DESIGN OF LOW COMPLEXITY UNSOURCED RANDOM ACCESS SCHEMES OVER WIRELESS CHANNELS

A DISSERTATION SUBMITTED TO

THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR

THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

ELECTRICAL AND ELECTRONICS ENGINEERING

By

Mert Özateş

November 2023

Design of Low Complexity Unsourced Random Access Schemes Over Wireless Channels

By Mert Özateş

November 2023

We certify that we have read this dissertation and that in our opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.

Tolga Mete Duman (Advisor)

Sinan Gezici

Erdal Arıkan

Ali Özgür Yılmaz

Ayşe Melda Yüksel Turgut

Approved for the Graduate School of Engineering and Science:

Orhan Arıkan
Director of the Graduate School

ABSTRACT

DESIGN OF LOW COMPLEXITY UNSOURCED RANDOM ACCESS SCHEMES OVER WIRELESS CHANNELS

Mert Özateş
Ph.D. in Electrical and Electronics Engineering
Advisor: Tolga Mete Duman
November 2023

The Sixth Generation and Beyond communication systems are expected to enable communications of a massive number of machine-type devices. The traffic generated by some of these devices will significantly deviate from those in conventional communication scenarios. For instance, for applications where a massive number of cheap sensors communicate with a base station (BS), the devices will only be sporadically active and there will be no coordination among them or with the BS. For such systems requiring massive random access solutions, a new paradigm called unsourced random access (URA) has recently been proposed. In URA, all the users employ the same codebook and there is no user identity. The destination is only interested in the list of messages being sent from the set of active users. While there are many interesting URA schemes developed in the recent literature, many significant challenges remain, in particular in designing low-complexity and energy-efficient solutions.

With the motivation of addressing the current challenges in URA, we develop practical solutions for several scenarios. First, we propose and study URA over frequency-selective channels via orthogonal frequency division multiplexing to mitigate the fading effects. The decoder employs a joint activity detection and channel estimation algorithm coupled with treating interference as noise and successive interference cancellation (SIC). Our results show that the proposed scheme offers competitive performance with grant-based frequency division multiple-access while the performance loss due to the estimated channel state information is limited. We then examine the scenario for which the receiver is equipped with a massive number of antennas and develop a simple yet energy-efficient solution by dividing the transmission frame into slots where each active user utilizes a non-orthogonal pilot sequence followed by its polar encoded codeword. At the receiver, we first detect the transmitted pilot sequences by a generalized orthogonal matching pursuit algorithm and utilize a linear minimum

mean square error (LMMSE) solution to estimate the channel vectors. We then perform iterative decoding based on maximal ratio combining and single-user decoding followed by SIC. Numerical examples and analysis results demonstrate that the proposed scheme either outperforms the existing approaches in the literature or has a competitive performance with lower complexity. We then adapt our solution to the scenarios with residual hardware impairments (HWIs) at the BS and the user equipment sides by developing a hardware-impairment aware LMMSE solution for channel estimation using the HWI statistics and observe that the newly proposed solution improves the energy efficiency and increases the number of supported active users. Finally, we study on-off division multiple access in the context of URA where each active user utilizes a small fraction of the transmission frame and show that the new approach is superior to the existing ones in terms of performance or complexity.

Keywords: Unsourced random access, polar codes, massive random access, frequency-selective channels, multiple-input-multiple-output systems, generalized orthogonal matching pursuit, hardware impairments, on-off division multiple access.

ÖZET

KABLOSUZ KANALLAR ÜZERİNDE DÜŞÜK KARMAŞIKLIKLI KAYNAKSIZ RASTGELE ERİŞİM ŞEMALARININ TASARIMI

Mert Özateş
Elektrik ve Elektronik Mühendisliği, Doktora
Tez Danışmanı: Tolga Mete Duman
Kasım 2023

Altıncı Nesil ve Ötesi iletişim sistemlerinin çok sayıda makine tipi cihazın iletişimini sağlaması beklenmektedir. Bu cihazlardan bazılarının oluşturduğu trafik, geleneksel iletişim senaryolarından önemli ölçüde farklı olacaktır. Örneğin, çok sayıda ucuz sensörün bir baz istasyonu (BS) ile iletişim kurduğu uygulamalarda, cihazlar yalnızca ara sıra aktif olacak ve aralarında veya BS ile herhangi bir koordinasyon olmayacaktır. Masif rastgele erişim çözümleri gerektiren sistemler için yakın zamanda kaynaksız rastgele erişim (URA) adı verilen yeni bir paradigma önerilmiştir. URA'da tüm kullanıcılar aynı kod kitabını kullanır ve kullanıcı kimliği yoktur. Hedef yalnızca aktif kullanıcı grubundan gönderilen mesajların listesiyle ilgilenir. Güncel literatürde geliştirilen birçok ilginç URA şeması olmasına rağmen, özellikle düşük karmaşıklıklı ve enerji açısından verimli çözümlerin tasarlanması konusunda birçok önemli zorluk devam etmektedir.

