u)}), \tag{6.15}$$

where k is the normalization constant. Remind that the prior pdf $p(\theta^{(u)})$ and $\mu_{g_{n'}\to\theta^{(u)}}$ have the form,

$$p(\theta^{(u)}) = (1-p_a^{(u)})\delta(\theta^{(u)}) + p_a^{(u)}\delta(\theta^{(u)}-1) \tag{6.16}$$

$$\mu_{g_{n'}\to\theta^{(u)}}(\theta^{(u)}) \propto \mathcal{CN}(\theta^{(u)}; m_{g_{n'}\to\theta^{(u)}}, \sigma^2_{g_{n'}\to\theta^{(u)}}), \tag{6.17}$$

where in Eq. (6.16) $\delta(.)$ is the Dirac delta function. The product of the messages $\mu_{g_{n'}\to\theta^{(u)}}(\theta^{(u)})$ in Eq. (6.15) is also a Gaussian distribution such that,

$$\prod_{n'=0}^{N-1}\mu_{g_{n'}\to\theta^{(u)}}(\theta^{(u)}) = \mathcal{CN}(\theta^{(u)}; \mathring{m}_{\theta^{(u)}}, \mathring{\sigma}^2_{\theta^{(u)}}), \tag{6.18}$$

where the mean $\mathring{m}_{\theta^{(u)}}$, and variance $\mathring{\sigma}^2_{\theta^{(u)}}$ are given by the following equations,

$$\mathring{\sigma}^{-2}_{\theta^{(u)}} = \sum_{n'=0}^{N-1}\sigma^{-2}_{g'_n\to\theta^{(u)}} \tag{6.19}$$

$$\mathring{m}_{\theta^{(u)}} = \mathring{\sigma}^2_{\theta^{(u)}}\sum_{n'=0}^{N-1}\sigma^{-2}_{g_{n'}\to\theta^{(u)}} m_{g_{n'}\to\theta^{(u)}}. \tag{6.20}$$

The normalization constant k in Eq. (6.15) is given by the following equation,

$$k = \frac{1}{(1-p_a^{(u)})\mathcal{CN}(0;\mathring{m}_{\theta^{(u)}},\mathring{\sigma}^2_{\theta^{(u)}}) + p_a^{(u)}\mathcal{CN}(1;\mathring{m}_{\theta^{(u)}},\mathring{\sigma}^2_{\theta^{(u)}})}. \tag{6.21}$$

It follows that, the mean and variance of *of the a posteriori PMF* of $\theta^{(u)}$, i.e. $p(\theta^{(u)}|\mathbf{y})$ are given by the following equations,

$$\begin{aligned} m_{\theta^{(u)}|\mathbf{y}} &= \frac{p_a^{(u)}\mathcal{CN}(1;\mathring{m}_{\theta^{(u)}},\mathring{\sigma}^2_{\theta^{(u)}})}{(1-p_a^{(u)})\mathcal{CN}(0;\mathring{m}_{\theta^{(u)}},\mathring{\sigma}^2_{\theta^{(u)}}) + p_a^{(u)}\mathcal{CN}(1;\mathring{m}_{\theta^{(u)}},\mathring{\sigma}^2_{\theta^{(u)}})} \\ &= \frac{1}{\frac{1-p_a^{(u)}}{p_a^{(u)}}e^{\frac{1-2\Re\left\{\mathring{m}_{\theta^{(u)}}\right\}}{\mathring{\sigma}^2_{\theta^{(u)}}}}+1}, \end{aligned} \tag{6.22}$$

$$\sigma^2_{\theta^{(u)}|\mathbf{y}} = m_{\theta^{(u)}|\mathbf{y}_n}(1 - m_{\theta^{(u)}|\mathbf{y}}). \tag{6.23}$$

Thus the output of the projection operator in Eq. (6.14) is the Gaussian distribution $\mathcal{CN}(\theta^{(u)}; m_{\theta^{(u)}|\mathbf{y}}, \sigma^2_{\theta^{(u)}|\mathbf{y}})$. Finally, the message $\mu_{\theta^{(u)}\to g_n}(\theta^{(u)})$ is a division of two Gaussian distributions whose mean and variance are given by the following equations,

$$\sigma^{-2}_{\theta^{(u)}\to g_n} = \sigma^{-2}_{\theta^{(u)}|\mathbf{y}} - \sigma^{-2}_{g_n\to\theta^{(u)}} \tag{6.24}$$

$$m_{\theta^{(u)}\to g_n} = \sigma^2_{\theta^{(u)}\to g_n}\left(\sigma^{-2}_{\theta^{(u)}|\mathbf{y}} m_{\theta^{(u)}|\mathbf{y}} - \sigma^{-2}_{g_n\to\theta^{(u)}} m_{g_n\to\theta^{(u)}}\right). \tag{6.25}$$

As a byproduct, the *a posteriori* LLR of $\theta^{(u)}$ becomes,

$$l_{\theta^{(u)}|\mathbf{y}} = \ln\frac{p(\theta^{(u)}|\mathbf{y})_{|\theta^{(u)}=0}}{p(\theta^{(u)}|\mathbf{y})_{|\theta^{(u)}=1}} = \ln\left(\frac{1-p_a^{(u)}}{p_a^{(u)}}\right) + \frac{1 - 2\Re\ \{\mathring{m}_{\theta^{(u)}}\}}{\mathring{\sigma}^2_{\theta^{(u)}}}. \tag{6.26}$$

It follows that hard UAD is obtained by performing hypothesis testing

$$\hat{\theta}^{(u)} = \begin{cases} 1 & \text{if } l_{\theta^{(u)}|\mathbf{y}} < 0 \\ 0 & \text{otherwise.} \end{cases} \tag{6.27}$$

which boils down to the simple and intuitive form

$$\hat{\theta}^{(u)} = \begin{cases} 0 & \text{if } \Re\{\mathring{m}_{\theta^{(u)}}\} \leq 0.5, \\ 1 & \text{otherwise.} \end{cases} \tag{6.28}$$

when $p_a^{(u)} = 1/2$.

6.5 Receiver implementation

In this section, the implementation of the receiver is presented, detailing each step of the process. The implementation begins with the initialization of the messages prior to the first iteration of EP. This initialization ensures that the starting point for the message-passing process is properly set, allowing the algorithm to converge efficiently.

Following the initialization, the message-passing schedule is described in detail. This schedule demonstrates the sequence in which EP messages are propagated through the factor graph, highlighting how information flows between the various nodes and factors. The correct sequence is crucial for ensuring the accuracy and performance of the algorithm, as the order of message updates impacts the convergence behavior.

6.5.1 Initialization

To begin the EP, it is essential to properly initialize certain messages to facilitate smooth convergence and reduce delays. Proper initialization not only accelerates convergence but also improves the stability of the algorithm during its iterations.

In the DEM subgraph, for a data (resp. a pilot) RE the message $\mu_{d_n^{(u)}\to g_n}(d_n^{(u)})$ is initialized as a complex Gaussian distribution with mean zero (resp. equal to the known pilot symbol) and unit (resp. zero) variance. Likewise, the LLRs at the decoder output $l_{c_{n,q}^{(u)}\to\chi_n^{(u)}}(c_{n,q}^{(u)})$ are initialized to zero. This ensures a uniform probability distribution over the values of $c_{n,q}^{(u)}$, implying equal likelihoods for $c_{n,q}^{(u)} = 0$ or $c_{n,q}^{(u)} = 1$.

In the CE subgraph, the messages $\mu_{\mathbf{x}_n^{(u)} \to g_n}(\mathbf{x}_n^{(u)})$ are initialized with a mean of zero and a variance of $E_s^{(u)}\mathbf{\Gamma}^{(u)}$, denoted as $\mu_{\mathbf{x}_n^{(u)} \to g_n}(\mathbf{x}_n^{(u)}) \sim \mathcal{CN}(\mathbf{x}_n^{(u)}; \mathbf{0}, E_s^{(u)}\mathbf{\Gamma}^{(u)})$.

For the UAD subgraph, the messages $\mu_{\theta^{(u)} \to g_n}(\theta^{(u)})$ are initialized with a mean of one and a variance of zero, expressed as $\mu_{\theta^{(u)} \to g_n}(\theta^{(u)}) \sim \mathcal{CN}(\theta^{(u)}; 1, 0)$, which is tantamount to forcing the existence of all users before the start of message-passing..

6.5.2 Message-passing schedule

We process all U user subgraphs in Fig. 3.4 sequentially, following this serial schedule:

1. First, process the CE subgraph by following the steps detaile in Sec. 4.6.
2. Next, process the DEM subgraph as described in Sec. 4.4.
3. Then, carry out the DEC subgraph processing, described in Sec. 4.5.
4. Finally, process the UAD subgraph by following the steps using the novel EP rule derived in the current chapter from the Wirtinger calculus approximation in Sec. 6.4.

6.5.3 Complexity analysis

As already mentioned, CE and DEM are the same for the receiver proposed in this chapter and the vectorized hybrid EP/BP algorithm introduced in chapter 4. Moreover, the per user per iteration complexity order of of UAD using Wirtinger-based EP as presented in Sec. 6.4 is dominated by the size-N_R matrix inversions in Eq. (6.10).
The complexity order of the receiver proposed in this chapter is summarized in Tab. 6.1.

Comparing with Tab. 5.1, this corresponds to the same complexity order as vectorized hybrid EP/BP.

Table 6.1: Per iteration and per user complexity order of the Wirtinger-based EP receiver.

Task	CE	DEM	UAD
Wirtinger-based EP	$\mathcal{O}(N_R^3)$	$\mathcal{O}(N_R^3 Q)$	$\mathcal{O}(N_R^3)$

6.6 Simulation results

We use the running example of grant-free OFDM-IDMA in Sec. 3.3 using the orthogonal pilots sequences described in Fig. 3.2 to evaluate the performance of the proposed method. For the sake of fair comparison with the methods in the previous chapters, the system model parameters are identical (see Tab. 4.1) under standard conditions of equal energy reception (see Tab. 4.2, except that $\rho^{(u)}$ can be non-zero) and known hyperparameters.

The system setup is consistent with that used in previous chapters. It includes 12 active users out of a total of 16, utilizing a 16-QAM modulation scheme. The coding structure comprises a recursive convolutional encoder with a generating function of $[5/7]_8$, along with a repetition encoder featuring a repetition factor of 4. Additionally, the system is equipped with $N_R = 4$ receive antennas and operates with $N = 1024$ total subcarriers, and antenna correlation is set to $\rho^{(u)} = 0.6$ for all user indices $u = 1, \ldots, U$.

We compare four types of EP algorithms in the simulations section:

1. **Proposed**: The iterative receiver presented in this chapter where UAD uses EP based on the Wirtinger calculus approximation in Sec. 6.4.

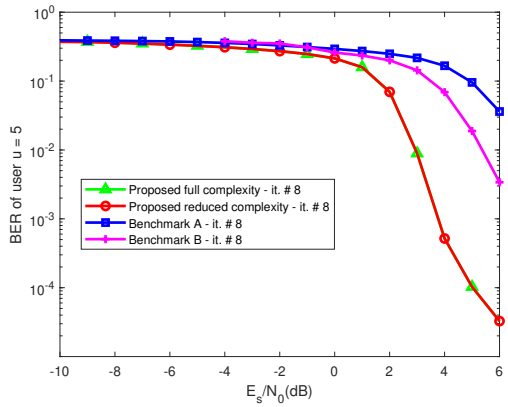

Figure 6.1: BER at convergence: 12 (out of U = 16) active equal energy users.

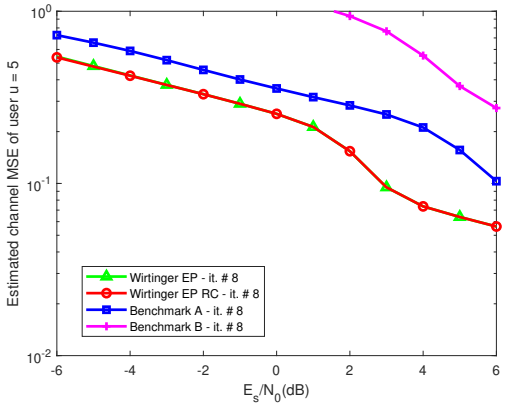

Figure 6.2: CFR estimation MSE at convergence: 12 (out of U = 16) active equal energy users.

2. **Proposed with RC**: A simplified version of the proposed receiver, where all messages on the u-th user subgraph are frozen as soon as $\hat{\theta}^{(u)} = 0$. The motivation behind this approach is that, since the proposed receiver can reliably detect inactive users (as will be demonstrated), ignoring them in subsequent iterations reduces computational complexity, which is proportional to the number of active users processed.

3. **Benchmark A**: For a fair comparison, we modify the receiver in Sec. 6.4 to use the competing Wirtinger calculus-based EP rule from [93] for UAD. The key distinction when applying Eq. (6.5) is that the Wirtinger calculus approximation in [93] is limited to scalar observations and hidden variables, which effectively ignores antenna correlation in our setup.

4. **Benchmark B**: We also evaluate a recent state-of-the-art competing hybrid EP/BP method for grant-free access [24], which uses scalar auxiliary variables.

6.6.1 12 active users out of 16

Fig. 6.1 and 6.2 illustrate the comparison of performance between the proposed full-complexity and RC algorithms with the Benchmark A and Benchmark B algorithms. The results clearly show that both versions of the proposed algorithms outperform the benchmarks in terms of BER and CFR estimation MSE. This is mainly due to the fact that the method proposed in this chapter is able to account for antenna correlation, while Benchmark A and Benchmark B are scalarized methods that are unable to do so by design. Importantly, the RC variant of the receiver presented in this chapter does not incur any penalty in terms of detection and channel estimation accuracy when compared to the full complexity method. Tab. 6.2 also shows that the number of user subgraphs that need further processing reduces to the correct number of active users on average after only 2 iterations for moderate to high SNR. This not only validates the proposed concept of complexity reduction, but also suggests great potential for the massive access scenario with low activity rate for which processing all potential users in the system is prohibitively complex. This issue will be the subject of the next chapter.

Fig. 6.3 presents a comparison of P_{fa} and the P_{md} for the proposed full and reduced complexity algorithms, as well as Benchmark A and Benchmark B algorithms.

In terms of P_{fa}, both the proposed full and reduced complexity algorithms outperform Benchmark A, achieving lower false alarm probabilities across the entire SNR range. This demonstrates their enhanced ability to accurately detect inactive users, minimizing false positives. While Benchmark B has vanishing P_{fa} across the entire SNR range (i.e. it almost

	E_s/N_0 (dB)			
iteration #	-10	-5	0	5
1	16	16	16	16
2	5	9.8	12	12
8	4.5	9.4	12	12

Table 6.2: Average number of tentative users processed by the proposed RC algorithm.

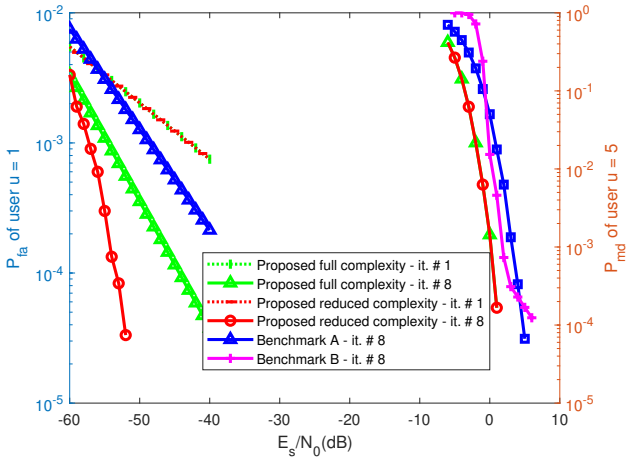

Figure 6.3: P_{fa} and P_{md} at convergence: 12 (out of U = 16) active equal energy users.

completely avoids false alarms), its non-vanishing P_{md} even at high SNR (see the error floor for SNRs beyond 4 dB) would lead to unwanted packet retransmissions that are very detrimental in the context of grant-free access. Also, the proposed algorithms consistently achieve better P_{md} than all benchmarks accross the SNR range.

The validity of the RC version of the proposed method is also exemplified by the fact P_{fa} is already extremely low at negative SNR even after the first iteration, meaning that early detection of user activity variables estimated as zero correspond to highly reliable decisions.

6.6.2 2 active users out of 16

user index	$\theta^{(u)}$	$\rho^{(u)}$	$E_s^{(u)}/N_0$ (dB)
$u \in \{1, \ldots, 14\}$	0	0.6	E_s/N_0 (dB)
$u \in \{15, 16\}$	1	0.6	E_s/N_0 (dB)

Table 6.3: Equal receive energy scenario with 2 active users out of $U = 16$.

Let us slightly modify the setup at the beginning of Sec. 6.6 to account for a lower user activity rate as given by Tab. 6.3.

In Fig. 6.4 and 6.5, we once again observe that the proposed full and reduced complexity algorithms significantly outperform both Benchmark A and Benchmark B, particularly in the case of Benchmark B. This further demonstrates the superiority of the proposed algorithms in terms of both BER and CFR estimation accuracy. The improved performance of the proposed algorithms, even in reduced-complexity form, highlights their effectiveness in handling symbol detection and channel estimation more reliably than the benchmark algorithms.

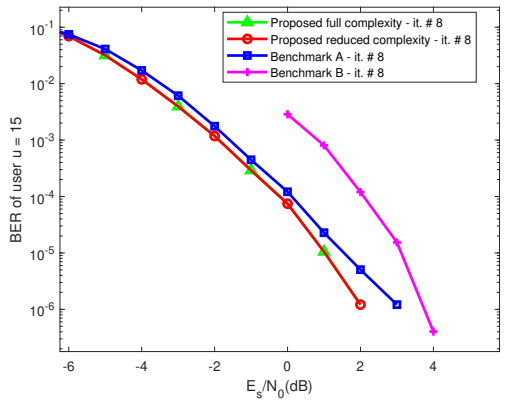

Figure 6.4: BER at convergence: 2 (out of U = 16) active equal energy users.