URA'daki mevcut zorlukların üstesinden gelme motivasyonuyla çeşitli senaryolar için pratik çözümler geliştiriyoruz. İlk olarak, sönümleme etkilerini azaltmak için dikey frekans bölmeli çoklama yoluyla frekans seçici kanallar üzerinden URA'yı öneriyoruz ve üzerinde çalışıyoruz. Kod çözücü, girişimi gürültü olarak ele alma ve ardışık girişim iptali (SIC) ile birleştirilmiş bir ortak aktivite algılama ve kanal kestirimi algoritması kullanır. Sonuçlarımız, önerilen şemanın, hibe bazlı frekans bölmeli çoklu erişim ile rekabetçi bir performans sunduğunu ve tahmini kanal durumu bilgisinden kaynaklanan performans kaybının sınırlı olduğunu göstermektedir. Daha sonra alıcının çok sayıda antenle donatıldığı senaryoyu inceliyoruz ve iletim çerçevesini, her aktif kullanıcının dikey olmayan bir pilot diziyi ve ardından kutupsal kodlanmış kod sözcüğünü kullandığı yarıklara bölerek basit ama enerji açısından verimli bir çözüm geliştiriyoruz. Alıcıda, ilk önce genelleştirilmiş bir dikey eşleştirme takip algoritması ile iletilen pilot dizileri

tespit ediyoruz ve kanal vektörlerini tahmin etmek için doğrusal bir minimum ortalama kare hatası (LMMSE) çözümü kullanıyoruz. Daha sonra maksimum oran birleştirme ve tek kullanıcılı kod çözmeye ve ardından SIC'ye dayalı yinelemeli kod çözme gerçekleştiriyoruz. Sayısal örnekler ve analiz sonuçları, önerilen şemanın literatürdeki mevcut yaklaşımlardan daha iyi performans gösterdiğini veya daha düşük karmaşıklıkla rekabetçi bir performansa sahip olduğunu göstermektedir. Daha sonra çözümümüzü, donanım bozuklukları (HWI) istatistiklerini kullanan kanal kestirimi için donanım bozukluğuna duyarlı bir LMMSE çözümü geliştirerek BS ve kullanıcı ekipmanı tarafında kalan donanım bozukluklarının olduğu senaryolara uyarlıyoruz ve yeni önerilen çözümün enerji verimliliğini ve desteklenen aktif kullanıcı sayısını artırdığını gözlemliyoruz. Son olarak, her aktif kullanıcının iletim çerçevesinin küçük bir bölümünü kullandığı açma-kapama bölmeli çoklu erişimi URA bağlamında inceliyoruz ve yeni yaklaşımın performans veya karmaşıklık açısından mevcut yaklaşımlardan üstün olduğunu gösteriyoruz.

Anahtar sözcükler: Kaynaksız rastgele erişim, kutupsal kodlar, masif rastgele erişim, frekans seçici kanallar, çoklu-girişli çoklu-çıkışlı sistemler, genelleştirilmiş dikey eşleme takibi, donanım bozuklukları, açma-kapama bölmeli çoklu erişim.

Acknowledgement

First and foremost, I would like to thank my supervisor Prof. Tolga Mete Duman. His patience, continuous support, and dedication into research have been a significant value for this thesis and contribute me a lot as a person.

I am grateful to Prof. Sinan Gezici and Prof. Ali Özgür Yılmaz for being a member of my thesis committee (TİK) and providing valuable comments throughout our committee meetings. I also would like to thank Prof. Erdal Arıkan and Prof. Ayşe Melda Yüksel Turgut for accepting to be a member of my thesis jury.

I would like to thank Dr. Mohammad Kazemi for his assistance and our fruitful discussions on my research. I feel lucky since I have collaborated him. I also thank the members of the Communications Theory and Applications Research Lab and my other friends in Bilkent for our enjoyable time.

This work was supported by Vodafone within the framework of 5G and Beyond Joint Graduate Support Programme coordinated by Information and Communication Technologies Authority and Scientific and Technological Research Council of Turkey (TUBITAK) under the grant 119E589. I am grateful to Vodafone and TUBITAK for this support.

Last, but not least, I would like to express my gratitude to my family for their love, support and motivation.

Contents

List of Figures

List of Tables

Abbreviations

AD	activity detection
AMP	approximate message passing
AWGN	additive white Gaussian noise
BCH	Bose–Chaudhuri–Hocquenghem
BP	belief propagation
BPSK	binary phase shift keying
BS	base station
CDMA	code division multiple access
CP	cyclic prefix
CRC	cyclic redundancy check
CS	compressed sensing
CSA	coded slotted ALOHA
CSI	channel state information
DFT	discrete Fourier transform
FBL	finite blocklength
FFT	fast Fourier transform

FDMA	frequency division multiple access
GMAC	Gaussian multiple access channel
gOMP	generalized orthogonal matching pursuit
HWI	hardware impairment
IoT	Internet-of-things
IRSA	irregular repetition slotted ALOHA
LDPC	low-density parity check
LLR	log-likelihood ratio
LMMSE	linear minimum mean square error
MAC	multiple access channel
MIMO	multiple-input multiple-output
ML	maximum likelihood
MMSE	minimum mean square error
mMTC	massive machine type communications
MRC	maximal ratio combining
MSE	mean square error
NNLS	non-negative least squares

NOMA	non-orthogonal multiple access
ODMA	on-off division multiple access
OFDM	orthogonal frequency division multiplexing
OMA	orthogonal multiple access
OMP	orthogonal matching pursuit
QPSK	quadrature phase shift keying
PUPE	per-user probability of error
RA	repeat accumulate
SCMA	sparse code multiple access
SCLD	successive cancellation list decoder
SDMA	space division multiple access
SIC	successive interference cancellation
SINR	signal-to-interference-noise ratio
SISO	soft-input soft-output
SKP	sparse kronecker product
SNR	signal-to-noise ratio
SPARC	sparse regression code

TDMA	time division multiple access
TIN	treating interference as noise
UE	user equipment
URA	unsourced random access

Chapter 1

Introduction

1.1 Overview

Next generation wireless communication systems require massive connectivity, low latency, and high spectral efficiency. As a result, massive machine type communications (mMTC) will become a key aspect of beyond 5G (B5G) wireless networks. In some applications of mMTC like the Internet-of-Things (IoT), there are a huge number of devices (e.g., millions of devices per km^2) with short payloads and limited computational power, and their transmissions have a sporadic and uncoordinated nature. This scenario is generally referred as massive access [1].

Multiple access is a building block for massive access, which allows multiple users share a communication medium for their transmissions simultaneously. A conventional way to serve multiple users in the same system is to assign them orthogonal system resources as in time division multiple access (TDMA) or frequency division multiple access (FDMA). However, these schemes require scheduling of the users by the base station (BS) and are infeasible when the number of the users becomes large due to the excessive delay and high signaling overhead.

A well-known solution for the communication of a large number of users is grant-free random access where the users transmit their data independently without any scheduling by the BS. In this scenario, the receiver does not have the information of the active user set; however, the users can pick unique code sequences prior to the transmission which can be utilized by the receiver to differentiate them via activity detection. However, this approach becomes insufficient in the case of massive access due to the lack of resources (code sequences), hence other solutions are needed.

In order to address the massive access problem, Polyanskiy introduced the unsourced random access (URA) paradigm in [2]. In URA, it is assumed that a potentially unbounded number of devices with short payloads sporadically communicate with the BS without any coordination. The devices share the same codebook, hence there is no user identity and the decoding is only up to a permutation of the transmitted messages. Since there is no user identity, the system can operate irrespective of the total number of devices. In addition, per-user probability of error (PUPE) is adopted as the main performance criterion rather than the stringent global error probability. These aspects lead to a substantial departure from the traditional multiple access scenario.

Polyanskiy also developed a random coding achievability bound on the energy efficiency of URA over additive white Gaussian noise (AWGN) channels. Since then, there have been substantial efforts to develop low-complexity coding schemes to approach the theoretical limits. The main approaches for this purpose are using a slotted ALOHA protocol that is capable of recovering up to T colliding users (T-fold ALOHA), coded compressed sensing, and random spreading that can be thought as an application of code division multiple access (CDMA) to the URA scenario.

In the Gaussian multiple access channel (MAC) model, the received powers of all the users are equal. However, due to the differences in their distances to the BS or different environment conditions, this model may be inadequate. With this motivation, fading channel models where the transmitted signals of the users are attenuated with a random fading coefficient are also considered in many recent

works in the URA context. The channel state information of the users is unknown at the receiver, which makes the estimation of the fading coefficients a part of the problem. However, approaches for designing low-complexity coding schemes in the Gaussian MAC scenario can be utilized by also incorporating suitable channel estimation algorithms.

Massive multiple-input multiple-output (MIMO) is a key technology for current wireless communication systems due to its potential to provide high spectral efficiencies and spatial multiplexing gains. Hence, it becomes a prominent candidate to increase the supported active user load and the energy efficiency of the URA systems. There are already several solutions addressing URA with a massive MIMO receiver adapting the similar design concepts with the case of single antenna receivers.

Motivated by the recent developments, in this thesis, we develop practical transmission schemes for different scenarios in the context of URA. Specifically, we consider frequency-selective channels for the first time in this context, which is a more realistic scenario caused by multipath fading, compared to flat fading. We propose to employ orthogonal frequency division multiplexing (OFDM) to overcome the detrimental effects of the multipath, i.e., to eliminate the inter-symbol interference. We then investigate URA with a massive MIMO receiver and develop an energy-efficient scheme with low complexity employing slotted transmissions and generalized orthogonal matching pursuit (gOMP) based activity detection, linear minimum mean square error (LMMSE) channel estimation, and polar coding. We also extend the proposed scheme for the case with transceiver hardware impairments, which is an inevitable practical scenario for URA. Finally, we study URA based on on-off division multiple access (ODMA) with polar coding employing single-antenna receivers.

1.2 Contributions

In the first part of the thesis, we propose and study URA over frequency-selective channels, which can be observed in many practical environments due to multipath propagation. To this end, we assume that the active users transmit their data over a frequency-selective channel and employ OFDM to overcome the multipath fading effects. We consider slotted transmissions in the frequency domain, and the pilot signal of each user is uniformly distributed in its OFDM word to be transmitted along with the data part encoded by a polar code. We propose a joint activity detection (AD) and channel estimation algorithm to estimate the channel taps of the users, followed by treating interference as noise (TIN) in conjunction with successive interference cancellation (SIC) to recover the message bits. We consider grant-based FDMA for comparison and observe that the proposed scheme offers a competitive performance. The results along this line of investigations have been published in [3].

In the second part of the thesis, we consider URA with a massive MIMO receiver and propose an energy-efficient scheme with low complexity based on slotted transmissions. We assume that the users employ non-orthogonal pilot sequences followed by polar encoded codewords to transmit their data. At the receiver side, we propose a decoding algorithm combining the ideas of gOMP for AD, LMMSE channel estimation, maximal ratio combining (MRC) for symbol estimation, single-user decoding and re-estimation of the channel vectors for SIC. We also analyze the performance of the proposed scheme using an analytical signal-to-interference-and-noise ratio (SINR) characterization and normal approximations (based on some results from finite length information theory). Through a set of numerical results, we demonstrate that the proposed solution either outperforms the schemes in the literature or has a competitive performance with lower complexity, and it has a small gap with its approximate performance limits. Furthermore, it is suitable for fast-fading scenarios due to its excellent performance in the short blocklength regime. The results along this line of investigations have been published in [4, 5].

We also extend our work on URA over wireless channels to the case with residual hardware impairments (HWIs) at both the BS and the UE side with the motivation that they are widely observed in practical massive MIMO systems and the devices in URA scenario are likely to have a cheap hardware due to their massive numbers. We propose a HWI-aware receiver using the HWI statistics and observe that the HWI-aware receiver can increase the energy efficiency and the number of the active users that can be supported. Our results in this context have been published in [6].