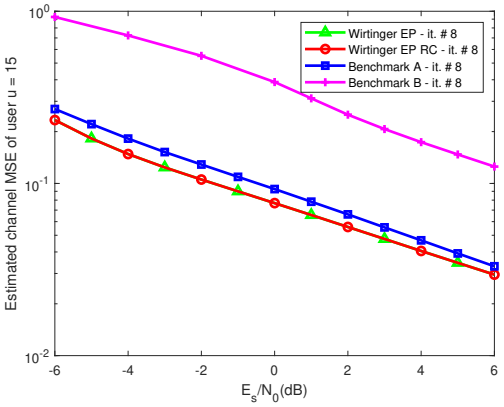

Figure 6.5: CFR estimation MSE at convergence: 2 (out of U = 16) active equal energy users.

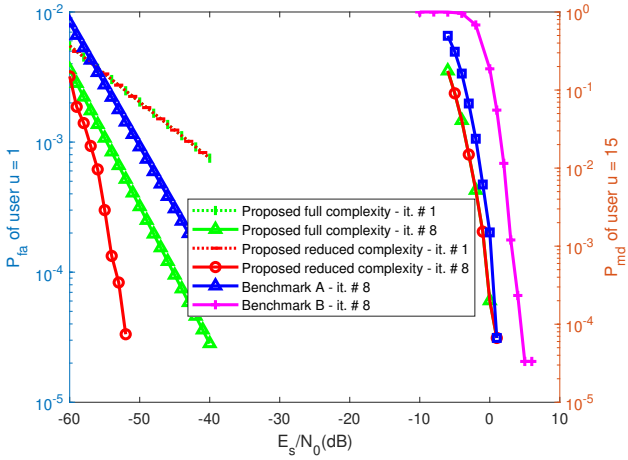

Figure 6.6: P_{fa} and P_{md} at convergence: 2 (out of U = 16) active equal energy users.

In Fig. 6.6, the comparison of probabilities for missed detection (P_{md}) and false alarm (P_{fa}) also reinforces the advantage of the proposed algorithms. Both the full and reduced complexity versions of the proposed methods consistently achieve better performance than Benchmark A and Benchmark B in terms of detecting active users with lower P_{md}, and avoiding false alarms with lower P_{fa}. Notably, Benchmark B performs the worst in terms of P_{md}, exhibiting a higher missed detection probability, which indicates its relative weakness in reliably detecting active users. However, it compensates for this by achieving the lowest P_{fa}, meaning it excels at avoiding false alarms, though at the cost of missing actual active users.

When comparing the two user activity scenarios—12 active users out of 16 versus 2 active users out of 16—it is evident from Fig. 6.4 and 6.1 that the latter scenario (2 active users) yields a substantial performance gain, with a nearly 5 dB SNR advantage in BER. This improvement can be attributed to the reduced level of interference and multi-user collisions when fewer users are active, allowing for more efficient detection and lower decoding errors in the system.

6.6.3 Comparison with vectorized hybrid EP/BP

Let us return to the original setup described at the beginning of Sec. 6.6.
In this section, we compare the results of Wirtinger calculus based EP with RC with vec-

torized hybrid EP/BP introduced in chapter 4. Let us recall from the introductory parapgraph in this chapter that the difference between the two algorithms lies only in the UAD subgraph processing, while over the rest of the factor graph, MUD, demodulation, decoding and CFR estimation, the same EP rules are applied.

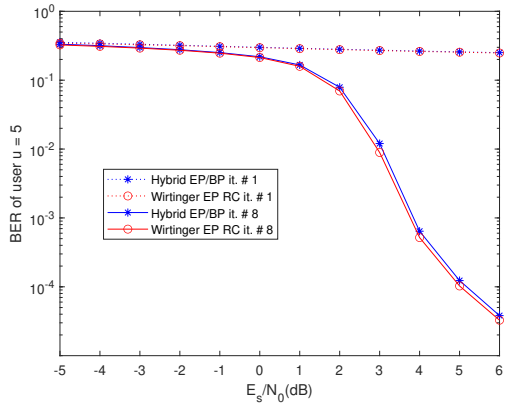

Figure 6.7: BER of hybrid EP/BP vs. Wirtinger based EP with RC.

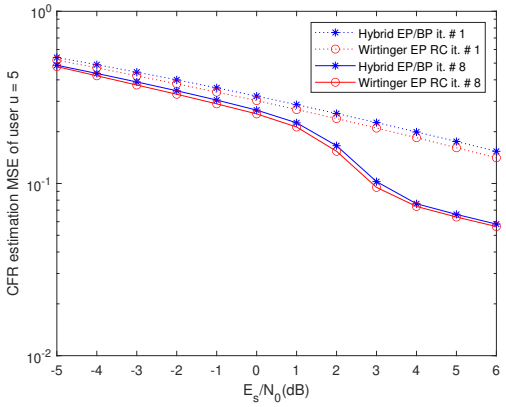

Figure 6.8: CFR estimation MSE of hybrid EP/BP vs. Wirtinger based EP with RC.

In Fig. 6.7 and 6.8, the BER and CFR estimation MSE of Wirtinger-based EP with reduced complexity (RC) are only marginally better than those of vectorized hybrid EP/BP . This similarity arises because both algorithms employ the same EP rules for demodulation, decoding, and CFR estimation.

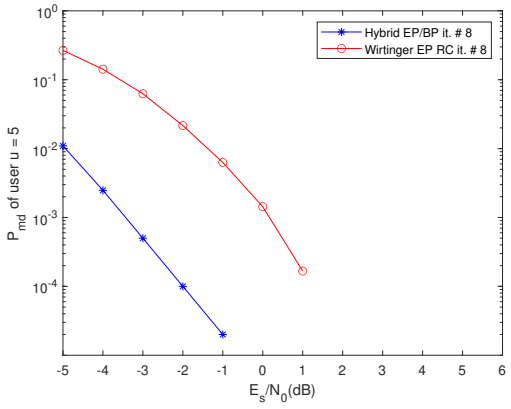

Figure 6.9: P_{md} of hybrid EP/BP vs. Wirtinger based EP with RC.

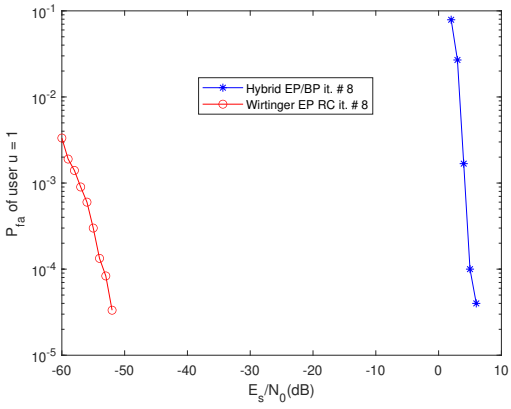

Figure 6.10: P_{fa} of hybrid EP/BP vs. Wirtinger based EP with RC.

However, Fig. 6.9 and Fig. 6.10, reveals that both algorithms have a different P_{fa} vs. P_{md} tradeoff. We believe that Wirtinger-based EP with RC achieves the better compromise, because both P_{fa} and P_{md} vanish for a SNR > 1 dB, while one has to wait until SNR > 4 dB for vectorized hybrid EP/BP. Also, note that the reduced-complexity mechanism would not work properly with vectorized hybrid EP/BP (i.e. without BER performance loss) since its P_{fa} does not vanish at convergence before an SNR > 4 dB. Thus, in the massive access scenario studied in the next chapter where reducing the complexity of joint CE/DEM/DEC/UAD with the maximum number of users U is of paramount interest, only Wirtinger-based EP will be considered.

6.7 Scalarized Wirtinger Calculus based EP

For the sake of complexity reduction, we also develop a scalarized version of the Wirtinger calculus based EP, similar to what has been done for hybrid EP/BP in chapter 5. Therefore, the scalarized model of channel is used ignoring the antenna correlation is Eq. (3.6), while the relevant scalarized observation model is Eq. (3.8). Also, the factor graph over which message passing applies is given by Fig. 3.5.

6.7.1 Message from $g_{n,r}$ to $\theta^{(u)}$

The message from $g_{n,r}$ to $\theta^{(u)}$, following the EP factor node rule Eq. (2.4), is expressed as:

$$\mu_{g_{n,r}\to\theta^{(u)}}(\theta^{(u)}) = \frac{\text{proj}_{\Phi}\left(k\cdot\mu_{\theta^{(u)}\to g_{n,r}}(\theta^{(u)})\tilde{p}(y_{n,r}|\theta^{(u)})\right)}{\mu_{\theta^{(u)}\to g_{n,r}}(\theta^{(u)})}, \tag{6.29}$$

where $\tilde{p}(y_{n,r}|\theta^{(u)})$ is the continuous Gaussian mixture defined in Eq. (5.26) and projected to the complex Gaussian,

$$\tilde{q}(y_{n,r}|\theta^{(u)}) = \mathcal{CN}(y_{n,r}; m_{y_{n,r}|\theta^{(u)}}(\theta^{(u)}), \sigma^2_{y_{n,r}|\theta^{(u)}}(\theta^{(u)})), \tag{6.30}$$

with mean vector $m_{y_{n,r}|\theta^{(u)}}(\theta^{(u)})$ in Eq. (5.28) and variance $\sigma^2_{y_{n,r}|\theta^{(u)}}(\theta^{(u)})$ in Eq. (5.29), rewritten as:

$$m_{y_{n,r}|\theta^{(u)}}(\theta^{(u)}) = h_{\theta^{(u)}\to g_{n,r}}\theta^{(u)} + I_{d_n^{(u)}\to g_{n,r}}, \tag{6.31}$$

$$\sigma^2_{y_{n,r}|\theta^{(u)}}(\theta^{(u)}) = |\theta^{(u)}|^2 A^{(u)}_{n,r} + B^{(u)}_{n,r}. \tag{6.32}$$

Here, the parameters in Eqs. (6.31) and (6.32) are defined as follows:

$$\begin{aligned}
h_{\theta^{(u)}\to g_{n,r}} &= m_{d_n^{(u)}\to g_{n,r}} m_{x_{n,r}^{(u)}\to g_{n,r}},\\
I_{d_n^{(u)}\to g_{n,r}} &= \sum_{u'\neq u} m_{d_n^{(u')}\to g_{n,r}} m_{x_{n,r}^{(u')}\to g_{n,r}} m_{\theta^{(u')}\to g_{n,r}},\\
A^{(u)}_{n,r} &= |m_{d_n^{(u)}\to g_{n,r}}|^2\sigma^2_{x_{n,r}^{(u)}\to g_{n,r}} + \sigma^2_{d_n^{(u)}\to g_{n,r}}\left(|m_{x_{n,r}^{(u)}\to g_{n,r}}|^2 + \sigma^2_{x_{n,r}^{(u)}\to g_{n,r}}\right),\\
B^{(u)}_{n,r} &= \sum_{u'\neq u}\sigma^2_{\theta^{(u')}\to g_{n,r}}\left(|m_{d_n^{(u')}\to g_{n,r}}|^2 + \sigma^2_{d_n^{(u')}\to g_{n,r}}\right)\left(|m_{x_{n,r}^{(u')}\to g_{n,r}}|^2 + \sigma^2_{x_{n,r}^{(u')}\to g_{n,r}}\right)\\
&\quad + \sum_{u'\neq u}|m_{\theta^{(u')}\to g_{n,r}}|^2\left[|m_{d_n^{(u')}\to g_{n,r}}|^2\sigma^2_{x_{n,r}^{(u')}\to g_{n,r}} + \sigma^2_{d_n^{(u')}\to g_{n,r}}\left(|m_{x_{n,r}^{(u')}\to g_{n,r}}|^2 + \sigma^2_{x_{n,r}^{(u')}\to g_{n,r}}\right)\right]\\
&\quad + N_0.
\end{aligned} \tag{6.33}$$

Using the second-order Taylor series approximation with Wirtinger calculus in Eq. (6.4) for the logarithm of the projection operator argument in Eq. (6.29) around a design parameter θ_0, the message becomes $\mu_{g_{n,r}\to\theta^{(u)}}(\theta^{(u)}) = \mathcal{CN}(\theta^{(u)}; m_{g_{n,r}\to\theta^{(u)}}, \sigma^2_{g_{n,r}\to\theta^{(u)}})$, with mean and variance given by:

$$
\begin{aligned}
\frac{1}{\sigma^2_{g_{n,r}\to\theta^{(u)}}} &= (|\theta_0|^2 A^{(u)}_{n,r} + B^{(u)}_{n,r})^{-2} A^{(u)}_{n,r} B^{(u)}_{n,r} \\
&+ |h_{\theta^{(u)}\to g_{n,r}}|^2 (|\theta_0|^2 A^{(u)}_{n,r} + B^{(u)}_{n,r})^{-2} B^{(u)}_{n,r} \\
&+ |y_{n,r} - I_{d^{(u)}_n \to g_{n,r}}|^2 (|\theta_0|^2 A^{(u)}_{n,r} + B^{(u)}_{n,r})^{-2} A^{(u)}_{n,r} \\
&- 2|y_{n,r} - I_{d^{(u)}_n \to g_{n,r}} - h_{\theta^{(u)}\to g_{n,r}}\theta_0|^2 (|\theta_0|^2 A^{(u)}_{n,r} + B^{(u)}_{n,r})^{-3} A^{(u)}_{n,r} B^{(u)}_{n,r}, \\
\frac{m_{g_{n,r}\to\theta^{(u)}}}{\sigma^2_{g_{n,r}\to\theta^{(u)}}} &= h^*_{\theta^{(u)}\to g_{n,r}} (|\theta_0|^2 A^{(u)}_{n,r} + B^{(u)}_{n,r})^{-1} (y_{n,r} - I_{d^{(u)}_n \to g_{n,r}} - h_{\theta^{(u)}\to g_{n,r}}\theta_0) \\
&+ \theta_0 \Bigg(\frac{1}{\sigma^2_{g_n\to\theta^{(u)}}} - A^{(u)}_{n,r} (|\theta_0|^2 A^{(u)}_{n,r} + B^{(u)}_{n,r})^{-1} \\
&\quad + |y_{n,r} - I_{d^{(u)}_n \to g_{n,r}} - h_{\theta^{(u)}\to g_{n,r}}\theta_0|^2 (|\theta_0|^2 A^{(u)}_{n,r} + B^{(u)}_{n,r})^{-2} A^{(u)}_{n,r} \Bigg),
\end{aligned}
\tag{6.34}
$$

6.7.2 Message from $\theta^{(u)}$ to $g_{n,r}$

Using the EP variable node rule for $\theta^{(u)}$

$$
\mu_{\theta^{(u)}\to g_{n,r}}(\theta^{(u)}) = \frac{\text{proj}_\Phi\Big(p(\theta^{(u)}|\mathbf{y})\Big)}{\mu_{g_{n,r}\to\theta^{(u)}}(\theta^{(u)})}. \tag{6.35}
$$

The argument of the projection operator represents the *a posteriori* pdf of $\theta^{(u)}$ such that

$$
p(\theta^{(u)}|\mathbf{y}) = k \cdot p(\theta^{(u)}) \prod_{n'=0,r'=1}^{N-1,N_R} \mu_{g_{n',r'}\to\theta^{(u)}}(\theta^{(u)}), \tag{6.36}
$$

where k is the normalization constant. Recall that the prior of $\theta^{(u)}$ is given by Eq. (6.16) and $\mu_{g_{n',r}\to\theta^{(u)}}(\theta^{(u)})$ is of the form,

$$
\mu_{g_{n',r'}\to\theta^{(u)}}(\theta^{(u)}) \propto \mathcal{CN}(\theta^{(u)}; m_{g_{n',r'}\to\theta^{(u)}}, \sigma^2_{g_{n',r'}\to\theta^{(u)}}). \tag{6.37}
$$

The product of the messages $\mu_{g_{n',r'}\to\theta^{(u)}}(\theta^{(u)})$ in Eq. (6.36) is also Gaussian, so

$$
\prod_{n'=0,r'=1}^{N-1,N_R} \mu_{g_{n',r'}\to\theta^{(u)}}(\theta^{(u)}) = \mathcal{CN}(\theta^{(u)}; \mathring{m}_{\theta^{(u)}}, \mathring{\sigma}^2_{\theta^{(u)}}), \tag{6.38}
$$

where the mean $\mathring{m}_{\theta^{(u)}}$ and variance $\mathring{\sigma}^2_{\theta^{(u)}}$ are given by

$$
\mathring{\sigma}^{-2}_{\theta^{(u)}} = \sum_{n'=0,r'=1}^{N-1,N_R} \sigma^{-2}_{g_{n',r'}\to\theta^{(u)}} \tag{6.39}
$$

$$
\mathring{m}_{\theta^{(u)}} = \mathring{\sigma}^2_{\theta^{(u)}} \sum_{n'=0,r'=1}^{N-1,N_R} \sigma^{-2}_{g_{n',r'}\to\theta^{(u)}} m_{g_{n',r'}\to\theta^{(u)}}. \tag{6.40}
$$

The normalization constant k in Eq. (6.36) is calculated as

$$
k = \frac{1}{(1-p^{(u)}_a)\mathcal{CN}(0; \mathring{m}_{\theta^{(u)}}, \mathring{\sigma}^2_{\theta^{(u)}}) + p^{(u)}_a \mathcal{CN}(1; \mathring{m}_{\theta^{(u)}}, \mathring{\sigma}^2_{\theta^{(u)}})}. \tag{6.41}
$$

Consequently, the mean and variance of the *a posteriori PMF* of $\theta^{(u)}$, i.e., $p(\theta^{(u)}|\mathbf{y})$, are expressed as

$$
\begin{aligned}
m_{\theta^{(u)}|\mathbf{y}} &= \frac{p_a^{(u)}\mathcal{CN}(1;\mathring{m}_{\theta^{(u)}},\mathring{\sigma}^2_{\theta^{(u)}})}{(1-p_a^{(u)})\mathcal{CN}(0;\mathring{m}_{\theta^{(u)}},\mathring{\sigma}^2_{\theta^{(u)}})+p_a^{(u)}\mathcal{CN}(1;\mathring{m}_{\theta^{(u)}},\mathring{\sigma}^2_{\theta^{(u)}})} \\
&= \frac{1}{\frac{1-p_a^{(u)}}{p_a^{(u)}}e^{\frac{1-2\Re\left\{\mathring{m}_{\theta^{(u)}}\right\}}{\mathring{\sigma}^2_{\theta^{(u)}}}}+1},
\end{aligned}
\tag{6.42}
$$

$$
\sigma^2_{\theta^{(u)}|\mathbf{y}} = m_{\theta^{(u)}|\mathbf{y}}(1-m_{\theta^{(u)}|\mathbf{y}}). \tag{6.43}
$$

Therefore, the output of the projection operator in Eq. (6.35) is the Gaussian distribution $\mathcal{CN}(\theta^{(u)};m_{\theta^{(u)}|\mathbf{y}},\sigma^2_{\theta^{(u)}|\mathbf{y}})$. Finally, the message $\mu_{\theta^{(u)}\to g_{n,r}}(\theta^{(u)})$ is the quotient of two Gaussian distributions, with mean and variance given by

$$
\sigma^{-2}_{\theta^{(u)}\to g_{n,r}} = \sigma^{-2}_{\theta^{(u)}|\mathbf{y}} - \sigma^{-2}_{g_{n,r}\to\theta^{(u)}} \tag{6.44}
$$

$$
m_{\theta^{(u)}\to g_{n,r}} = \sigma^2_{\theta^{(u)}\to g_{n,r}}\left(\sigma^{-2}_{\theta^{(u)}|\mathbf{y}}m_{\theta^{(u)}|\mathbf{y}} - \sigma^{-2}_{g_{n,r}\to\theta^{(u)}}m_{g_{n,r}\to\theta^{(u)}}\right). \tag{6.45}
$$

As a byproduct, the *a posteriori* LLR of $\theta^{(u)}$ is calculated as

$$
l_{\theta^{(u)}|\mathbf{y}} = \ln\frac{p(\theta^{(u)}|\mathbf{y})_{|\theta^{(u)}=0}}{p(\theta^{(u)}|\mathbf{y})_{|\theta^{(u)}=1}} = \ln\left(\frac{1-p_a^{(u)}}{p_a^{(u)}}\right)+\frac{1-2\Re\left\{\mathring{m}_{\theta^{(u)}}\right\}}{\mathring{\sigma}^2_{\theta^{(u)}}}. \tag{6.46}
$$

This enables hard UAD through hypothesis testing as follows:

$$
\hat{\theta}^{(u)} = \begin{cases} 1 & \text{if } l_{\theta^{(u)}|\mathbf{y}} < 0, \\ 0 & \text{otherwise.} \end{cases} \tag{6.47}
$$

Under the condition $p_a^{(u)} = 1/2$, this test simplifies to

$$
\hat{\theta}^{(u)} = \begin{cases} 0 & \text{if } \Re\left\{\mathring{m}_{\theta^{(u)}}\right\} \leq 0.5, \\ 1 & \text{otherwise.} \end{cases} \tag{6.48}
$$

6.7.3 Simulation results

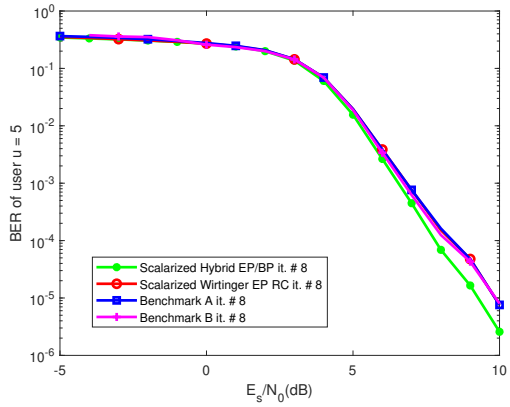

Figure 6.11: BER of scalarized hybrid EP/BP vs. scalarized Wirtinger based EP with RC.

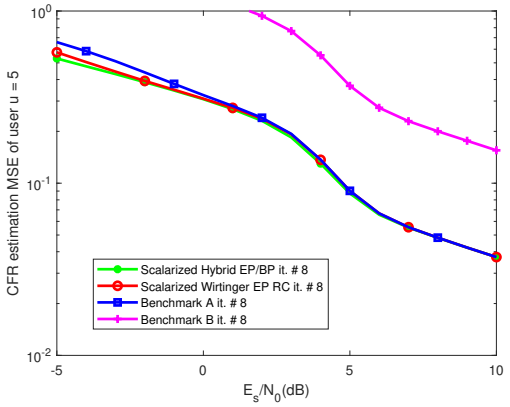

Figure 6.12: CFR estimation MSE of scalarized hybrid EP/BP vs. scalarized Wirtinger based EP with RC.

In Fig. 6.11 and 6.12, the BER and CFR estimation MSE of all the algorithms scalarized hybrid EP/BP, scalarized Wirtinger EP with RC, Benchmark A and Benchmark B are approximately the same (except that benchmark B has again a higher CFR estimation MSE). This similarity arises because all of these algorithms employ the same EP rules for demodulation, decoding, and CFR estimation.

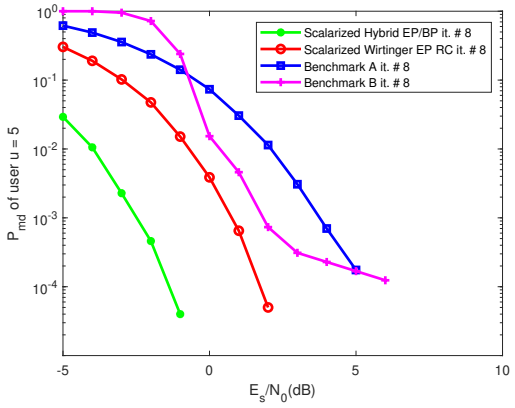

Figure 6.13: P_{md} of scalarized hybrid EP/BP vs. scalarized Wirtinger based EP with RC.

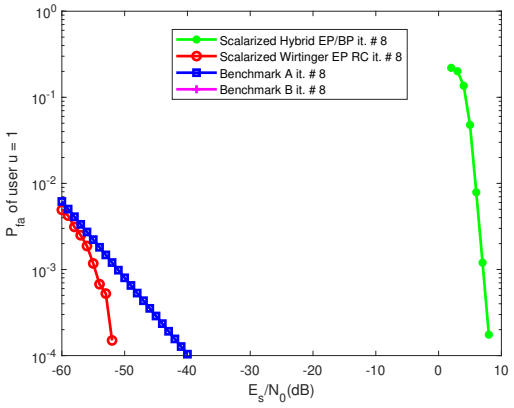

Figure 6.14: P_{fa} of scalarized hybrid EP/BP vs. scalarized Wirtinger based EP with RC.

In Fig. 6.13, the scalarized hybrid EP/BP achieves the lowest P_{md}, followed by scalarized Wirtinger-based EP with RC, Benchmark B, and Benchmark A. Conversely, in Fig. 6.14, Benchmark B shows a near-zero P_{fa} across low SNR values. Among other methods, scalarized Wirtinger-based EP with RC achieves the next lowest P_{fa}, followed by Benchmark A and scalarized hybrid EP/BP. These results highlight that the Wirtinger-based EP with RC stands out among these algorithms, providing both low P_{md} and P_{fa} while maintaining reduced computational complexity. This efficiency enables inactive users to be processed within a single EP cycle over the factor graph, optimizing performance with minimized processing effort.

Chapter 7

Massive grant-free access with non-orthogonal pilots

In the IoT ecosystem, devices like sensors and actuators gather and transmit data over wireless networks to improve efficiency, safety, and maintenance across various sectors, including transportation, healthcare, and manufacturing. A key characteristic of IoT is massive connectivity, where numerous devices sporadically send short data packets. This feature necessitates a rethinking of traditional scheduling protocols. Massive grant-free [98] access emerges as a promising solution, enabling a large number of uncoordinated devices to coexist without explicit resource allocation, thus eliminating the signaling overhead found in conventional contention-based protocols. While facing the same challenges as the grant-free access systems considered in the previous chapters, supporting massive connectivity needs sufficient initialization so that the methods proposed so far remain applicable. The later initialization problem is investigated in the form of efficient pilot-only joint UAD and CE in the present chapter.

7.1 Related work

Current solutions can be categorized by the type of pilot sequences used. Preamble-based methods assume that pilots and data share the same quasi-static channel. Initial approaches relied on orthogonal preambles [99], but it soon became evident that the number of orthogonal pilots is limited by their length, which can lead to pilot collisions [100] or necessitate the use of non-orthogonal Zadoff-Chu sequences [101], though at a cost to performance. Alternatively, non-orthogonal pseudo-random preambles can reduce the ratio between pilot symbols and data payload, making them attractive for short packet transmissions. This approach has been explored in several works [102, 103, 104, 105, 106] by leveraging advanced receivers based on AMP techniques (see [107]). Extensions to periodic pilots, meanwhile, have proven useful for tracking time-varying channels across both the time [108] or frequency [109, 110, 111].

Various techniques have been explored in the literature to address the challenge of joint UAD and CE with known data. These include computationally intensive methods such as maximum likelihood [112] or maximum a posteriori estimation [113], as well as machine learning approaches that require large amounts of labeled data [114, 115].

In this work, we focus on compressive sensing methods [116], which offer a good trade-off between complexity, pilot overhead, and performance, making them particularly suitable for massive access scenarios. Earlier solutions, based on greedy algorithms [117] or convex relaxation [118], typically provided hard estimates of user activity and channels.

On the other hand, iterative Bayesian learning techniques not only offer activity detection and channel estimation but also provide reliability measures, all while minimizing the number of non-payload data, which is critical for short packet, grant-free transmissions. Applications of these methods to our problem have been studied before. Sparse Bayesian Learning (SBL) [110, Sec. V] models the channels using conditional Gaussian distributions, where the variances are treated as hyperparameters and estimated through expectation maximization (EM) to capture the sparsity.

Another related approach with reduced complexity is EP [119], which simplifies the process by assuming factorizations of the joint distribution. Further simplifications, as explained in [120], result in AMP-like algorithms with even lower complexity [102, 103, 104, 105, 106]. However, these AMP-based algorithms are typically limited by the fact that their approximations rely on large system dimension assumptions.

7.2 Main Contributions

The key contributions in this chapter, along with its advancements over prior work, can be summarized as follows:

- We propose a SBL algorithm to address the problem of clustered sparsity in the rows of signals within Multiple Measurement Vector (MMV) models. This framework is applied to estimate users' multi-antenna Channel Impulse Response (CIR)s, making the proposed algorithm ideal for pilot-only joint UAD/CE in the presence of Inter-user Interference (IUI).
- Grant-free access relies on uncoordinated signal transmission between users and the receiver, making it difficult to obtain parameters like users' receive energy and channel correlation. Our SBL estimator provides a solution to this often-overlooked challenge as a byproduct of its estimation process.
- We consider using the vectorized Wirtinger-based EP receiver with reduced complexity (RC) introduced in chapter 6, to enhance UAD and CE while also decoding the data. This EP-based receiver is dependent on proper initialization, which accounts for channel uncertainty provided by the SBL estimator. In standalone mode, SBL alone leads to suboptimal user support and CIR recovery. Therefore, the symbiotic use of both algorithms within a two-stage receiver architecture is an effective solution to the challenges of Massive Grant-free Access NOMA (mGF-NOMA).

7.3 Problem formulation

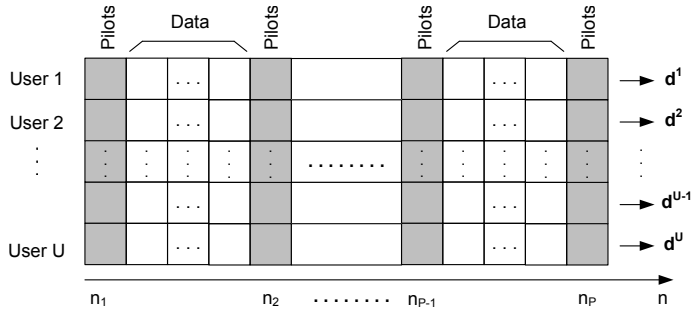

Figure 7.1: mGF-NOMA frequency domain RE sharing: data symbols (white) and non-orthogonal scattered pilot sequences (shaded).

The retained system model for mGF-NOMA is the one described in chapter 3, where for simplicity we restrict the applicability to CD-NOMA codebooks sent over frequency-domain

REs, so that the dynamic channel model is the CFR derived from the frequency-selective block fading model introduced in Sec. 3.1.3 b) (see Eq. (3.4)). Consequently, the usual memoryless vectorized observation model in Eq. (3.7) applies. While the general principle of this access model is not essentially different from the ones used in previous chapters, the key difference of massive access lies in the fact that there is a large number of potential users indexed by $u = 1, \ldots, U$, each having low activity rate $p_a^{(u)}$.

Notations for sub-vector and sub-matrix manipulations are introduced as follows. Considering a vector $\mathbf{v}$, $\mathbf{v}_{\mathcal{S}}$ denotes the subvector indexed by the columns in the set $\mathcal{S}$. Considering a matrix $\mathbf{A}$, $\mathbf{A}_{:,c}$ denotes its c-th column, $\mathbf{A}_{l_1:l_2,c_1:c_2}$ denotes the submatrix with line (resp. column) indices $l_1 \leq l \leq l_2$ (resp. $c_1 \leq c \leq c_2$) and $\mathrm{diag}(\mathbf{A})$ is the column vector containing the main diagonal of matrix $\mathbf{A}$. Moreover, the Kronecker product is denoted by $\otimes$.

There is another key distinction with respect to the previous chapters: the non-orthogonal pilot sequences shown in Fig. 3.3 are employed for each user, whereas in the previous chapters, the orthogonal pilot sequences shown in Fig. 3.2 were utilized for all users, thus limiting drastically the total number of potential users in the system, U.

Non-orthogonal pilot sequences differ in that, instead of each user having distinct, orthogonal pilots, all users simultaneously transmit their pilot symbols on the same subcarriers. This simultaneous transmission leads to collisions between pilot symbols on these subcarriers, making MUD more challenging but also allowing for more flexible system design and increased capacity in terms of number of users. A visual representation of the non-orthogonal pilot structure in Fig. 3.3 is depicted in Fig. 7.1 in the particular case of frequency domain REs, where pilot subcarriers are uniformly spaced and shared by all potential users.

The uniform spacing of pilot subcarriers ensures that the channel can be effectively captured, but the use of non-orthogonal pilots introduces interference that must be accounted for in the estimation process. This creates a more complex scenario for joint UAD and CE, as collisions at the pilot subcarriers require advanced signal processing techniques to separate users and estimate their channels. This non-orthogonal structure aligns with the massive connectivity requirements of mGF-NOMA, where a large number of devices need to access the network without pre-coordination, thus enabling efficient grant-free transmission with reduced pilot overhead.

Considering the problem of pilot-only initial acquisition of the channels and user activity, we denote the u-th user pilot symbol subvector indexed by the pilot set $\mathcal{P} = \{n_1, n_2, ..., n_P\}$ in Fig. 7.1 by

$$\mathbf{d}_{\mathcal{P}}^{(u)} = \begin{bmatrix} d_{n_1}^{(u)} \\ d_{n_2}^{(u)} \\ \vdots \\ d_{n_P}^{(u)} \end{bmatrix}, \tag{7.1}$$

and we assume a PDP common to all users written in vector form as

$$\mathbf{p} = \begin{bmatrix} p_0 \\ p_1 \\ \vdots \\ p_{L-1} \end{bmatrix}. \tag{7.2}$$

The problem at hand is described under the framework of sparse signal processing. Considering the sparsity of the vector

$$\boldsymbol{\gamma} = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_U \end{bmatrix}, \tag{7.3}$$

where $\gamma^{(u)} = \theta^{(u)} E_s^u$ for $u = 1, \ldots, U$, the desired signal matrix corresponding to the u-th user is denoted by

$$\mathbf{X}^{(u)} = \sqrt{\gamma^{(u)}} \begin{bmatrix} \sqrt{p_0}(\mathbf{\Gamma}^{(u)\frac{1}{2}} \mathbf{g}_0^{(u)})^T \\ \sqrt{p_1}(\mathbf{\Gamma}^{(u)\frac{1}{2}} \mathbf{g}_1^{(u)})^T \\ \vdots \\ \sqrt{p_{L-1}}(\mathbf{\Gamma}^{(u)\frac{1}{2}} \mathbf{g}_{L-1}^{(u)})^T \end{bmatrix} \in \mathbb{C}^{L \times N_R}, \tag{7.4}$$

giving rise to the compound block sparse signal matrix

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} \\ \vdots \\ \mathbf{X}^{(U)} \end{bmatrix} \in \mathbb{C}^{LU \times N_R}. \tag{7.5}$$

Combining Eqs. (3.4) and (3.7), the observation matrix over the pilot subcarriers can be written as a matrix in in $\mathbb{C}^{P \times N_R}$

$$\mathbf{Y} = \begin{bmatrix} y_{n_1,1} & y_{n_1,2} & \cdots & y_{n_1,N_R} \\ y_{n_2,1} & y_{n_2,2} & \cdots & y_{n_2,N_R} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n_P,1} & y_{n_P,2} & \cdots & y_{n_P,N_R} \end{bmatrix} = \sum_{u=1}^{U} \theta^{(u)} \begin{bmatrix} d_{n_1}^{(u)} \mathbf{H}_{n_1}^{(u)T} \\ d_{n_2}^{(u)} \mathbf{H}_{n_2}^{(u)T} \\ \vdots \\ d_{n_P}^{(u)} \mathbf{H}_{n_{P-1}}^{(u)T} \end{bmatrix} + \mathbf{W}, \tag{7.6}$$

affected by the noise matrix with independent Gaussian zero-mean entries with variance N_0 over the pilot subcarriers

$$\mathbf{W} = \begin{bmatrix} \mathbf{w}_{n_1}^T \\ \mathbf{w}_{n_2}^T \\ \vdots \\ \mathbf{w}_{n_P}^T \end{bmatrix}. \tag{7.7}$$

Consequently, we obtain a MMV model,

$$\mathbf{Y} = \underbrace{\mathbf{A}}_{\in \mathbb{C}^{P \times LU}} \underbrace{\mathbf{X}}_{\in \mathbb{C}^{LU \times N_R}} + \mathbf{W}, \tag{7.8}$$

whose the sensing matrix is given by

$$\mathbf{A} = \left[\operatorname{diag}(\mathbf{d}_{\mathcal{P}}^{(1)}) \mathbf{F} \,\middle|\, \operatorname{diag}(\mathbf{d}_{\mathcal{P}}^{(2)}) \mathbf{F} \,\middle|\, \ldots \,\middle|\, \operatorname{diag}(\mathbf{d}_{\mathcal{P}}^{(U)}) \mathbf{F} \right], \tag{7.9}$$

and where $\mathbf{F}$ is a DFT submatrix given by,

$$\mathbf{F} = \begin{bmatrix} 1 & e^{-j\frac{2\pi}{N} n_1} & \cdots & e^{-j\frac{2\pi}{N} n_1 (L-1)} \\ 1 & e^{-j\frac{2\pi}{N} n_2} & \cdots & e^{-j\frac{2\pi}{N} n_2 (L-1)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & e^{-j\frac{2\pi}{N} n_P} & \cdots & e^{-j\frac{2\pi}{N} n_P (L-1)} \end{bmatrix}. \tag{7.10}$$

7.4 Proposed structured multiple SBL (S-MSBL) solution

Assigning different Gaussian priors to the signal matrix $\mathbf{X}$ leads to various learning algorithms. The original Multiple Sparse Bayesian Learning (MSBL) algorithm, introduced in [121], assumes that $\mathbf{X}$ is row-sparse, with independent Gaussian variables in each row. An extension to this was proposed in [122], where correlated vectors within each row are considered, though the covariance matrix for each row must remain identical to prevent overfitting.