In the last part of the thesis, we investigate on-off division multiple access (ODMA) for URA combined with polar coding. Namely, we assume that the active users distribute their polar codewords to the transmission frame based on a transmission pattern determined by a part of their message bits. We propose a low complex pattern detection method based on the received signal energy followed by single-user decoding and SIC. We observe that the newly proposed scheme offers a promising performance in both Gaussian and fading MAC scenarios with low complexity.

1.3 Thesis Outline

The rest of the thesis is organized as follows. We review the existing literature on multiple access and unsourced random access in Chapter 2. We present our proposed coding scheme based on OFDM and TIN-SIC for URA over frequency-selective channels in Chapter 3. We then focus on URA with a massive MIMO receiver and propose a slotted transmission scheme in Chapter 4, which we extend for the scenario with HWIs in Chapter 5. We introduce a transmission scheme based on ODMA and polar codes for URA in Chapter 6, and conclude the thesis in Chapter 7 along with some future research directions.

Chapter 2

Review of Multiple Access and Massive Random Access

In this chapter, we start with an overview of multiple access and then present an extensive literature review of unsourced random access. For the part on URA, we first provide the system model and fundamental performance limits along with a summary of the existing coding schemes considering Gaussian MAC and fading MAC with a single antenna receiver, and then we extend our coverage to URA over fading channels with a receiver equipped with a massive number of antennas.

2.1 Multiple Access Techniques

Multiple access refers to the scenario in which more than one user simultaneously share the system resources. One way to serve multiple users is to assign dedicated resources (e.g., time, frequency, or code) to them by a central coordination unit through a prior handshaking procedure. The conventional techniques in this framework are TDMA, FDMA, code division multiple access (CDMA), and space division multiple access (SDMA). In TDMA and FDMA, non-overlapping time slots or frequency sub-channels are assigned to the different users, respectively,

and the data of each user is detected in its allocated time frame or frequency band. On the other hand, in CDMA, all the users can utilize the entire time-frequency resources simultaneously through different (nearly) orthogonal code sequences coupled with a low-complexity decorrelation based receiver. SDMA relies on the principle of creating (nearly) orthogonal spatial channels for different users to minimize the inter-user interference by employing beamforming and spatial multiplexing.

All of the approaches described in the previous paragraph are in the class of orthogonal multiple access (OMA), where the number of users is strictly limited due to the orthogonal resource assignment. In order to increase the number of supported users, non-orthogonal multiple access (NOMA) techniques are investigated in the literature, where multiple users can utilize the same resources to transmit their data, allowing the system to support more users. However, this brings an inter-user interference to the system, and more sophisticated interference cancellation receivers are needed with an increased complexity compared to OMA.

The NOMA schemes are divided into two main categories, which are called power-domain NOMA and code-domain NOMA. In power-domain NOMA [7], the users employ different power levels to transmit their data using the same resources, and the receiver exploits the received power difference for detection through SIC. On the other hand, code-domain NOMA is similar to classical CDMA in the sense that the users employ user-specific code sequences to transmit their data, however, the code sequences are sparse sequences or non-orthogonal sequences with low cross-correlation [8]. Here, the receiver can employ message passing algorithms utilizing the code structure to differentiate the users. Some examples of code-domain NOMA schemes in the literature are low-density spreading code division multiple access [9], low-density spreading aided orthogonal frequency division multiplexing [10], sparse code multiple access (SCMA) [11], and multi-user shared access [12].

All of the described schemes up to now are grant-based. That is, they require scheduling grants and coordination by the BS prior to transmission, and it is

assumed that the number of the users, their activity patterns etc. are available at the BS. However, when the number of users become large, they become infeasible due to the excessive delays and signaling overhead. For example, typical number of users per signal dimension in time-frequency domain (per degree of freedom) is between 1.5-3 according to 3GPP studies, as the system performance significantly degrades beyond that threshold [1].

Grant-free multiple access is a well-known candidate to enable communication of a large number of users where the users transmit their data without any prior grant by the BS using the common system resources, which reduces the latency and signaling overhead. Nevertheless, the users can have specific signatures/preambles or they can utilize different codebooks pre-configured by the BS. In this line, grant-free NOMA is the prominent solution for future communication systems due to its low latency, improved spectral efficiency, and its potential to support massive connectivity since the transmission resources are non-orthogonal. Note that the main difference between grant-based and grant-free NOMA is that, the BS does not have the information of the set of active users in the latter, which brings many challenges such as blind activity detection and data recovery, channel estimation, and synchronization with minimal overhead [13].

Power-domain NOMA and code-domain NOMA represent the two main categories of NOMA. Among them, the power-domain NOMA is hard to extend to the grant-free scenario as the successful recovery at the receiver heavily depends on the power difference of the users, which may not be maintained due to the grant-free nature of the transmissions. On the other hand, signature based schemes in code-domain NOMA can be adapted to the grant-free case with a proper activity detection prior to the data detection or joint activity detection and data recovery. For instance, a grant-free version of SCMA is proposed in [14]. Furthermore, the inherent sparsity of the user activity in grant-free multiple access is utilized in [15, 16, 17, 18] through compressive sensing algorithms for multi-user detection.