Motivated by the concept of structured sparsity, where consecutive rows of $\mathbf{X}$ share the same support, we propose a structured multiple SBL (S-MSBL) algorithm tailored to the model described in Eq. (7.8). Although [123] attempted to incorporate correlation within each row of $\mathbf{X}$, it encountered the same overfitting issue as in [122], requiring the correlation to be independent of the row index.

In our model, however, the actual correlation within each row of $\mathbf{X}$ is user-dependent. To address this, we simplify the model by assuming $\rho^{(u)} = 0$ for all users $u = 1, \ldots, U$, which allows us to derive the proposed S-MSBL algorithm. This simplification is sufficient to address IUI and perform initial joint UAD and CE. We later compensate for this model mismatch by considering inter-antenna correlation during channel refinement, using the theory of Wirtinger-based expectation propagation with RC, as discussed in the previous chapter (refer to Sec. 7.5 for more details).

Finally, we explain how to construct a custom estimator for $\rho^{(u)}$ and $E_s^{(u)}$ based on the channel impulse response (CIR) estimates provided by S-MSBL for all users $u = 1, \ldots, U$.

From the Eq. (7.8), the observation vector for an individual receive antenna i can be written as,

$$\mathbf{Y}_{:,i} = \mathbf{A}\mathbf{X}_{:,i} + \mathbf{W}_{:,i}, \tag{7.11}$$

so that,

$$p(\mathbf{Y}_{:,i}|\mathbf{X}_{:,i}) = \mathcal{CN}(\mathbf{Y}_{:,i}; \mathbf{A}\mathbf{X}_{:,i}, N_0\mathbf{I}_P). \tag{7.12}$$

Assuming that all the $\mathbf{X}_{:,i}$ elements are independent for $i = 1, ..., N_R$, then it has the prior pdf,

$$p(\mathbf{X}_{:,i}|\boldsymbol{\gamma}, \mathbf{p}) = \mathcal{CN}(\mathbf{X}_{:,i}; \mathbf{0}_{LU\times 1}, \boldsymbol{\Sigma}_{\mathbf{X}}), \tag{7.13}$$

where $\boldsymbol{\Sigma}_{\mathbf{X}}$ is given by,

$$\boldsymbol{\Sigma}_{\mathbf{X}} = \begin{bmatrix} \gamma_1 \operatorname{diag}(\mathbf{p}) & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \gamma_2 \operatorname{diag}(\mathbf{p}) & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \gamma_U \operatorname{diag}(\mathbf{p}) \end{bmatrix} = \operatorname{diag}(\boldsymbol{\gamma}) \otimes \operatorname{diag}(\mathbf{p}) \in \mathbb{R}^{LU\times LU}, \tag{7.14}$$

The joint conditional pdfs of $p(\mathbf{Y}|\mathbf{X})$ and $p(\mathbf{X}|\boldsymbol{\gamma}, \mathbf{p})$ are given by,

$$\begin{aligned} p(\mathbf{Y}|\mathbf{X}) &= \prod_{i=1}^{N_R} p(\mathbf{Y}_{:,i}|\mathbf{X}_{:,i}) \\ p(\mathbf{X}|\boldsymbol{\gamma}, \mathbf{p}) &= \prod_{i=1}^{N_R} p(\mathbf{X}_{:,i}|\boldsymbol{\gamma}, \mathbf{p}). \end{aligned} \tag{7.15}$$

In order to estimate the CIR based on the observations using the Minimum-mean-squared-error (MMSE) estimator, we need to compute the *a posteriori* pdf of $\mathbf{X}$,

$$p(\mathbf{X}|\mathbf{Y}, \boldsymbol{\gamma}, \mathbf{p}) = \frac{p(\mathbf{Y}|\mathbf{X})p(\mathbf{X}|\boldsymbol{\gamma}, \mathbf{p})}{p(\mathbf{Y}|\boldsymbol{\gamma}, \mathbf{p})} \propto \prod_{i=1}^{N_R} \mathcal{CN}(\mathbf{X}_{:,i}; \boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{A}^H\boldsymbol{\Sigma}_{\mathbf{Y}}^{-1}\mathbf{Y}_{:,i}, \boldsymbol{\Sigma}_{|\mathbf{Y}}), \tag{7.16}$$

where

$$\begin{aligned} \boldsymbol{\Sigma}_{\mathbf{Y}} &= \mathbf{A}^H\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{A}^H + N_0\mathbf{I}_P \\ \boldsymbol{\Sigma}_{|\mathbf{Y}} &= \boldsymbol{\Sigma}_{\mathbf{X}} - \boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{A}^H\boldsymbol{\Sigma}_{\mathbf{Y}}^{-1}\mathbf{A}\boldsymbol{\Sigma}_{\mathbf{X}}, \end{aligned} \tag{7.17}$$

and the mean of the normal distribution is the MMSE estimate of the CIR matrix **X** is given by,

$$\hat{\mathbf{X}}_{|\mathbf{Y}} = \mathbf{\Sigma}_{\mathbf{X}}\mathbf{A}^H\mathbf{\Sigma}_{\mathbf{Y}}^{-1}\mathbf{Y}. \tag{7.18}$$

7.4.1 Hyperparameter Estimation using the EM Framework

The EM algorithm [124] is commonly used to compute the ML estimate of parameters when both observed and unobserved (or hidden) data are involved.

In this context, the hyperparameters to be estimated using EM are $\boldsymbol{\gamma}$ and $\mathbf{p}$. To apply the EM algorithm, we require both the observed (incomplete) dataset $\mathbf{Y}$ and the unobserved (missing) dataset $\mathbf{X}$. The EM process consists of two steps: the expectation (E-step) and maximization (M-step).

The E-step can be written as:

$$\begin{aligned} Q(\boldsymbol{\gamma}, \mathbf{p}, \boldsymbol{\gamma}^{(t-1)}, \mathbf{p}^{(t-1)}) &= \int \ln p(\mathbf{Y}, \mathbf{X}|\boldsymbol{\gamma}, \mathbf{b}) p(\mathbf{Y}, \mathbf{X}|\boldsymbol{\gamma}^{(t-1)}, \mathbf{p}^{(t-1)}) d\mathbf{X} \\ &= \text{const} - N_R L \sum_{u=1}^{U} \log \gamma^{(u)} - N_R U \sum_{l=0}^{L-1} \log p_l - \sum_{i=1}^{N_R}\sum_{u=1}^{U} \frac{\operatorname{tr}\left(\operatorname{diag}(\mathbf{p})^{-1}(\mathbf{\Sigma}_{|\mathbf{Y}}^{(u),(t)} + \hat{\mathbf{X}}_{:,i|\mathbf{Y}}^{(u)}\hat{\mathbf{X}}_{:,i|\mathbf{Y}}^{(u)H})\right)}{\gamma^{(u)}}, \end{aligned} \tag{7.19}$$

where the superscript t denotes the EM iteration index.

The M-step then maximizes $Q(\boldsymbol{\gamma}, \mathbf{p}, \boldsymbol{\gamma}^{(t-1)}, \mathbf{b}^{(t-1)})$ with respect to $\boldsymbol{\gamma}$ and $\mathbf{p}$, providing updated estimates for these parameters:

$$\begin{aligned} \gamma^{(u),(t)} &= \frac{1}{N_R L}\sum_{i=1}^{N_R} \operatorname{tr}\left(\operatorname{diag}(\mathbf{p}^{(t-1)})^{-1}(\mathbf{\Sigma}_{|\mathbf{Y}}^{(u),(t)} + \hat{\mathbf{X}}_{:,i|\mathbf{Y}}^{(u)}\hat{\mathbf{X}}_{:,i|\mathbf{Y}}^{(u)H})\right),\ u = 1, \ldots, U \\ \mathbf{p}^{(t)} &= \frac{1}{N_R U}\sum_{i=1}^{N_R}\sum_{u=1}^{U} \frac{\operatorname{diag}(\mathbf{\Sigma}_{|\mathbf{Y}}^{(u),(t)} + \hat{\mathbf{X}}_{:,i|\mathbf{Y}}^{(u)}\hat{\mathbf{X}}_{:,i|\mathbf{Y}}^{(u)H})}{\gamma^{(u),(t)}}. \end{aligned} \tag{7.20}$$

Here, $\hat{\mathbf{X}}_{:,i}^{(u)}$ and $\mathbf{\Sigma}_{|\mathbf{Y}}^{(u),(t)}$ are defined as follows:

$$\begin{aligned} \mathbf{\Sigma}_{|\mathbf{Y}}^{(u),(t)} &= \mathbf{\Sigma}_{(u-1)L+1:uL,(u-1)L+1:uL|\mathbf{Y}}, \\ \hat{\mathbf{X}}_{:,i}^{(u)} &= \hat{\mathbf{X}}_{((u-1)L+1:uL,i)|\mathbf{Y}}. \end{aligned} \tag{7.21}$$

At the end of each EM iteration, $\mathbf{p}^{(t)}$ is renormalized to unit power to avoid ambiguities:

$$\mathbf{p}^{(t)} = \frac{\mathbf{p}^{(t)}}{\sum_{l=0}^{L-1} p_l^{(t)}}. \tag{7.22}$$

7.4.2 Estimation of Channel Model Parameters

When applying the S-MSBL algorithm, we initially assume that the antenna correlation $\rho^{(u)} = 0$. However, in practice, this is not the case, and it must be estimated via post-processing. We propose an estimator, similar to the one discussed in the hybrid EP/BP in chapter 4, to estimate $\rho^{(u)}$ as follows:

$$\hat{\rho}^{(u)} = \sum_{i=1}^{N_R-1} \frac{\text{trace}\left(\text{diag}\left(\mathbf{p}^{(t_{max})}\right)^{-1}(\hat{\mathbf{X}}^{(u),(t_{max})}_{:,i|\mathbf{Y}}\hat{\mathbf{X}}^{(u),(t_{max})H}_{:,i+1|\mathbf{Y}})\right)}{(N_R-1)L\gamma^{(u),(t_{max})}} \tag{7.23}$$

where t_{max} denotes the final iteration of EM. The estimated symbol energies are assigned at the end of S-MSBL according to $\hat{E}_s^{(u)} = \gamma^{(u),(t_{max})}$, for $u = 1, \ldots, U$.

The correlation coefficient between two consecutive discrete frequencies can be estimated as:

$$\hat{\sigma} = \sqrt{2\sum_{l=0}^{L-1} p_l^{(t_{max})}\left(1 - \cos\left(2\pi l/N\right)\right)}. \tag{7.24}$$

7.4.3 Tentative Hard UAD

For UAD, a standard approach is to classify users with high energy as active and those with low energy as inactive at the final iteration of the EM algorithm. The detection is performed as:

$$\hat{\theta}^{(u)}_{S-MSBL} = \begin{cases} 0, \text{ if } \hat{E}_s^{(u)} < \gamma_{\text{th}} \\ 1, \text{ otherwise,} \end{cases} \tag{7.25}$$

where $\gamma_{\text{th}} > 0$ is a pruning threshold chosen to minimize both the miss detection probability P_{md} and the false alarm probability P_{fa} at the output of the S-MSBL.

7.5 Receiver implementation

Different receivers can be implemented using the initial pilot-only UAD/CE described in Sec. 7.4. One approach could be separating the pilot-only joint UAD/CE from MUD/DEC. However, this method requires either high SNR or a high pilot-to-payload ratio, which is inefficient in terms of energy or bandwidth, respectively.

Another option is to integrate the joint UAD/CE into a code-aided receiver that generates virtual pilots, as discussed in [125]. However, this leads to an increasing number of virtual pilots N_p with each outer iteration of the code-aided receiver. The complexity of joint UAD/CE in this scenario becomes $O(t_{max}(LUN_p(N_p+N_R)+N_p^3))$. This growing complexity makes the approach impractical.

Therefore, focus shifts to a complete receiver implementation where initial UAD/CE, as performed using the S-MSBL outlined in Sec. 7.4, plays the role of a first stage. A second stage iteratively refines UAD/CE/DEM/DEC using the vectorized EP based on a Wirtinger calculus approximation along with the reduced complexity (RC) mechanism in chapter 6 (in the sequel, EP receiver will be the denomination for the vectorized method described in Sec. 6.4). This approach is later called Two-stage receiver and is illustrated in Fig. 7.2.

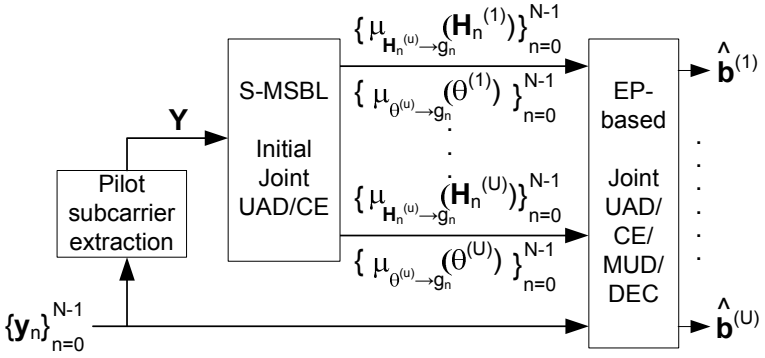

Figure 7.2: Complete description of the proposed two-stage architecture.

The EP receiver discussed in chapter 6 faces challenges related to incorrect message initialization within each user subgraph if the IUI problem is not adequately addressed. This issue, however, is effectively resolved by the proposed S-MSBL in Sec. 7.4, which mitigates pilot contamination resulting from the reuse of non-orthogonal pilot sequences on the same REs by multiple users.

First, the S-MSBL estimates the posterior distribution of the CIR for each user $u = 1, \ldots, U$ and then transforms it into its CFR counterpart over subcarrier n using the deterministic linear transformation

$$\mathbf{f}_n = [(e^{-j2\pi n/N})^0, (e^{-j2\pi n/N})^1, \ldots, (e^{-j2\pi n/N})^{L-1}], \tag{7.26}$$

which relates the two.

Now refering to the factor graph in Fig. 3.4 over which the EP receiver with RC in chapter 6 operates, recalling that the variable node $\mathbf{x}_n^{(u)}$ has been identified in this chapter to the CFR over the n-th subcarrier $\mathbf{H}_n^{(u)}$, the messages reaching g_n from the CE subgraph within the EP framework are initialized with improved accuracy:

$$\mu_{\mathbf{H}_n^{(u)} \to g_n}(\mathbf{H}_n^{(u)}) = \mathcal{CN}(\mathbf{H}_n^{(u)} : \mathbf{m}_{\mathbf{H}_n^{(u)}}, \mathbf{C}_{\mathbf{H}_n^{(u)}}), \tag{7.27}$$

where

$$\mathbf{m}_{\mathbf{H}_n^{(u)}} = (\mathbf{f}_n^T \hat{\mathbf{X}}_{|\mathbf{Y}}^{(u),(t_{max})})^T, \; \mathbf{\Sigma}_{\mathbf{H}_n^{(u)}} = \left(\mathbf{f}_n^T \hat{\mathbf{\Sigma}}_{|\mathbf{Y}}^{(u),(t_{max})} \mathbf{f}_n^*\right) \mathbf{I}_{N_R}.$$

Furthermore, we introduce the RC mechanism at the onset of the EP-based receiver by adjusting the messages over all UAD subgraphs for $n = 0, \ldots, N-1$ and $u = 1, \ldots, U$ as follows:

$$\mu_{\theta^{(u)} \to g_n}(\theta^{(u)}) = \mathcal{CN}(\theta^{(u)}; \hat{\theta}_{S-MSBL}^{(u)}, 0), \tag{7.28}$$

thus pre-cancelling the contributions of users already identified as inactive by S-MSBL, while leaving the UAD initialization for active users unchanged.

It is important to note that only a minor algorithmic adjustment is made to the EP-based receiver in chapter 6, specifically in message initialization, due to the enhanced CFR initialization Eq. (7.27) and the updated UAD initialization Eq. (7.28) for all users.

However, since most inactive users are eliminated early by S-MSBL, there is no need to process their subgraphs in the EP-based receiver, leading to significant complexity reduction from the very first iteration.

7.6 Simulation results

In the simulation results section, we first outline the detailed system setup used to evaluate the performance of the proposed receiver. This includes the key parameters such as the number of users, subcarriers, and antennas, as well as the pilot and payload structure, channel model assumptions, and noise characteristics. We then present the results obtained using the S-MSBL-based receiver.

Next, we compare the performance of the S-MSBL-based receiver with other leading algorithms from the state-of-the-art. These include block Simultaneous Orthogonal Matching Pursuit (SOMP), which is commonly used for sparse signal recovery, and EM-AMP, a popular technique for joint user activity detection and channel estimation in massive MIMO systems. The comparison highlights the strengths and limitations of each algorithm, particularly in challenging scenarios involving high levels of interference, low SNR, or limited pilot resources.

Finally, we analyze the computational complexity and convergence behavior of the S-MSBL algorithm in comparison to block SOMP (B-SOMP) and EM-AMP, discussing how these factors impact their suitability for real-time implementations in large-scale wireless communication systems.

7.6.1 Setup

We use the running example of grant-free OFDM-IDMA in Sec. 3.3 using the non-orthogonal pilots sequences described in Fig. 3.3 to evaluate the performance of the proposed two-stage receiver. The pilot subcarriers are spaced every 7 subcarriers, such that $\mathcal{P} = \{0, 7, 14, ..., 1023\}$. For the sake of fair comparison with the previous chapters, the system model parameters are identical (see Tab. 4.1), except that U is set to 156 with $K = 12$ randomly selected active users to simulate a massive access scenario and the channel modeling error parameter set to $\zeta = 25$. Equal energy levels ($E_s^{(u)} = E_s,\ u = 1, 2, ..., U$) are assumed, with antenna correlation coefficients $\rho^{(u)} = \rho = 0.6$ for $u = 1, \ldots, U$ and the channel model hyperparameters are unknown.

The channel tap length at the transmitter is $L = 128$, but it is truncated to $L = 9$ at the receiver side, so as to avoid overfitting while not compromising the performance. In the S-MSBL algorithm, a maximum of $t_{max} = 400$ iterations is used. For UAD inside the EM process, a threshold of $\gamma_{th} = 0.168 \times \exp(0.28 \times (E_s/N_0)_{dB}) - 0.009$ is applied, aiming to minimize both $P_{md}^{(S-MSBL)}$ and $P_{fa}^{(S-MSBL)}$ at the output of S-MSBL.

7.6.2 Validation of the proposed receiver

Different receiver architectures can be constructed using the approaches previously discussed. The following outlines three potential designs:

1. **Standalone-EP:**
   In this setup, only EP receiver with RC in chapter 6 is used, without leveraging any prior UAD or CE from S-MSBL. This approach means EP handles all estimation tasks independently, without any assistance from pre-estimated user activity or channel conditions.

2. **Conventional:**
   This design separates UAD/CE from MUD/DEC. First, UAD/CE is performed using S-MSBL, and then the estimates for both user activity and the channel are passed to an EP receiver restricted to MUD/DEC. The EP process operates based on these predetermined estimates, which improves accuracy.

3. **Two-Stage Receiver:**
   This is the advocated architecture in Sec. 7.5 where the initial UAD/CE estimates from

S-MSBL are used to initialize the EP receiver with RC in chapter 6. At each EP iteration, both UAD/CE and the MUD/DEC processes are refined. This iterative refinement improves overall performance by continuously updating UAD/CE along with demodulation and decoding.

These architectures differ in their reliance on S-MSBL and offer various trade-offs in complexity and performance. the corresponding performance results are illustrated in Fig. 7.3 to 7.6.

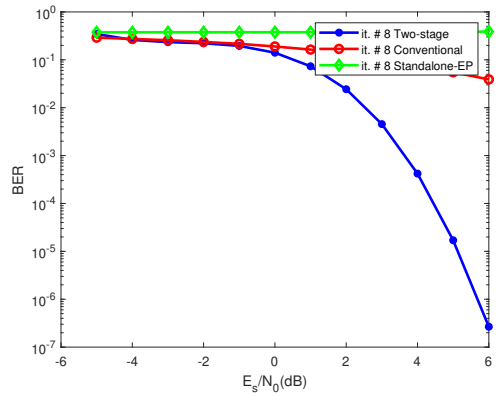

Figure 7.3: BER comparison for standalone-EP, conventional and two-stage receiver.

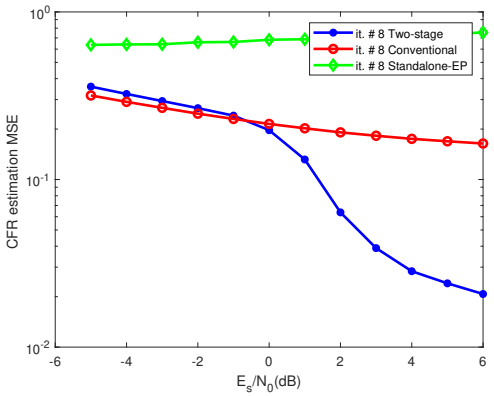

Figure 7.4: CFR MSE comparison for standalone-EP, conventional and two-stage receiver.

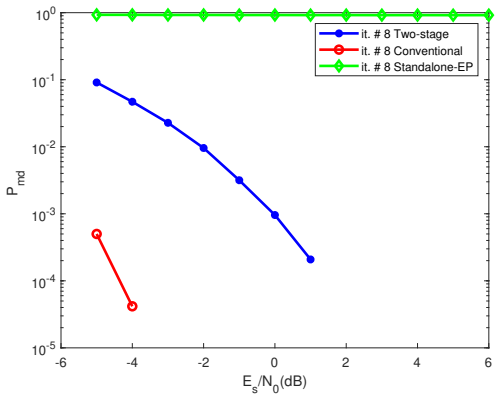

Figure 7.5: P_{md} comparison for standalone-EP, conventional and two-stage receiver.

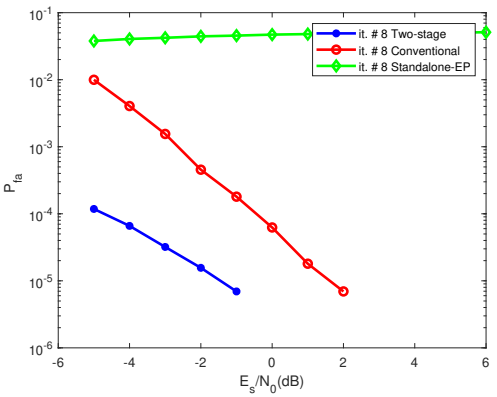

Figure 7.6: P_{fa} comparison for standalone-EP, conventional and two-stage receiver.

In the standalone-EP receiver, CE/UAD/MUD/DEC are jointly performed using EP rules, with each user subgraph processed sequentially. This approach struggles with IUI at pilot subcarriers, leading to poor UAD/CE quality during the first iteration, and the subsequent iterations do not improve performance. In the conventional receiver, UAD/CE is handled separately by S-MSBL before EP, but since EP only performs MUD/DEC, the improved data symbol estimates are not utilized for re-estimating the channel and user activities, resulting in suboptimal overall performance. While UAD at the output of S-MSBL can be fairly accurate at high SNR, the channel estimation suffers due to unavoidable CIR truncation, impacting the MUD/DEC performance in EP. The two-stage receiver succeeds because EP leverages the initial channel estimates and UAD from S-MSBL to initialize its UAD/CE subgraph messages. As the itera-

tions proceed, both the channel estimates and user activities are refined using the estimated data symbols, improving overall performance with each iteration.

7.6.3 Comparison with existing initial joint UAD/CE methods

In the following, we focus on the proposed two-stage architecture by comparing different algorithms for initial joint UAD/CE:

1. **B-SOMP**: This algorithm, as outlined in [126], is applied under known sparsity K, suitable for the MMV problem with structured sparsity in (7.8). Since only hard CIR estimates are obtained, the missing posterior covariance required to initialize the EP-based receiver is set to $\mathbf{\Sigma}_{|\mathbf{Y}} = 10^{-3}\mathbf{I}_{LU}$. Additionally, channel model parameters $(E_s^{(u)}, \rho^{(u)})$ for $u = 1, \ldots, U$ are estimated as described in [127, Sec. IV.A] due to the absence of $\boldsymbol{\gamma}$ and $\mathbf{p}$ estimates.
2. **EM-AMP**: The columns of the signal matrix $\mathbf{X}$ are estimated separately using 10 inner iterations of AMP [120]. Hyperparameters $\boldsymbol{\gamma}, \mathbf{p}$, and CIR coefficient activity probabilities are recovered using 400 outer EM iterations [128]. Channel model parameter estimation is performed as outlined in Sec. 7.4.2.
3. **S-MSBL**: The proposed receiver architecture is fully described in Sec. 7.4 and shown in Fig. 7.2.

The overall complexity of each method is summarized in Tab. 5.1 and the corresponding performance results are shown in Fig. 7.7 to 7.10.

Table 7.1: Overall complexity order for each initial joint UAD/CE.

Algorithm	Complexity
B-SOMP	$\mathcal{O}(KULPN_R)$
EM-AMP	$\mathcal{O}(t_{\text{max}}IULPN_R)$
S-MSBL	$\mathcal{O}(t_{\text{max}}ULP^2)$

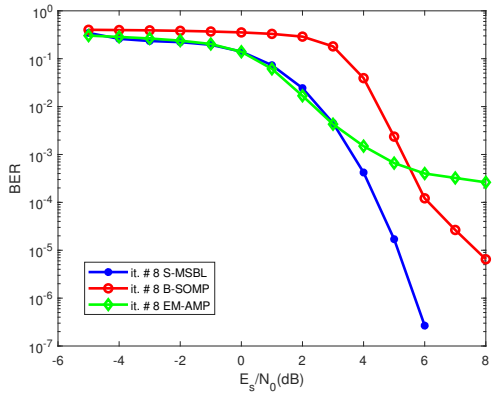

Figure 7.7: BER comparison of two-stage receivers with different initial joint UAD/CE.

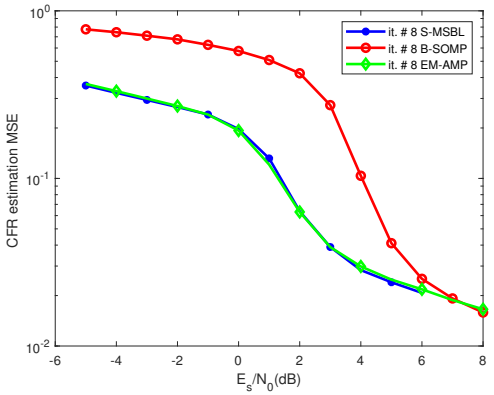

Figure 7.8: CFR MSE comparison of two-stage receivers with different initial joint UAD/CE.

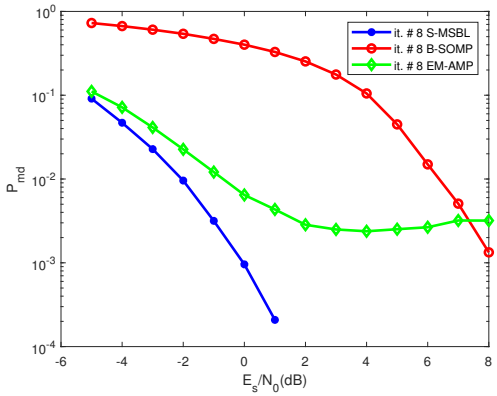

Figure 7.9: P_{md} comparison of two-stage receivers with different initial joint UAD/CE.

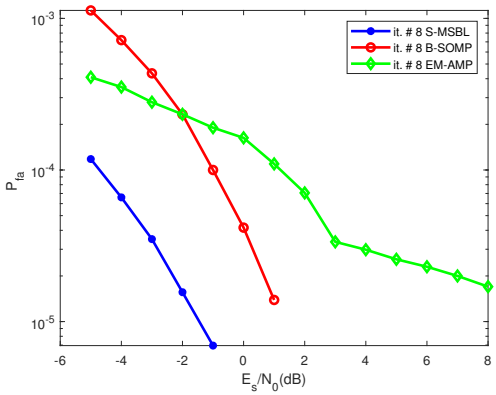

Figure 7.10: P_{fa} comparison of two-stage receivers with different initial joint UAD/CE.

At high SNR levels (e.g., ≥ 4 dB), the performance can be interpreted as follows:

First, a significant missed detection rate persists at the output of B-SOMP. For OFDM blocks affected by this issue, the EP receiver encounters residual IUI, leading to poor CE for some users in the first iteration. This, in turn, negatively impacts other tasks like MUD, DEC, and UAD as iterations progress. Second, EM-AMP experiences a high error floor in the false alarm rate. Although this is mitigated in later stages by the EP receiver, the inclusion of non-existent user channel estimates degrades the overall quality of the initial CE, which affects the EP stage. Lastly, the proposed initial S-MSBL exhibits negligible missed detection and false alarm rates. This ensures consistent joint CE for all users, providing an excellent foundation for the EP stage and enhancing performance.

Chapter 8

Conclusion and research perspectives

8.1 Conclusion

Grant-free access and NOMA are promising techniques to address the key challenges in 5G and beyond communication, such as meeting low-latency requirements, managing sporadic user activity, improving bandwidth utilization, and supporting high data rates. These technologies are crucial for applications involving a massive number of devices, particularly in the context of the IoT and URLLC.

The algorithms proposed in this thesis are applicable to any grant-free access wireless communication system that employs CD-NOMA, and they provide significant improvements in system performance. Notably, the hybrid EP/BP algorithm developed here surpasses other state-of-the-art algorithms, such as conventional BP, as demonstrated in Chapter 3.

Similarly, the Wirtinger calculus-based EP algorithm developed in this thesis outperforms existing methods that also use Wirtinger calculus-based EP. Its key advantage lies in effectively accounting for spatial antenna correlation, crucial for modern multi-antenna systems in (sub-)terahertz bandwidths considered for future 6G standards. This results in more accurate channel estimation and enhanced overall performance, particularly in scenarios with correlated fading channels. Additionally, the algorithm is designed to mitigate numerical instabilities, ensuring reliable convergence and robust performance even in complex environments. This not only improves UAD/CE/MUD/DEC accuracy but also enhances the receiver's overall efficiency.

In addition to its improved performance, the later algorithm excels in computational efficiency. They not only offer enhanced accuracy in UAD, CE and decoding but do so with lower computational complexity, making them well-suited for large-scale systems with stringent latency and processing requirements. This balance of performance and complexity is essential for the scalability and practical deployment of 5G and beyond communication systems.

Multiantenna grant-free access, which includes a large number of users with low-latency requirements and limited bandwidth, can be efficiently managed using the proposed approach. This approach involves pre-estimating user activity and channel conditions (UAD/CE), followed by the execution of the EP-based receiver. This two-stage strategy not only improves the system's ability to handle sporadic user activity and dense environments but also ensures robust channel estimation, leading to reliable communication in challenging conditions.

In conclusion, we believe that the contributions of this thesis provide significant advancements in the field of grant-free access with NOMA in the contexts under study. The proposed algorithms not only meet the rigorous demands of 5G systems but also lay the foundation for

future research and development in next-generation communication technologies.

8.2 Perspectives

8.2.1 Convergence analysis of the proposed algorithms

Extrinsic information transfer (EXIT) charts [129] [130] have been used with great success to better understand the convergence behavior and gain insight into the performance of iteratively decoded code combinations and, by extension, iterative 'turbo' receivers, without requiring time consuming Monte Carlo simulations. They are obtained by calculating transfer functions based on mutual information for the individual probabilistic modules involved in the receiver. EXIT charts have been used to analyze iterative demapping [131] and channel estimation [132]. Unlike DEM/DEC and CE, UAD is not easy to model and deserves special attention. Moreover, massive multiple access and iterative MUD using GaBP and EP form a high dimensional (nonlinear) system for which the application of the EXIT charts framework does not seem straightforward. If we can solve this problem, we will be able to better understand the dynamics of performance metrics of the algorithms proposed in the thesis and predict semi-analytically the maximum tolerable activity rates under a certain quality of service constraint.

8.2.2 Other system models for massive GF-NOMA

Correlated user activity. Wireless networks that support the use cases of massive MTC and URLLC are densely and massively populated. The existing algorithms - including those proposed in this thesis - assume that the activity of each transmitting device is homogeneous and independent, which is not the case in many applications (for example, because sensors observe a common phenomenon). To answer this question, [133] introduces a new flexible model taking into account heterogeneous group activity, using the framework of copula theory. It is then exploited by a hybrid generalized message pass-through algorithm to solve the problem of UAD and CE. It would be very interesting to take the ideas of this study and extend them to the algorithms proposed in this thesis.

Asynchronism. In its most general and demanding form, grant-free access is expected to be supported without any closed-loop time alignment signaling or predefined random access procedure. To account for this total lack of coordination among transmitting devices, we propose to complicate our system model and consider the case where the identities and the number of active users change dynamically during transmission, according to a stochastic model whose parameters are known or need to be learned, where the users' signals are asynchronously received at the destination, and where the channel state information at the receiver is either unknown or partially known, and may vary according to a time-evolving model, whose parameters are known or need to be learned. Various models and algorithms have been proposed to address some of these cases [135] [134] [22]. In particular, in [135], the authors aim to solve the problem of users' CE and DEM/DEC in a fully unsupervised way, without the need of signaling data. They employ a multiple access model based on an infinite factorial finite-state machine which does not impose any constraint on the number of users in the system, whether users are synchronized, whether they use a preamble, or whether their channel is known. More importantly, due to its non-parametric nature, this model allows the number of users to be unbounded. We believe that there is still some algorithmic research to be done to improve receiver performance in such challenging scenarios.

8.2.3 Unsourced massive random access

When one aims massive scalability, a new random access problem called unsourced massive access can be considered, where the receiver's task is to decode messages transmitted by a small fraction of active users on each transmission slot or resource block, whose identity

is not known at first, and who all use the same transmission protocol (including channel signaling and coding). In this context, senders who want to identify themselves must include their ID in the information message itself and the receiver's job is to decode the list of active user messages up to permutations. This new information-theoretic problem has been first formulated by [136] and later investigated in [137]. It poses a number of interesting problems in the design of the receiver in realistic communications scenarios, see for example [138] [139].

8.2.4 Deep learning in GF-NOMA

Recent research has shown the powerful capabilities of machine learning - especially deep learning - to improve the efficiency of transmitter/receiver models in wireless communications. In general, machine learning can solve NP-hard optimization problems faster, more accurately and more robustly than traditional approaches. Instead of relying on models and equations, machine learning algorithms look for patterns in data to make the best possible, almost optimal decisions. For example, in [140], a deep learning architecture, to effectively solve the joint UAD and CE problem for grant-free NOMA, by exploiting the framework of the compressive sensing-based algorithm. It would be interesting to compare the performance of this algorithm with those proposed in this thesis, or even to hybridize approaches based on different learning methods, i.e., deep learning and graphical probabilistic models, exploiting the strengths of each.