Another candidate solution to address the drawbacks of the coordinated multiple access is uncoordinated multiple access, where the users exploit the system

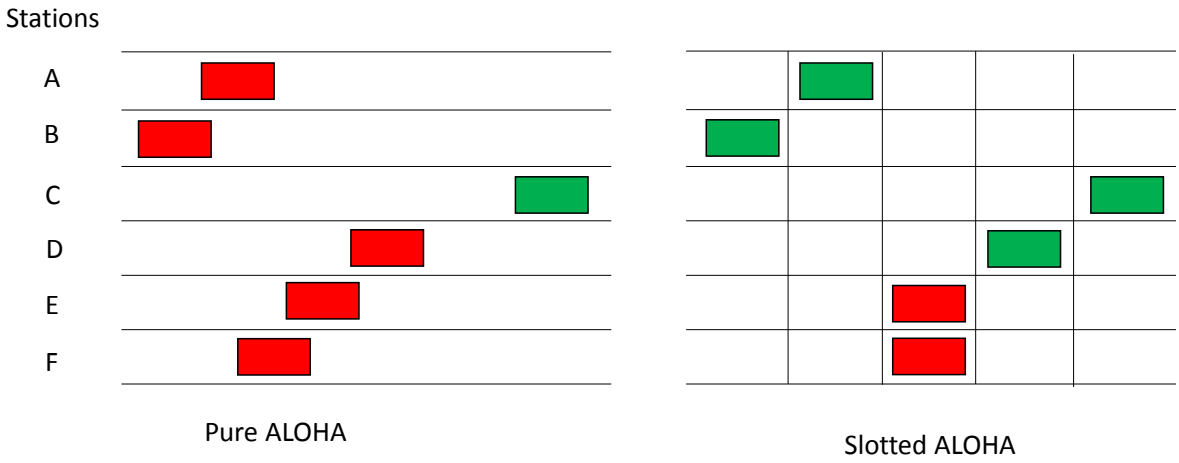

Figure 2.1: An illustration of pure and slotted ALOHA. Red boxes show the collided packets and green boxes are the successfully transmitted ones.

resources independently to transmit their data using the same transmission protocol without any coordination by the BS. This scheme is particularly useful when there are stringent latency requirements or the user activity is random/sporadic, which renders assigning dedicated system resources ineffective.

The first approach for uncoordinated multiple access is the ALOHA protocol [19] proposed in the 1970s, where the users transmit the data packets whenever they have data to send. The data packets are assumed to be lost in the case of a collision and they should be re-transmitted. ALOHA is simple; however, it is inefficient as the system throughput is $T = Le^{-2L}$ with a maximum of $1/2e \approx 18\%$ where T is the system throughput and L is the system load. An improvement to the pure version of ALOHA is called slotted ALOHA. In slotted ALOHA, it is assumed that the time is divided into slots and the users can start the transmission only at the beginning of the time slots. Thus, there are no partial collisions, and a packet is assumed as successfully received if there are no other transmissions in the same slot. In this case, the throughput can be calculated as $T = Le^{-L}$, with a maximum of $1/e \approx 37\%$. The transmission structures in pure ALOHA and slotted ALOHA are illustrated in Figure 2.1, and their throughput with respect to the system load is depicted in Figure 2.2.

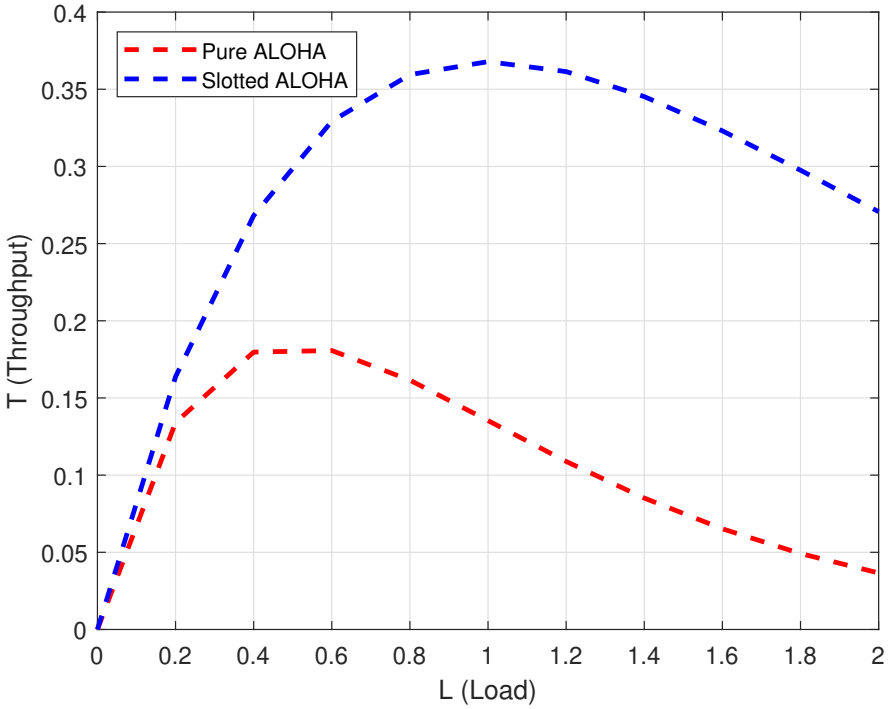

Figure 2.2: Throughput comparison of pure and slotted ALOHA.

Although slotted ALOHA doubles the throughput of pure ALOHA, it is still less than 40 %, which is limited. With the motivation to increase the throughput, several variants of ALOHA are proposed in the recent literature. For instance, in Contention Resolution Diversity Slotted ALOHA (CRDSA) [20], two copies of the packets are transmitted in two different slots, and in the case that a packet is successfully recovered in one slot, the effect of its replica in the other slot is subtracted. In this way, the maximum throughput can be increased to 52%. This scheme is improved by irregular repetition slotted ALOHA (IRSA) [21], where the packets are repeated different number of times rather than twice. With the optimization of the repetition distributions by density evolution techniques, IRSA achieves $T \simeq 0.8$ in practical implementations. An extension of IRSA exploiting forward error correction is called Coded Slotted ALOHA (CSA) [22], which improves the efficiency even further, approaching a throughput of almost 1 when sufficiently low coding rates are employed. IRSA and CSA are further studied in [23] - [26].