8.2.5 GF-NOMA and sensing

The proposed message-passing algorithms may be extended to support joint multi-user communication and sensing [141], where wireless devices not only transmit data but also actively monitor their environments. This integration opens up opportunities for applications such as autonomous vehicles, smart cities, and advanced industrial automation.

8.2.6 GF-NOMA and RIS

Additionally, the algorithms may be applied to [142],[143] Reconfigurable Intelligent Surfaces (RIS)s, an emerging technology that enables the dynamic control and manipulation of electromagnetic waves to enhance signal propagation. These advancements are poised to significantly shape future 6G networks, improving efficiency, signal reliability, and overall communication performance. AMP is a powerful iterative algorithm that can be employed for joint channel estimation, UAD, and data decoding in communication systems with large dimensions, such as in grant-free massive MIMO setups. In these systems, particularly relevant to future 6G networks, the need for efficient and scalable signal processing techniques is critical due to the vast number of users and the large volume of data being transmitted.

Appendix A

Projection operator for a continuous mixture of Gaussian distributions

Consider that we wish to project any conditional density $p(\mathbf{z}|\boldsymbol{\pi})$ to a desired Gaussian target density of the form $q(\mathbf{z}|\boldsymbol{\pi}) = \mathcal{CN}(\mathbf{z}; \mathbf{m}(\boldsymbol{\pi}), \boldsymbol{\Sigma})$, by minimizing the newly introduced criterion in Eq. (4.2).

Let us develop the criterion to be minimized as

$$\begin{aligned} E_{p(\boldsymbol{\pi})}[KL(p||q)] &= \int_{\boldsymbol{\pi}} \left[\int_{\mathbf{z}} p(\mathbf{z}|\boldsymbol{\pi}) \ln \frac{p(\mathbf{z}|\boldsymbol{\pi})}{q(\mathbf{z}|\boldsymbol{\pi})} d\mathbf{z} \right] p(\boldsymbol{\pi}) d\boldsymbol{\pi} \\ &= -\int_{\boldsymbol{\pi}} \left[\int_{\mathbf{z}} p(\mathbf{z}|\boldsymbol{\pi}) \ln q(\mathbf{z}|\boldsymbol{\pi}) d\mathbf{z} \right] p(\boldsymbol{\pi}) d\boldsymbol{\pi} \\ &\quad + \int_{\boldsymbol{\pi}} \left[\int_{\mathbf{z}} p(\mathbf{z}|\boldsymbol{\pi}) \ln p(\mathbf{z}|\boldsymbol{\pi}) d\mathbf{z} \right] p(\boldsymbol{\pi}) d\boldsymbol{\pi} \\ &= \int_{\boldsymbol{\pi}} \left[\int_{\mathbf{z}} -p(\mathbf{z}|\boldsymbol{\pi}) \ln q(\mathbf{z}|\boldsymbol{\pi}) d\mathbf{z} \right] p(\boldsymbol{\pi}) d\boldsymbol{\pi} + C. \end{aligned} \tag{A.1}$$

Indeed, since the minimization is performed w.r.t. $q(.|.)$, the second double integral in (Eq. (A.1)) can be considered as a constant independent of $\boldsymbol{\pi}$ that we call C. Also, replacing $-\ln q(\mathbf{z}|\boldsymbol{\pi})$ by its expresssion, we obtain

$$\begin{aligned} -\ln q(\mathbf{z}|\boldsymbol{\pi}) =& d \ln \pi + \ln |\boldsymbol{\Sigma}| + (\mathbf{z} - \mathbf{m}(\boldsymbol{\pi}))^H \boldsymbol{\Sigma}^{-1} (\mathbf{z} - \mathbf{m}(\boldsymbol{\pi})) \\ =& d \ln \pi + \ln |\boldsymbol{\Sigma}| + \mathbf{z}^H \boldsymbol{\Sigma}^{-1} \mathbf{z} - \mathbf{m}(\boldsymbol{\pi})^H \boldsymbol{\Sigma}^{-1} \mathbf{z} \\ & - \mathbf{z}^H \boldsymbol{\Sigma}^{-1} \mathbf{m}(\boldsymbol{\pi}) + \mathbf{m}(\boldsymbol{\pi})^H \boldsymbol{\Sigma}^{-1} \mathbf{m}(\boldsymbol{\pi}). \end{aligned} \tag{A.2}$$

Using the property $\mathrm{trace}(\mathbf{AB}) = \mathrm{trace}(\mathbf{BA})$ for any matrices of suitable dimensions **A** and **B**, we obtain an alternative expression as

$$-\ln q(\mathbf{z}|\boldsymbol{\pi}) = d \ln \pi + \ln |\boldsymbol{\Sigma}| + \mathrm{trace}(\boldsymbol{\Sigma}^{-1} (\mathbf{z} - \mathbf{m}(\boldsymbol{\pi}))(\mathbf{z} - \mathbf{m}(\boldsymbol{\pi}))^H). \tag{A.3}$$

Injecting (A.2) into the last line of (A.1), since $p(\boldsymbol{\pi}) \geq 0 \quad \forall \boldsymbol{\pi}$, optimization w.r.t. $\mathbf{m}(\boldsymbol{\pi})$ is equivalent to minimizing the inner integral inside the brackets, which results in solving

$$\int_{\mathbf{z}} \frac{\partial}{\partial \mathbf{m}(\boldsymbol{\pi})} \Big[-\mathbf{m}(\boldsymbol{\pi})^H \boldsymbol{\Sigma}^{-1} \mathbf{z} - \mathbf{z}^H \boldsymbol{\Sigma}^{-1} \mathbf{m}(\boldsymbol{\pi}) + \mathbf{m}(\boldsymbol{\pi})^H \boldsymbol{\Sigma}^{-1} \mathbf{m}(\boldsymbol{\pi}) \Big] p(\mathbf{z}|\boldsymbol{\pi}) d\mathbf{z} = \mathbf{0}, \tag{A.4}$$

whose unique solution is (see [84, Tab. 20.4])

$$\mathbf{m}(\boldsymbol{\pi}) = \int_{\mathbf{z}} \mathbf{z} p(\mathbf{z}|\boldsymbol{\pi}) d\mathbf{z}, \tag{A.5}$$

which in turn admits the closed form given in Eq. (4.3) for the continuous Gaussian mixture $p(\mathbf{z}|\boldsymbol{\pi})$ [87, p. 106-108].

Similarly, injecting (A.3) into the last line of (A.1), results in solving

$$\int_{\boldsymbol{\pi}} \left[\int_{\mathbf{z}} \frac{\partial}{\partial \boldsymbol{\Sigma}} \Big(\operatorname{trace}(\boldsymbol{\Sigma}^{-1}(\mathbf{z}-\mathbf{m}(\boldsymbol{\pi}))(\mathbf{z}-\mathbf{m}(\boldsymbol{\pi}))^H) + \ln|\boldsymbol{\Sigma}| \Big) p(\mathbf{z}|\boldsymbol{\pi}) d\mathbf{z} \right] p(\boldsymbol{\pi}) d\boldsymbol{\pi} = \mathbf{0}, \tag{A.6}$$

whose unique solution is (see [84, Tab. 20.4 and Tab. 20.5])

$$\boldsymbol{\Sigma} = \int_{\boldsymbol{\pi}} \left[\int_{\mathbf{z}} (\mathbf{z}-\mathbf{m}(\boldsymbol{\pi}))(\mathbf{z}-\mathbf{m}(\boldsymbol{\pi}))^H p(\mathbf{z}|\boldsymbol{\pi}) d\mathbf{z} \right] p(\boldsymbol{\pi}) d\boldsymbol{\pi}, \tag{A.7}$$

which in turn admits the closed form given in Eq. (4.3) for the continuous Gaussian mixture $p(\mathbf{z}|\boldsymbol{\pi})$ [87, p. 106-108].

Appendix B

Proof of the Wirtinger calculus based EP rule for UAD

Consider the the argument of the projection operator in (6.5)

$$H(\theta^{(u)}) \approx k \cdot \mu_{\theta^{(u)} \to g_n}(\theta^{(u)}) \tilde{q}(\mathbf{y}_n | \theta^{(u)}), \tag{B.1}$$

that we wish to project onto the set of densities Φ chosen as the family of circularly symmetric Gaussian densities $\mathcal{G}$, in the form $\mathcal{CN}(\theta^{(u)}; m_{\theta^{(u)}}, \sigma^2_{\theta^{(u)}})$. We thus need up to a constant C

$$\log H(\theta^{(u)}) \approx C - \frac{|\theta^{(u)}|^2}{\sigma^2_{\theta^{(u)}}} + 2\frac{\mathrm{Re}(\theta^{(u)} m^*_{\theta^{(u)}})}{\sigma^2_{\theta^{(u)}}}. \tag{B.2}$$

Against the background of [93] that introduces additional intermediate variables, we use a simpler direct second-order expansion of the log of (B.1) wrt $\theta^{(u)}$ around θ_0 using Eq. (6.4) [92, Eq. (99)]. Ignoring the term $\mathrm{Re}\{\frac{\partial}{\partial \theta^{(u)*}} \left(\frac{\partial \log H(\theta^{(u)})}{\partial \theta^{(u)}} \right)^* \Big|_{\theta^{(u)}=\theta_0} (\theta^{(u)*} - \theta_0^*)^2\}$ that will disappear under projection onto $\mathcal{G}$, we obtain

$$\begin{aligned} \log H(\theta^{(u)}) \approx & \log H(\theta_0) + 2\,\mathrm{Re}\left(\frac{\partial \log H(\theta^{(u)})}{\partial \theta^{(u)}} \Big|_{\theta^{(u)}=\theta_0} (\theta^{(u)} - \theta_0) \right) \\ & + \frac{\partial}{\partial \theta^{(u)}} \left(\frac{\partial \log H(\theta^{(u)})}{\partial \theta^{(u)}} \right)^* \Big|_{\theta^{(u)}=\theta_0} |\theta^{(u)} - \theta_0|^2 \end{aligned} \tag{B.3}$$

so that by identification with (B.2) we get

$$\begin{cases} \dfrac{1}{\sigma^2_{\theta^{(u)}}} = -\dfrac{\partial}{\partial \theta^{(u)}} \left(\dfrac{\partial \log H(\theta^{(u)})}{\partial \theta^{(u)}} \right)^* \Big|_{\theta^{(u)}=\theta_0} \\ \dfrac{m_{\theta^{(u)}}}{\sigma^2_{\theta^{(u)}}} = \dfrac{\theta_0}{\sigma^2_{\theta^{(u)}}} + \left(\dfrac{\partial \log H(\theta^{(u)})}{\partial \theta^{(u)}} \Big|_{\theta^{(u)}=\theta_0} \right)^*. \end{cases} \tag{B.4}$$

Now, rewriting $\mu_{\theta^{(u)} \to g_n}(\theta^{(u)})$ in the desired Gaussian form $\mathcal{CN}(\theta^{(u)}; m_{\theta^{(u)} \to g_n}, \sigma^2_{\theta^{(u)} \to g_n})$ and injecting the expression of (6.6) into (B.1) leads (up to some constant C) to

$$\begin{aligned} \log H(\theta^{(u)}) = & C - \frac{|\theta^{(u)} - m_{\theta^{(u)} \to g_n}|^2}{\sigma^2_{\theta^{(u)} \to g_n}} - \log\det(|\theta^{(u)}|^2 \mathbf{A}_n^{(u)} + \mathbf{B}_n^{(u)}) \\ & - \mathrm{trace}\{(\mathbf{y}_n - \mathbf{h}_{\theta^{(u)} \to g_n} \theta^{(u)} - \mathbf{I}_{d_n^{(u)} \to g_n})^H (|\theta^{(u)}|^2 \mathbf{A}_n^{(u)} + \mathbf{B}_n^{(u)})^{-1} (\mathbf{y}_n - \mathbf{h}_{\theta^{(u)} \to g_n} \theta^{(u)} - \mathbf{I}_{d_n^{(u)} \to g_n})\}. \end{aligned} \tag{B.5}$$

Using the identities [95, App. D.2.3 and D.2.4], $\forall(\mathbf{A},\mathbf{B})$

$$\begin{aligned}\frac{d\log\det(\mathbf{A}_n^{(u)}t+\mathbf{B}_n^{(u)})}{dt} &= \operatorname{trace}\{(\mathbf{A}_n^{(u)}t+\mathbf{B}_n^{(u)})^{-1}\mathbf{A}_n^{(u)}\}\\ \frac{d\operatorname{trace}\{(\mathbf{B}^H(\mathbf{A}_n^{(u)}t+\mathbf{B}_n^{(u)})^{-1}\mathbf{A}\}}{dt} &= -\operatorname{trace}\{\mathbf{B}^H(\mathbf{A}_n^{(u)}t+\mathbf{B}_n^{(u)})^{-1}\mathbf{A}_n^{(u)}(\mathbf{A}_n^{(u)}t+\mathbf{B}_n^{(u)})^{-1}\mathbf{A}\}\end{aligned} \tag{B.6}$$

and applying the product and chain rule of complex derivatives [96, pp. 405-413] leads to the missing first and second order derivatives of (B.5) expressed as follows

$$\begin{aligned}-\frac{\partial}{\partial\theta^{(u)}}\left(\frac{\partial\log H(\theta^{(u)})}{\partial\theta^{(u)}}\right)^* &= \frac{1}{\sigma^2_{\theta^{(u)}\to g_n}}\\ &+\operatorname{trace}\left\{(|\theta^{(u)}|^2\mathbf{A}_n^{(u)}+\mathbf{B}_n^{(u)})^{-1}\mathbf{A}_n^{(u)}(|\theta^{(u)}|^2\mathbf{A}_n^{(u)}+\mathbf{B}_n^{(u)})^{-1}\mathbf{B}_n^{(u)}\right\}\\ &+\mathbf{h}^H_{\theta^{(u)}\to g_n}(|\theta^{(u)}|^2\mathbf{A}_n^{(u)}+\mathbf{B}_n^{(u)})^{-1}\mathbf{B}_n^{(u)}(|\theta^{(u)}|^2\mathbf{A}_n^{(u)}+\mathbf{B}_n^{(u)})^{-1}\mathbf{h}_{\theta^{(u)}\to g_n}\\ &+(\mathbf{y}_n-\mathbf{I}_{d_n^{(u)}\to g_n})^H(|\theta^{(u)}|^2\mathbf{A}_n^{(u)}+\mathbf{B}_n^{(u)})^{-1}\mathbf{A}_n^{(u)}(|\theta^{(u)}|^2\mathbf{A}_n^{(u)}+\mathbf{B}_n^{(u)})^{-1}(\mathbf{y}_n-\mathbf{I}_{d_n^{(u)}\to g_n})\\ &-2(\mathbf{y}_n-\mathbf{I}_{d_n^{(u)}\to g_n}-\mathbf{h}_{\theta^{(u)}\to g_n}\theta^{(u)})^H(|\theta^{(u)}|^2\mathbf{A}_n^{(u)}+\mathbf{B}_n^{(u)})^{-1}\mathbf{A}_n^{(u)}(|\theta^{(u)}|^2\mathbf{A}_n^{(u)}+\mathbf{B}_n^{(u)})^{-1}\dots\\ &\times\mathbf{B}_n^{(u)}(|\theta^{(u)}|^2\mathbf{A}_n^{(u)}+\mathbf{B}_n^{(u)})^{-1}(\mathbf{y}_n-\mathbf{I}_{d_n^{(u)}\to g_n}-\mathbf{h}_{\theta^{(u)}\to g_n}\theta^{(u)})\\ \frac{\partial\log H(\theta^{(u)})}{\partial\theta^{(u)}}^* &= \frac{m_{\theta^{(u)}\to g_n}}{\sigma^2_{\theta^{(u)}\to g_n}}+\mathbf{h}^H_{\theta^{(u)}\to g_n}(|\theta^{(u)}|^2\mathbf{A}_n^{(u)}+\mathbf{B}_n^{(u)})^{-1}(\mathbf{y}_n-\mathbf{I}_{d_n^{(u)}\to g_n}-\mathbf{h}_{\theta^{(u)}\to g_n}\theta^{(u)})+\\ &\theta^{(u)}\Bigg(-\frac{1}{\sigma^2_{\theta^{(u)}\to g_n}}-\operatorname{trace}\left\{\mathbf{A}_n^{(u)}(|\theta^{(u)}|^2\mathbf{A}_n^{(u)}+\mathbf{B}_n^{(u)})^{-1}\right\}+\\ &\qquad(\mathbf{y}_n-\mathbf{I}_{d_n^{(u)}\to g_n}-\mathbf{h}_{\theta^{(u)}\to g_n}\theta^{(u)})^H(|\theta^{(u)}|^2\mathbf{A}_n^{(u)}+\mathbf{B}_n^{(u)})^{-1}\mathbf{A}_n^{(u)}(|\theta^{(u)}|^2\mathbf{A}_n^{(u)}+\mathbf{B}_n^{(u)})^{-1}\dots\\ &\qquad\times(\mathbf{y}_n-\mathbf{I}_{d_n^{(u)}\to g_n}-\mathbf{h}_{\theta^{(u)}\to g_n}\theta^{(u)})\Bigg).\end{aligned}$$

Importantly, our Wirtinger calculus method for projecting the pdf of a latent variable onto $\mathcal{G}$ allows $\mathbf{y}_n$ to be a vector, unlike [93].

Bibliography

[1] C. Bockelmann, "Iterative Soft Interference Cancellation for Sparse BPSK Signals," in IEEE Communications Letters, vol. 19, no. 5, pp. 855-858, May 2015.

[2] W. Kim, Y. Ahn and B. Shim, "Deep Neural Network-Based Active User Detection for Grant-Free NOMA Systems," in IEEE Transactions on Communications, vol. 68, no. 4, pp. 2143-2155, April 2020.

[3] H. F. Schepker, C. Bockelmann and A. Dekorsy, "Efficient Detectors for Joint Compressed Sensing Detection and Channel Decoding," in IEEE Transactions on Communications, vol. 63, no. 6, pp. 2249-2260, June 2015.

[4] B. Wang, L. Dai, T. Mir and Z. Wang, "Joint User Activity and Data Detection Based on Structured Compressive Sensing for NOMA," in IEEE Communications Letters, vol. 20, no. 7, pp. 1473-1476, July 2016.

[5] B. Wang, L. Dai, Y. Zhang, T. Mir and J. Li, "Dynamic Compressive Sensing-Based Multi-User Detection for Uplink Grant-Free NOMA," in IEEE Communications Letters, vol. 20, no. 11, pp. 2320-2323, Nov. 2016.

[6] Y. Mei et al., "Compressive Sensing-Based Joint Activity and Data Detection for Grant-Free Massive IoT Access," in IEEE Transactions on Wireless Communications, vol. 21, no. 3, pp. 1851-1869, March 2022.

[7] K. Senel and E. G. Larsson, "Grant-Free Massive MTC-Enabled Massive MIMO: A Compressive Sensing Approach," in IEEE Transactions on Communications, vol. 66, no. 12, pp. 6164-6175, Dec. 2018.

[8] A. Bayesteh, E. Yi, H. Nikopour and H. Baligh, "Blind detection of SCMA for uplink grant-free multiple-access," 2014 11th International Symposium on Wireless Communications Systems (ISWCS), 2014, pp. 853-857.

[9] Q. Zou, H. Zhang, D. Cai and H. Yang, "A Low-Complexity Joint User Activity, Channel and Data Estimation for Grant-Free Massive MIMO Systems," in IEEE Signal Processing Letters, vol. 27, pp. 1290-1294, 2020.

[10] A. T. Abebe and C. G. Kang, "Joint Channel Estimation and MUD for Scalable Grant-Free Random Access," in IEEE Communications Letters, vol. 23, no. 12, pp. 2229-2233, Dec. 2019.

[11] J. Ding and J. Choi, "Comparison of Preamble Structures for Grant-Free Random Access in Massive MIMO Systems," in IEEE Wireless Communications Letters, vol. 9, no. 2, pp. 166-170, Feb. 2020.

[12] T. Ding, X. Yuan and S. C. Liew, "Sparsity Learning-Based Multiuser Detection in Grant-Free Massive-Device Multiple Access," in IEEE Transactions on Wireless Communications, vol. 18, no. 7, pp. 3569-3582, July 2019.

[13] Y. Han, B. D. Rao and J. Lee, "Massive Uncoordinated Access With Massive MIMO: A Dictionary Learning Approach," in IEEE Transactions on Wireless Communications, vol. 19, no. 2, pp. 1320-1332, Feb. 2020.

[14] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen and L. Hanzo, "A Survey of Non-Orthogonal Multiple Access for 5G," in IEEE Communications Surveys & Tutorials, vol. 20, no. 3, pp. 2294-2323, thirdquarter 2018.

[15] J. Goseling, M. Gastpar and J. H. Weber, "Random Access With Physical-Layer Network Coding," in IEEE Transactions on Information Theory, vol. 61, no. 7, pp. 3670-3681, July 2015,

[16] E. Paolini, C. Stefanovic, G. Liva and P. Popovski, "Coded random access: applying codes on graphs to design random access protocols," in IEEE Communications Magazine, vol. 53, no. 6, pp. 144-150, June 2015.

[17] M. J. Wainwright and M. I. Jordan, "Graphical Models, Exponential Families, and Variational Inference", Foundations and Trends in Machine Learning: Vol. 1: No. 1–2, pp 1-305, 2008.

[18] S. Sharma, K. Deka and Y. Hong, "User Activity Detection-Based Large SCMA System for Uplink Grant-Free Access," 2019 IEEE International Conference on Communications Workshops (ICC Workshops), 2019, pp. 1-6.

[19] Y. Zhang, Q. Guo, Z. Wang, J. Xi and N. Wu, "Block Sparse Bayesian Learning Based Joint User Activity Detection and Channel Estimation for Grant-Free NOMA Systems," in IEEE Transactions on Vehicular Technology, vol. 67, no. 10, pp. 9631-9640, Oct. 2018.

[20] W. Zhu, M. Tao, X. Yuan and Y. Guan, "Message Passing-Based Joint User Activity Detection and Channel Estimation for Temporally-Correlated Massive Access," in IEEE Transactions on Communications, vol. 71, no. 6, pp. 3576-3591, June 2023.

[21] F. Lehmann, "Joint User Activity Detection, Channel Estimation, and Decoding for Multiuser/Multiantenna OFDM Systems," in IEEE Transactions on Vehicular Technology, vol. 67, no. 9, pp. 8263-8275, Sept. 2018

[22] W. Yuan, N. Wu, Q. Guo, D. W. K. Ng, J. Yuan and L. Hanzo, "Iterative Joint Channel Estimation, User Activity Tracking, and Data Detection for FTN-NOMA Systems Supporting Random Access," in IEEE Transactions on Communications, vol. 68, no. 5, pp. 2963-2977, May 2020.

[23] Z. Yuan, C. Zhang, Z. Wang, Q. Guo and J. Xi, "An Auxiliary Variable-Aided Hybrid Message Passing Approach to Joint Channel Estimation and Decoding for MIMO-OFDM," in IEEE Signal Processing Letters, vol. 24, no. 1, pp. 12-16, Jan. 2017.

[24] Y. Zhang, Z. Yuan, Q. Guo, Z. Wang, J. Xi and Y. Li, "Bayesian Receiver Design for Grant-Free NOMA With Message Passing Based Structured Signal Estimation," in IEEE Transactions on Vehicular Technology, vol. 69, no. 8, pp. 8643-8656, Aug. 2020.

[25] F. Wei, W. Chen, Y. Wu, J. Ma and T. A. Tsiftsis, "Message-Passing Receiver Design for Joint Channel Estimation and Data Decoding in Uplink Grant-Free SCMA Systems," in IEEE Transactions on Wireless Communications, vol. 18, no. 1, pp. 167-181, Jan. 2019.

[26] L. Sanguinetti, E. Björnson and J. Hoydis, "Toward Massive MIMO 2.0: Understanding Spatial Correlation, Interference Suppression, and Pilot Contamination," in IEEE Transactions on Communications, vol. 68, no. 1, pp. 232-257, Jan. 2020.

[27] M. J. Wainwright and M. I. Jordan, "Graphical Models, Exponential Families, and Variational Inference", Foundations and Trends in Machine Learning: Vol. 1: No. 1–2, pp 1-305, 2008.

[28] C. Knievel, P. A. Hoeher, A. Tyrrell and G. Auer, "Multi-Dimensional Graph-Based Soft Iterative Receiver for MIMO-OFDM," in IEEE Transactions on Communications, vol. 60, no. 6, pp. 1599-1609, June 2012.

[29] C. Novak, G. Matz and F. Hlawatsch, "IDMA for the Multiuser MIMO-OFDM Uplink: A Factor Graph Framework for Joint Data Detection and Channel Estimation," in IEEE Transactions on Signal Processing, vol. 61, no. 16, pp. 4051-4066, Aug.15, 2013.

[30] D. Bickson and D. Dolev, O. Shental, P.H. Siegel and J.K. Wolf, "Gaussian belief propagation based multiuser detection," Proc. ISIT 08, pp. 1878-1882, Toronto, Canada, Jul. 2008.

[31] W. Yuan, N. Wu, Q. Guo, D. W. K. Ng, J. Yuan and L. Hanzo, "Iterative Joint Channel Estimation, User Activity Tracking, and Data Detection for FTN-NOMA Systems Supporting Random Access," in IEEE Transactions on Communications, vol. 68, no. 5, pp. 2963-2977, May 2020.

[32] Deepender, Manoj, U. Shrivastava and J. K. Verma, "A Study on 5G Technology and Its Applications in Telecommunications," 2021 International Conference on Computational Performance Evaluation (ComPE), Shillong, India, 2021, pp. 365-371.

[33] X. Li, W. Xie and C. Hu, "Research on 5G URLLC Standard and Key Technologies," 2022 3rd Information Communication Technologies Conference (ICTC), Nanjing, China, 2022, pp. 243-249.

[34] P. Popovski, K. F. Trillingsgaard, O. Simeone and G. Durisi, "5G Wireless Network Slicing for eMBB, URLLC, and mMTC: A Communication-Theoretic View," in IEEE Access, vol. 6, pp. 55765-55779, 2018.

[35]

[36] M. Girmay et al., "Enabling Uncoordinated Dynamic Spectrum Sharing Between LTE and NR Networks," in IEEE Transactions on Wireless Communications, vol. 23, no. 6, pp. 5953-5968, June 2024.

[37] Khan, Koffka & Goodridge, Wayne. (2021). Ultra-HD Video Streaming in 5G Fixed Wireless Access Bottlenecks.

[38] "5G Technology and Networks (Speed, Use Cases, Rollout)." Thales Group, www.thalesgroup.com/en/markets/digital-identity-and-security/mobile/inspired/5G

[39] Zaidi, Ali, et al. "Whitepaper on Cellular IOT Connectivity in the 5G ERA." Ericsson, www.ericsson.com/en/reports-and-papers/white-papers/cellular-iot-in-the-5g-era. Accessed 20 Oct. 2024.

[40] Springer, Johannes. "5G Automotive Association, Pioneering Digital Transformation in the Automotive Industry." ITU, www.itu.int/en/ITU-T/extcoop/cits/Documents/Meeting-20190308-Geneva/14_5GAA-progress_report.pdf. Accessed 20 Oct. 2024.

[41] "Cisco Data Center Certifications." Cisco, 19 July 2024, www.cisco.com/site/us/en/learn/training-certifications/certifications/datacenter/index.html.

[42] 3GPP Low Power Wide Area Technologies, www.gsma.com/solutions-and-impact/technologies/internet-of-things/wp-content/uploads/2016/10/3GPP-Low-Power-Wide-Area-Technologies-GSMA-White-Paper.pdf. Accessed 14 Oct. 2024.

[43] V. Saxena, J. Bergman, Y. Blankenship, A. Wallen and H. S. Razaghi, "Reducing the Modem Complexity and Achieving Deep Coverage in LTE for Machine-Type Communications," 2016 IEEE Global Communications Conference (GLOBECOM), Washington, DC, USA, 2016, pp. 1-7.

[44] "Procedures for the 5G System (3GPP TS 23.502 version 15.2.0 Release 15)." ETSI, June 2018, www.etsi.org/deliver/etsi_ts/123500_123599/123502/15.02.00_60/ts_123502v150200p.pdf.

[45] Atterwall, Cecilia, and Eva Hedfors. “Late for 5G Launch: Non-Launched Operators?” Ericsson, Apr. 2023, www.ericsson.com/en/blog/2023/4/are-service-providers-that-havent-launched-5g-yet-running-late.

[46] “Technical Realization of Service Based Architecture (3GPP TS 29.500 version 16.4.0 Release 16).” ETSI, Nov. 2020, www.etsi.org/deliver/etsi_ts/129500_129599/129500/16.04.00_60/ts_129500v160400p.pdf.

[47] J. Choi, J. Ding, N. -P. Le and Z. Ding, ”Grant-Free Random Access in Machine-Type Communication: Approaches and Challenges,” in IEEE Wireless Communications, vol. 29, no. 1, pp. 151-158, February 2022.

[48] Z. Zhao, Q. Du and G. K. Karagiannidis, ”Improved Grant-Free Access for URLLC via Multi-Tier-Driven Computing: Network-Load Learning, Prediction, and Resource Allocation,” in IEEE Journal on Selected Areas in Communications, vol. 41, no. 3, pp. 607-622, March 2023.

[49] ‘’Beyond Netflix: 5G and the Internet of Things.” Thales Group, 6 Dec. 2021, www.thalesgroup.com/en/worldwide/group/magazine/beyond-netflix-5g-and-internet-things.

[50] Grami, Ali. Introduction to Digital Communications. Academic Press, 2016.

[51] Torrieri, Don. Principles of Spread-Spectrum Communication Systems. Springer International Publishing, 2018.

[52] Q. Wu et al., “A Comprehensive Overview on 5G-and-Beyond Networks With UAVs: From Communications to Sensing and Intelligence,” in IEEE Journal on Selected Areas in Communications, vol. 39, no. 10, pp. 2912-2945, Oct. 2021.

[53] Haesik Kim, ”Massive Machine Type Communication Systems,” in Design and Optimization for 5G Wireless Communications , IEEE, 2020, pp.343-395.

[54] Vaezi, Mojtaba, et al. Multiple Access Techniques for 5G Wireless Networks and Beyond. Springer, 2019.

[55] Almudayni, Ziyad, et al.“A comprehensive study on the energy efficiency of IOT from four angles: Clustering and routing in WSNS, Smart Grid, fog computing and MQTT & Coap Application Protocols.” Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, 2022, pp. 54–70.

[56] Yuan, Zhifeng, et al. “Multi-user shared access for internet of things.” 2016 IEEE 83rd Vehicular Technology Conference (VTC Spring), May 2016.

[57] Majumdar, P., Bhattacharya, D., Mitra, S. (2023). Utilities of 5G Communication Technologies for Promoting Advancement in Agriculture 4.0: Recent Trends, Research Issues and Review of Literature. In: Bhushan, B., Sharma, S.K., Kumar, R., Priyadarshini, I. (eds) 5G and Beyond. Springer Tracts in Electrical and Electronics Engineering. Springer, Singapore.

[58] S. M. R. Islam, N. Avazov, O. A. Dobre and K.-S. Kwak, ”Power-Domain Non-Orthogonal Multiple Access (NOMA) in 5G Systems: Potentials and Challenges,” in IEEE Communications Surveys & Tutorials, vol. 19, no. 2, pp. 721-742, Secondquarter 2017.

[59] ”Network Slicing for 5G Networks,” in 5G Networks: Fundamental Requirements, Enabling Technologies, and Operations Management , IEEE, 2018, pp.327-370.

[60] B. D. Van Veen and K. M. Buckley, ”Beamforming: a versatile approach to spatial filtering,” in IEEE ASSP Magazine, vol. 5, no. 2, pp. 4-24, April 1988.

[61] W. van Etten, ”Maximum Likelihood Receiver for Multiple Channel Transmission Systems,” in IEEE Transactions on Communications, vol. 24, no. 2, pp. 276-283, February 1976.

[62] Biglieri, Ezio. MIMO Wireless Communications. Cambridge University Press, 2010.

[63] M. B. Shahab, R. Abbas, M. Shirvanimoghaddam and S. J. Johnson, "Grant-Free Non-Orthogonal Multiple Access for IoT: A Survey," in IEEE Communications Surveys & Tutorials, vol. 22, no. 3, pp. 1805-1838, thirdquarter 2020.

[64] R. Hoshyar, F. P. Wathan and R. Tafazolli, "Novel Low-Density Signature for Synchronous CDMA Systems Over AWGN Channel," in IEEE Transactions on Signal Processing, vol. 56, no. 4, pp. 1616-1626, April 2008.

[65] H. Nikopour and H. Baligh, "Sparse code multiple access," in Proc. IEEE 24th Annu. Int. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC), London, U.K., Sep. 2013, pp. 332–336.

[66] S. Chen, B. Ren, Q. Gao, S. Kang, S. Sun and K. Niu, "Pattern Division Multiple Access—A Novel Nonorthogonal Multiple Access for Fifth-Generation Radio Networks," in IEEE Transactions on Vehicular Technology, vol. 66, no. 4, pp. 3185-3196, April 2017.

[67] Z. Yuan, G. Yu, W. Li, Y. Yuan, X. Wang and J. Xu, "Multi-User Shared Access for Internet of Things," 2016 IEEE 83rd Vehicular Technology Conference (VTC Spring), Nanjing, China, 2016, pp. 1-5.

[68] L. Ping, L. Liu, K. Wu and W. K. Leung, "Interleave division multiple-access," in IEEE Transactions on Wireless Communications, vol. 5, no. 4, pp. 938-947, April 2006.

[69] Yedidia, J.S., Freeman, W.T., Weiss, Y., 2003. Understanding belief propagation and its generalizations. Exploring artificial intelligence in the new millennium 8, 236–239.

[70] Pearl, Judea. "Reverend Bayes on Inference Engines: A Distributed Hierarchical Approach." *Proceedings of the Second AAAI Conference on Artificial Intelligence (AAAI'82)*, AAAI Press, 1982, pp. 133-136.

[71] Pearl, Judea. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publishers Inc., 1988.

[72] Y. Weiss, "Correctness of Local Probability Propagation in Graphical Models with Loops," in Neural Computation, vol. 12, no. 1, pp. 1-41, 1 Jan. 2000.

[73] T.P. Minka, "Expectation propagation for approximate," Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence (UAI2001), 2001, pp. 362–369.

[74] T. Minka, "Divergence Measures and Message Passing," Microsoft Resaerch Cambridge, MSR-TR-2005-173, pp. 1-17, January 2005.

[75] T. . -S. Chu and L. J. Greenstein, "A semi-empirical representation of antenna diversity gain at cellular and PCS base stations," in IEEE Transactions on Communications, vol. 45, no. 6, pp. 644-646, June 1997.

[76] F. R. Kschischang, B. J. Frey and H.-.A Loeliger, "Factor graphs and the sum-product algorithm," in IEEE Transactions on Information Theory, vol. 47, no. 2, pp. 498-519, Feb 2001.

[77] N. Czink, B. Bandemer, C. Oestges, T. Zemen and A. Paulraj, "Analytical Multi-User MIMO Channel Modeling: Subspace Alignment Matters," in IEEE Transactions on Wireless Communications, vol. 11, no. 1, pp. 367-377, January 2012.

[78] G. L. Stüber, Principles of Mobile Communication (2nd Ed.), Norwell, MA: Kluwer Academic Publishers, 2001.

[79] Kai Yu, M. Bengtsson, B. Ottersten, D. McNamara, P. Karlsson and M. Beach, "Modeling of wide-band MIMO radio channels based on NLoS indoor measurements," in IEEE Transactions on Vehicular Technology, vol. 53, no. 3, pp. 655-665, May 2004.

[80] L. Ping, Q. Guo and J. Tong, "The OFDM-IDMA approach to wireless communication systems," in IEEE Wireless Communications, vol. 14, no. 3, pp. 18-24, June 2007.