In some mMTC applications, the total number of devices can be in the order of millions, and they are only active very sporadically. This makes achieving any level of coordination infeasible. As a result, none of the multiple access

techniques summarized in this section is suitable for enabling the communication of these devices. To address this problem, Polyanskiy proposed URA paradigm in [2] where the devices share the same codebook, hence there is no user identity, and the system operation becomes independent of the total number of devices. The decoding is done only up to a permutation of the transmitted messages, and the main performance criterion is the per-user probability of error. The existing literature on URA will be presented in the following sections.

2.2 Unsourced Random Access over Gaussian MAC

2.2.1 Basic Model and Random Coding Bound

In his pioneering work [2], Polyanskiy introduces the URA paradigm and focuses on the following Gaussian MAC model:

$$\mathbf{y} = \mathbf{x}_1 + \cdots + \mathbf{x}_{K_a} + \mathbf{z}, \tag{2.1}$$

where K_a is the number of active users, $\mathbf{x}_i$ is the transmitted signal of the i-th user, $\mathbf{y}$ is the received signal and $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ is the AWGN. Assuming $\|\mathbf{x}_i\|^2 \leq nP$ is satisfied, the energy-per-bit becomes

$$\frac{E_b}{N_0} = \frac{nP}{2\log_2 M}, \tag{2.2}$$

where M is the cardinality of the set of the possible messages $[M]$ and n is the length of the transmission frame. In [2], an (M, n, ϵ) random access code is defined as follows

Definition 1. An (M, n, ϵ) random access code for the K_a-user channel $P_{Y|X^{K_a}}$ is a pair of (possibly randomized) maps - the encoder $f : [M] \to \mathcal{X}^n$ and the decoder $g : \mathcal{Y}^n \to \binom{[M]}{K_a}$ - satisfying

$$\frac{1}{K_a}\sum_{j=1}^{K_a}\mathbb{P}[E_j] \leq \epsilon, \tag{2.3}$$

where $E_j \triangleq \{W_j \notin \mathcal{L}\} \cup \{W_j = W_i \text{ for some } i \neq j\}$ is the error event of the j-th user, $W_1, \ldots, W_{K_a}$ are independent and uniform on $[M]$, and $\mathcal{L}$ is the list of the estimated user messages that is produced by the decoder.

A random coding achievability bound is also obtained in [2]. This bound can be used as a benchmark for the energy efficiency of URA over real-valued AWGN channel as it assumes that there are no complexity constraints. It is given by the following result.

Theorem 1 (**Random coding achievability [2]**). *Fix $P' < P$. There exists an (M, n, ϵ) random-access code for K_a-user GMAC satisfying power-constraint P and*

$$\epsilon \leq \sum_{i=1}^{K_a}\frac{t}{K_a}min(p_t, q_t) + p_0, \tag{2.4}$$

■

where

$$p_0 = \frac{\binom{Ka}{2}}{M} + K_a\mathbb{P}\left[\frac{1}{n}\sum_{j=1}^{n}Z_j^2 > \frac{P}{P'}\right], \quad p_t = e^{-nE(t)}, \tag{2.5}$$

where $E(t) = \max\limits_{0\leq\rho,\rho_1\leq 1} -\rho\rho_1 tR_1 - \rho_1 R_2 + E_0(\rho,\rho_1)$, $E_0 = \rho_1 a + \frac{1}{2}\log(1-2b\rho_1)$, $a = \frac{\rho}{2}\log(1+2P't\lambda) + \frac{1}{2}\log(1+2P't\mu)$, $b = \rho\lambda - \frac{\mu}{1+2P't\mu}$, $\mu = \frac{\rho\lambda}{1+2P't\lambda}$, $\lambda = \frac{P't-1+\sqrt{D}}{4(1+\rho_1\rho)P't}$, $D = (P't-1)^2 + 4P't\frac{1+\rho\rho_1}{1+\rho}$, $R_1 = \frac{1}{n}\log M - \frac{1}{nt}\log(t!)$, $R_2 = \frac{1}{n}\log\binom{K_a}{t}$, and q_t is defined as

$$q_t = \inf_{\gamma}\mathbb{P}[I_t \leq \gamma] + exp(n(tR_1 + R_2) - \gamma). \tag{2.6}$$

The random variable I_t is defined as $I_t = \min_{S_0} i_t(c(S_0); Y|c(S_0^c)$ where $c(S_0) \triangleq \sum_{j\in S_0} c_j$ for $c_1, \ldots, c_M \sim \mathcal{N}(0, P')$ and the minimum is over all t- subsets of $[K_a]$. The information density i_t is defined as

$$i_t(a; y|b) = nC_t + \frac{\log e}{2}\Big(\frac{\|y-b\|_2^2}{1+P't} - \|y-a-b\|_2^2\Big), \tag{2.7}$$

where $C_t = \frac{1}{2}\log(1+P't)$. ■

The idea behind the proof of this theorem is as follows: Let $S = \{W_1, \ldots, W_{K_a}\}$ be a random K_a-subset of $[M]$, the proof of the theorem is based on bounding the error event $F_t \triangleq \{|S \setminus \hat{S}| = t\}$ with two different methods, one considers a Gallager-type bound and the other uses the information density in (2.7). The details can be found in [2].

The random coding bound is an information-theoretic performance limit, and it is derived without any complexity constraints as it assumes that all active users are jointly decoded, which has a prohibitive complexity. Practical coding schemes for URA are needed to approach the results predicted by the bound.