[81] P. Hammarberg, F. Rusek and O. Edfors, "Channel Estimation Algorithms for OFDM-IDMA: Complexity and Performance," in IEEE Transactions on Wireless Communications, vol. 11, no. 5, pp. 1722-1732, May 2012.

[82] C. Novak, G. Matz and F. Hlawatsch, "IDMA for the Multiuser MIMO-OFDM Uplink: A Factor Graph Framework for Joint Data Detection and Channel Estimation," in IEEE Transactions on Signal Processing, vol. 61, no. 16, pp. 4051-4066, Aug.15, 2013.

[83] S. Kim, H. Kim, H. Noh, Y. Kim and D. Hong, "Novel Transceiver Architecture for an Asynchronous Grant-Free IDMA System," in IEEE Transactions on Wireless Communications, vol. 18, no. 9, pp. 4491-4504, Sept. 2019.

[84] P. Vaidyanathan, S. Phoong and Y. Lin, Signal Processing and Optimization for Transceiver Systems, Cambridge: Cambridge University Press, 2010.

[85] Shynk, John Joseph, Probability, Random Variables, and Random Processes: Theory and Signal Processing Applications, Hoboken, N.J.: John Wiley & Sons, 2013.

[86] D.C. Fraser and J.E. Potter, "The optimum linear smoother as a combination of two optimum linear filters," IEEE Trans. Automat. Contr., pp. 387-390, vol. 14, no. 4, Aug. 1969.

[87] H. Tanizaki, Nonlinear filters: estimation and applications, Berlin, Germany: Springer, 1996.

[88] M. Senst and G. Ascheid, "How the Framework of Expectation Propagation Yields an Iterative IC-LMMSE MIMO Receiver," 2011 IEEE Global Telecommunications Conference - GLOBECOM 2011, 2011, pp. 1-6.

[89] M. Ruan, M.C. Reed, and Z. Shi, "Successive multiuser detection and interference cancelation for contention based OFDMA ranging channel," IEEE Trans. Wireless Commun., vol. 9, no. 2, pp. 481-487, Feb. 2010.

[90] Q. Wang and G. Ren, "Iterative maximum likelihood detection for initial ranging process in 802.16 OFDMA systems," IEEE Trans. Wireless Commun., vol. 14, no. 5, pp. 2778-2787, May 2015.

[91] J. Dauwels and H.-A. Loeliger, "Phase estimation by message passing," 2004 IEEE International Conference on Communications (ICC), Paris, France, 2004, pp. 523-527 Vol.1.

[92] K. Kreutz-Delgado, "The complex gradient operator and the CR-calculus." , 2009 [Online] Available: `https://arxiv.org/abs/0906.4835`.

[93] S. Wu, L. Kuang, Z. Ni, D. Huang, Q. Guo and J. Lu, "Message-Passing Receiver for Joint Channel Estimation and Decoding in 3D Massive MIMO-OFDM Systems," IEEE Transactions on Wireless Communications, vol. 15, no. 12, pp. 8122-8138, Dec. 2016.

[94] Hjørungnes, Are. Complex-Valued Matrix Derivatives: With Applications in Signal Processing and Communications. Cambridge University Press, 2011.

[95] J. Dattorro, Convex optimization & Euclidian geometry, Palo Alto, CA: Meboo Publishing, 2005.

[96] R. F. H. Fischer, Precoding and Signal Shaping for Digital Transmission, New York, NY: John Wiley & Sons, 2002.

[97] S. Bochkanov. Bound and linear equality/inequality constrained optimization. (Online; accessed in April 2023): `http://www.alglib.net/optimization/boundandlinearlyconstrained.php`.

[98] S. R. Pokhrel, J. Ding, J. Park, O. -S. Park and J. Choi, “Towards Enabling Critical mMTC: A Review of URLLC Within mMTC,” in IEEE Access, vol. 8, pp. 131796-131813, 2020.

[99] A. Bayesteh, E. Yi, H. Nikopour and H. Baligh, “Blind detection of SCMA for uplink grant-free multiple-access,” 2014 11th International Symposium on Wireless Communications Systems (ISWCS), Barcelona, Spain, 2014, pp. 853-857.

[100] J. Ding and J. Choi, “Comparison of Preamble Structures for Grant-Free Random Access in Massive MIMO Systems,” in IEEE Wireless Communications Letters, vol. 9, no. 2, pp. 166-170, Feb. 2020.

[101] A. T. Abebe and C. G. Kang, “Joint Channel Estimation and MUD for Scalable Grant-Free Random Access,” in IEEE Communications Letters, vol. 23, no. 12, pp. 2229-2233, Dec. 2019.

[102] L. Liu and W. Yu, ”Massive Connectivity With Massive MIMO—Part I: Device Activity Detection and Channel Estimation,” in IEEE Transactions on Signal Processing, vol. 66, no. 11, pp. 2933-2946, 1 June1, 2018.

[103] K. Senel and E. G. Larsson, “Grant-Free Massive MTC-Enabled Massive MIMO: A Compressive Sensing Approach,” in IEEE Transactions on Communications, vol. 66, no. 12, pp. 6164-6175, Dec. 2018.

[104] Q. Zou, H. Zhang, D. Cai and H. Yang, “Message Passing Based Joint Channel and User Activity Estimation for Uplink Grant-Free Massive MIMO Systems With Low-Precision ADCs,” in IEEE Signal Processing Letters, vol. 27, pp. 506-510, 2020.

[105] Q. Zou, H. Zhang, D. Cai and H. Yang, “A Low-Complexity Joint User Activity, Channel and Data Estimation for Grant-Free Massive MIMO Systems,” in IEEE Signal Processing Letters, vol. 27, pp. 1290-1294, 2020.

[106] H. Djelouat, L. Marata, M. Leinonen, H. Alves and M. Juntti, “User Activity Detection and Channel Estimation of Spatially Correlated Channels via AMP in Massive MTC,” 2021 55th Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, USA, 2021, pp. 1200-1204.

[107] O. Y. Feng, R. Venkataramanan, C. Rush, R. J. Samworth, “A Unifying Tutorial on Approximate Message Passing,” Foundations and Trends in Machine Learning, vol. 15, no. 4, pp. 335–536, 2022.

[108] Y. Cheng, L. Liu and L. Ping, “Orthogonal AMP for Massive Access in Channels With Spatial and Temporal Correlations,” in IEEE Journal on Selected Areas in Communications, vol. 39, no. 3, pp. 726-740, March 2021.

[109] S. Srivastava, C. S. K. Patro, A. K. Jagannatham and L. Hanzo, “Sparse, Group-Sparse, and Online Bayesian Learning Aided Channel Estimation for Doubly-Selective mmWave Hybrid MIMO OFDM Systems,” in IEEE Transactions on Communications, vol. 69, no. 9, pp. 5843-5858, Sept. 2021.

[110] Y. Zhu et al., “OFDM-Based Massive Grant-Free Transmission Over Frequency-Selective Fading Channels,” in IEEE Transactions on Communications, vol. 70, no. 7, pp. 4543-4558, July 2022.

[111] B. Tahir, S. Schwarz and M. Rupp, “Impact of Channel Correlation on Subspace-Based Activity Detection in Grant-Free NOMA,” 2022 IEEE 95th Vehicular Technology Conference: (VTC2022-Spring), Helsinki, Finland, 2022, pp. 1-6.

[112] U. K. Ganesan, E. Björnson and E. G. Larsson, “Clustering-Based Activity Detection Algorithms for Grant-Free Random Access in Cell-Free Massive MIMO,” in IEEE Transactions on Communications, vol. 69, no. 11, pp. 7520-7530, Nov. 2021.

[113] W. Jiang, Y. Jia and Y. Cui, “Statistical Device Activity Detection for OFDM-Based Massive Grant-Free Access,” in IEEE Transactions on Wireless Communications, vol. 22, no. 6, pp. 3805-3820, June 2023.

[114] Y. Cui, S. Li and W. Zhang, “Jointly Sparse Signal Recovery and Support Recovery via Deep Learning With Applications in MIMO-Based Grant-Free Random Access,” in IEEE Journal on Selected Areas in Communications, vol. 39, no. 3, pp. 788-803, March 2021.

[115] Y. Zou, Z. Qin and Y. Liu, “Joint User Activity and Data Detection in Grant-Free NOMA using Generative Neural Networks,” ICC 2021 - IEEE International Conference on Communications, Montreal, QC, Canada, 2021, pp. 1-6.

[116] J. W. Choi, B. Shim, Y. Ding, B. Rao and D. I. Kim, “Compressed Sensing for Wireless Communications: Useful Tips and Tricks,” in IEEE Communications Surveys & Tutorials, vol. 19, no. 3, pp. 1527-1550, thirdquarter 2017.

[117] H. F. Schepker, C. Bockelmann and A. Dekorsy, “Efficient Detectors for Joint Compressed Sensing Detection and Channel Decoding,” in IEEE Transactions on Communications, vol. 63, no. 6, pp. 2249-2260, June 2015.

[118] Q. He, T. Q. S. Quek, Z. Chen, Q. Zhang and S. Li, “Compressive Channel Estimation and Multi-User Detection in C-RAN With Low-Complexity Methods,” in IEEE Transactions on Wireless Communications, vol. 17, no. 6, pp. 3931-3944, June 2018.

[119] J. Ahn, B. Shim and K. B. Lee, “EP-Based Joint Active User Detection and Channel Estimation for Massive Machine-Type Communications,” in IEEE Transactions on Communications, vol. 67, no. 7, pp. 5178-5189, July 2019.

[120] Q. Zou, H. Yang, “A Concise Tutorial on Approximate Message Passing.”, 2022 [Online] Available: `https://arxiv.org/abs/2201.07487`.

[121] D. P. Wipf and B. D. Rao, “An Empirical Bayesian Strategy for Solving the Simultaneous Sparse Approximation Problem,” in IEEE Transactions on Signal Processing, vol. 55, no. 7, pp. 3704-3716, July 2007.

[122] Z. Zhang and B. D. Rao, “Sparse Signal Recovery With Temporally Correlated Source Vectors Using Sparse Bayesian Learning,” in IEEE Journal of Selected Topics in Signal Processing, vol. 5, no. 5, pp. 912-926, Sept. 2011.

[123] N. Han, Z. Song, “Bayesian multiple measurement vector problem with spatial structured sparsity patterns,” in Digital Signal Processing, vol. 75, pp 184-201, 2018.

[124] Dempster, A. P., et al. “Maximum likelihood from incomplete data via the EM algorithm.” Journal of the Royal Statistical Society Series B: Statistical Methodology, vol. 39, no. 1, 1 Sept. 1977, pp. 1–22.

[125] H. Wymeersch, Iterative receiver designs, Cambridge, UK: Cambridge University Press, 2007.

[126] H. Ma, X. Yuan, L. Zhou, B. Li and R. Qin, ”Joint Block Support Recovery for Sub-Nyquist Sampling Cooperative Spectrum Sensing,” in IEEE Wireless Communications Letters, vol. 12, no. 1, pp. 85-88, Jan. 2023.

[127] F. Sagheer, F. Lehmann and A. O. Berthet, ”A New Hybrid Message Passing Algorithm for Joint User Activity Detection, Channel Estimation and Data Decoding in Grant-Free OFDM-IDMA,” in IEEE Transactions on Vehicular Technology, vol. 73, no. 7, pp. 10365-10380, July 2024.

[128] J. Vila and P. Schniter, “Expectation-maximization Bernoulli-Gaussian approximate message passing,” 2011 Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR), Pacific Grove, CA, USA, 2011, pp. 799-803.

[129] S. ten Brink, “Convergence of iterative decoding,” in Electronic Letters, vol. 35, no. 10, pp. 806–808, May 1999.

[130] S. ten Brink, “Convergence behavior of iteratively decoded parallel concatenated codes,” in IEEE Transactions on Communications, vol. 49, no. 10, pp. 1727–1737, Oct. 2001.

[131] S. ten Brink, J. Speidel, R. Yan, “Iterative demapping and decoding for multilevel modulation,” Proc. IEEE GLOBECOM 1998, Washington, Sydney, Australia, Nov. 1998.

[132] D.P. Sheperd, Z. Shi, M. Anderson, M.C. Reed, “EXIT chart analysis of an iterative receiver with channel estimation,” Proc. IEEE GLOBECOM 2007, Washington, DC, USA, Nov. 2007.

[133] L. Chetot, M. Egan and J. -M. Gorce, ”Hybrid Generalized Approximate Message Passing for Active User Detection and Channel Estimation With Correlated Group-Heterogeneous Activity,” in IEEE Transactions on Communications, vol. 72, no. 7, pp. 3919-3933, July 2024.

[134] S. Kim, H. Kim, H. Noh, Y. Kim and D. Hong, ”Novel Transceiver Architecture for an Asynchronous Grant-Free IDMA System,” in IEEE Transactions on Wireless Communications, vol. 18, no. 9, pp. 4491-4504, Sept. 2019.

[135] F.J.R Ruiz, I. Valera, L. Svensson, F. Perez-Cruz, “Infinite factorial finite state machine for blind multiuser channel estimation,” in IEEE Transactions on Communications, vol. 4, no. 3, pp. 177-191, Jun. 2018.

[136] Y. Polyanskiy, ”A perspective on massive random-access,” 2017 IEEE International Symposium on Information Theory (ISIT), Aachen, Germany, 2017, pp. 2523-2527.

[137] K.-H. Ngo, A. Lancho, G. Durisi and A. Graell i Amat, ”Unsourced Multiple Access With Random User Activity,” in IEEE Transactions on Information Theory, vol. 69, no. 7, pp. 4537-4558, July 2023.

[138] X. Liu, P. P. Cobo and R. Venkataramanan, ”Many-user multiple access with random user activity,” 2024 IEEE International Symposium on Information Theory (ISIT), Athens, Greece, 2024, pp. 2993-2998.

[139] M. Ozates, M. Kazemi and T. M. Duman, ”A Slotted Pilot-Based Unsourced Random Access Scheme With a Multiple-Antenna Receiver,” in IEEE Transactions on Wireless Communications, vol. 23, no. 4, pp. 3437-3449, April 2024.

[140] H. Yu, Z. Fei, Z. Zheng, N. Ye and Z. Han, ”Deep Learning-Based User Activity Detection and Channel Estimation in Grant-Free NOMA,” in IEEE Transactions on Wireless Communications, vol. 22, no. 4, pp. 2202-2214, April 2023.

[141] X. Tong, Z. Zhang, J. Wang, C. Huang and M. Debbah, ”Joint Multi-User Communication and Sensing Exploiting Both Signal and Environment Sparsity,” in IEEE Journal of Selected Topics in Signal Processing, vol. 15, no. 6, pp. 1409-1422, Nov. 2021.

[142] L. Wei, C. Huang, G. C. Alexandropoulos, C. Yuen, Z. Zhang and M. Debbah, ”Channel Estimation for RIS-Empowered Multi-User MISO Wireless Communications,” in IEEE Transactions on Communications, vol. 69, no. 6, pp. 4144-4157, June 2021.

[143] L. Wei et al., ”Joint Channel Estimation and Signal Recovery for RIS-Empowered Multiuser Communications,” in IEEE Transactions on Communications, vol. 70, no. 7, pp. 4640-4655, July 2022.

[144] A. Sahbafard, R. Muzaffar, R. Schmidt, Florian-Kaltenberger, A. Springer and H. -P. Bernhard, ”Enhanced Modeling of Uplink Configured Grant Transmissions for URLLC,” GLOBECOM 2023 - 2023 IEEE Global Communications Conference, Kuala Lumpur, Malaysia, 2023, pp. 6062-6066.

ECOLE DOCTORALE

Titre: Méthodes statistiques bayésiennes pour la détection conjointe des activités des utilisateurs, l'estimation des canaux et le décodage des données dans les réseaux sans fil dynamiques

Mots clés: NOMA, grant-free, accès massif, OFDM, graphes factoriels, algorithmes à passage de messages, propagation de croyance, propagation de l'espérance.

Résumé: L'accès multiple non orthogonal grant-free (GF-NOMA) devient un élément crucial des futurs systèmes d'accès radio, améliorant l'efficacité spectrale et permettant des communications ultra-fiables à faible latence. Cependant, la technique GF-NOMA introduit un nouveau défi : la détection de l'activité des utilisateurs. Le récepteur de la station de base doit classer les utilisateurs en utilisateurs actifs (qui émettent) ou inactifs.

Cette thèse se concentre sur le développement de nouvelles méthodes statistiques basées sur des algorithmes à passage de messages sur des graphes factoriels pour gérer conjointement les tâches d'estimation au niveau du récepteur. Les méthodes comprennent une inférence bayésienne hybride basée sur l'algorithme de propagation de croyance (BP) et l'algorithme de propagation d'espérance (EP), et un nouvel algorithme bayésien EP basé sur une approximation par calcul de Wirtinger. De plus, un récepteur à deux étages est conçu, utilisant une méthode d'acquisition bayésienne compressée pour l'estimation initiale du canal et de l'activité des utilisateurs en accès massif avec des séquences pilotes non orthogonales. Les simulations de ces méthodes mettent en évidence des performances prometteuses par rapport aux méthodes traditionnelles.

Title: Joint user activity detection, channel estimation and data decoding in dynamic wireless networks

Keywords: NOMA, grant-free, massive access, OFDM, factor graphs, message passing algorithms, belief propagation, expectation propagation.

Abstract: Grant-free non-orthogonal multiple access (GF-NOMA) is becoming a crucial part of future radio access systems, improving spectral efficiency and enabling ultra-reliable low latency communications. However, GF-NOMA introduces a new challenge: user activity detection. The base station receiver must classify users into active (i.e. transmitting) and non-active ones.

This thesis focuses on developing new statistical methods based on message passing algorithms on factor graphs to jointly handle all estimation tasks at the receiver level. The methods include hybrid Bayesian inference based on the belief propagation algorithm (BP) and the expectation propagation algorithm (EP), and a new Bayesian EP algorithm based on a Wirtinger calculus approximation. Moreover a two-stage receiver is designed, using a Bayesian compressed acquisition method for initial channel and user activity estimation in massive access with non-orthogonal pilot sequences. Simulations of these methods show promising performance compared to traditional methods.

Institut Polytechnique de Paris
91120 Palaiseau, France