Since the communication in URA is uncoordinated and the number of total users is huge, grant-based schemes like TDMA or FDMA are infeasible as they require scheduling of the users. Hence, the random access schemes such as slotted ALOHA and TIN become the first candidate solutions.

In slotted ALOHA, the transmission frame is divided into sub-frames and each user transmits randomly in an independently selected sub-frame. It is assumed that the slot decoding works if there is no collision and the single-user decoding is successful. Moreover, as shown in [2], TIN coding satisfies the following approximation

$$\log M \approx nC_{TIN}(P) - \sqrt{\frac{nP\log^2 e}{1+K_aP}}Q^{-1}(\epsilon), \tag{2.8}$$

where $C_{TIN}(P) = \frac{1}{2}\log\Big(1+\frac{P}{1+(K_a-1)P}\Big)$.

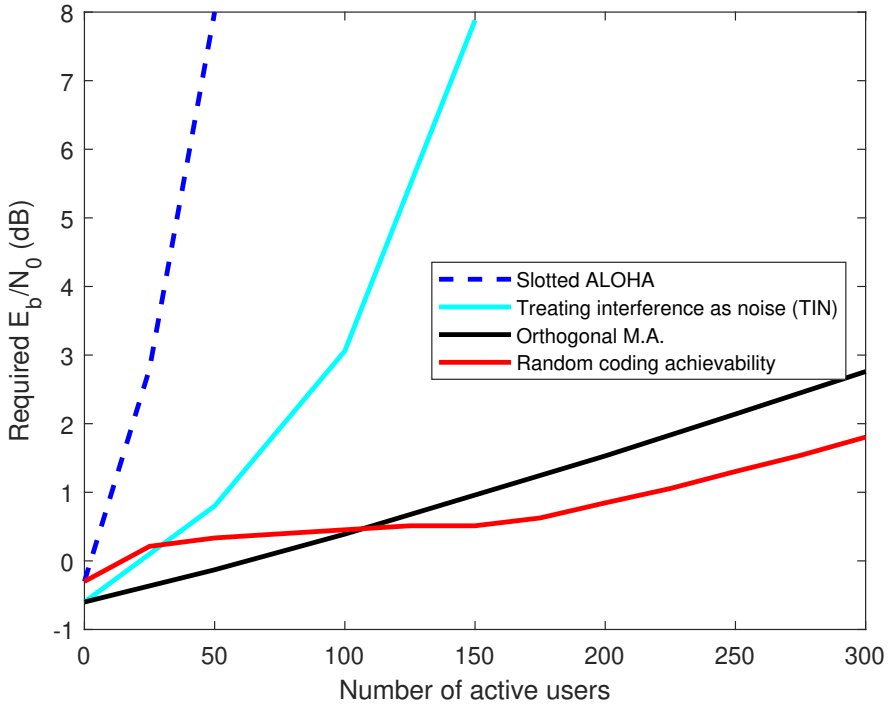

Figure 2.3: Required E_b/N_0 for versus the number of active users.

In [2], the performances of these approaches are evaluated assuming that there are K_a active users transmitting $k = 100$ bits of information through a transmission frame of length $n = 30000$, with a target PUPE of $\epsilon = 0.1$. The required E_b/N_0 versus number of active users for these schemes are plotted in Figure 2.3 along with the random coding bound derived in [2]. The results in Figure 2.3 show that the conventional random access techniques (i.e., slotted ALOHA and TIN) are energy inefficient for the URA scenario as there is a huge performance gap between these and the random coding bound, and the development of low-complexity coding schemes is needed. The performance of the grant-based TDMA scheme where the frame is divided into sub-frames of length $\frac{n}{K_a}$ and each active user transmits in its dedicated sub-frame referred as *Orthogonal M.A.* is also plotted in the figure (though it is infeasible for URA). Note that in all of the curves, it is assumed that the single-user codes satisfying the normal approximation in [27] are employed.

2.2.2 Practical Coding Schemes for URA

As indicated in the previous subsection, it is shown in [2] that conventional random access techniques are energy inefficient for URA. As a result, there have been substantial efforts to develop low-complexity and energy-efficient schemes in the subsequent literature. One of the design strategies for this purpose is slotting the transmission frame to reduce the multiuser interference, where each active user picks one or a few slots to transmit its data depending on the exact setting. In this direction, Ordentlich and Polyanskiy propose the first practical scheme for URA in [29], where a concatenated coding scheme is considered combined with the T-fold ALOHA protocol. T-fold ALOHA is an approach similar to slotted ALOHA in the sense that the users choose a random sub-block (slot) to transmit their packets, but in the slotted ALOHA, packets are lost in the case of a collision. However, in T-fold ALOHA, T or fewer users can be simultaneously decoded in each slot. In [29], the users choose one slot randomly to transmit their message in that slot. The concatenated coding scheme consists of an inner binary linear code which is used to decode the modulo-2 sum of the codewords transmitted within a slot and an outer code to recover the individual messages. Many off-the-shelf codes can be used for the inner code. The authors construct the outer code from the columns of a T-error correcting BCH code. This scheme suffers in terms of energy efficiency as the performance gap with the random coding bound is about 20 dB when the number of active users is 300.

The authors of [30] also consider a slotted structure where the users can repeat their codewords across sub-blocks, which is referred as T-fold IRSA. They propose to split the message into two parts and employ a coding scheme based on a combination of compressed sensing (CS) and low-density parity check (LDPC) coding at the slot level. The replicas of the recovered codewords are peeled from the other slots through successive interference cancellation (SIC). Note that the slotted transmission structure in [30] is depicted in Figure 2.4. T-fold IRSA is also utilized in [31] by replacing the LDPC codes with the polar codes, which provides a significant performance gain over [30]. In [32], the authors employ random signatures to convert the equal-gain channel into a more favorable independent

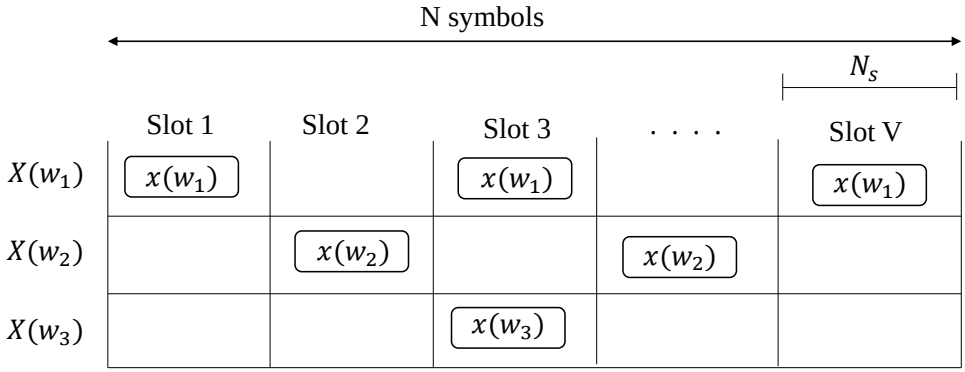

Figure 2.4: The slotted transmission structure in [30] based on T-fold IRSA.

and identically distributed (i.i.d.) fading scenario along with the trellis-based codes in a T-fold ALOHA structure. Their proposed approach provides a low complexity solution with moderate performance.

In another line of investigation, the problem is addressed through an approach that is called coded compressed sensing [33] - [37]. Namely, the message is divided into sub-blocks and each block is individually encoded by a combination of an inner CS code and an outer code to add redundancy. The encoded sub-blocks are then transmitted in different slots. At the receiver side, each sub-block is recovered through compressed sensing algorithms, and the recovered segments are stitched together by an outer tree code. In this line, a recent scheme employs sparse regression codes (SPARCs) as inner codes [37]. Then, the sub-blocks are CRC-encoded and they are connected via block Markov superposition transmission. At the receiver, a hybrid decoder combining successive cancellation (SC) and approximate message passing (AMP) is employed for inner decoding, and tree decoding is utilized to piece the inner-decoded sub-blocks together. This scheme improves the performance of the proposed scheme in [34] with lower complexity and provides the best performance among the works based on idea of coded compressed sensing for URA over Gaussian MAC. The general encoding-decoding structure in coded compressed sensing is illustrated in Figure 2.5.

In another design strategy, there are no slots and the active users utilize the whole transmission frame. In some of the works following this strategy, the idea

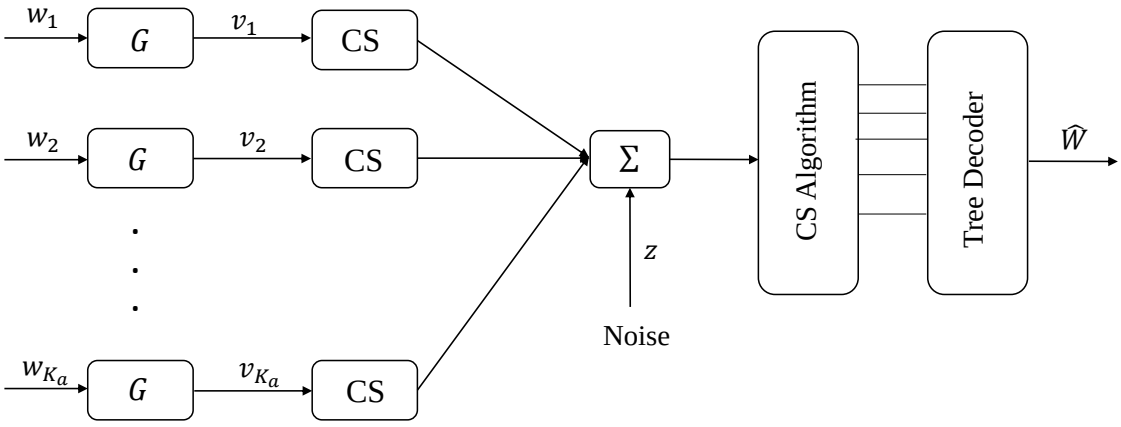

Figure 2.5: Encoding-decoding structure in coded compressed sensing.

is to spread the channel coded bits through the transmission frame using random signatures [38] - [40]. The message is divided into two parts: the first part determines the signature sequence while the second part is encoded by a polar code in [38]. At the receiver side, first, the selected signature sequences are detected by a correlation-based energy detector. Then, an iterative decoding algorithm consisting of MMSE filtering followed by single-user polar decoding is employed to recover the channel coded bits. Also used is SIC at the end of each iteration as illustrated in Figure 2.6. The authors of [39] propose to assign different power levels to the active users in the proposed scheme in [38]. This approach creates an artificial fading-like scenario, where the users with higher received power can be decoded first, and the ones with lower received power are decoded after SIC. This approach provides a performance improvement over [38] for $K_a \geq 150$. The authors of [40] replaces the polar codes in [38] with LDPC codes to exploit the soft information produced by the belief propagation decoder to perform soft-input-soft-output (SISO) MMSE filtering. This approach improves the decoding performance since the information about the partially decoded codewords can be utilized, and it provides the best performance among the proposed schemes based on random spreading.

In [38] - [40], each active user occupies the entire frame for data transmission.

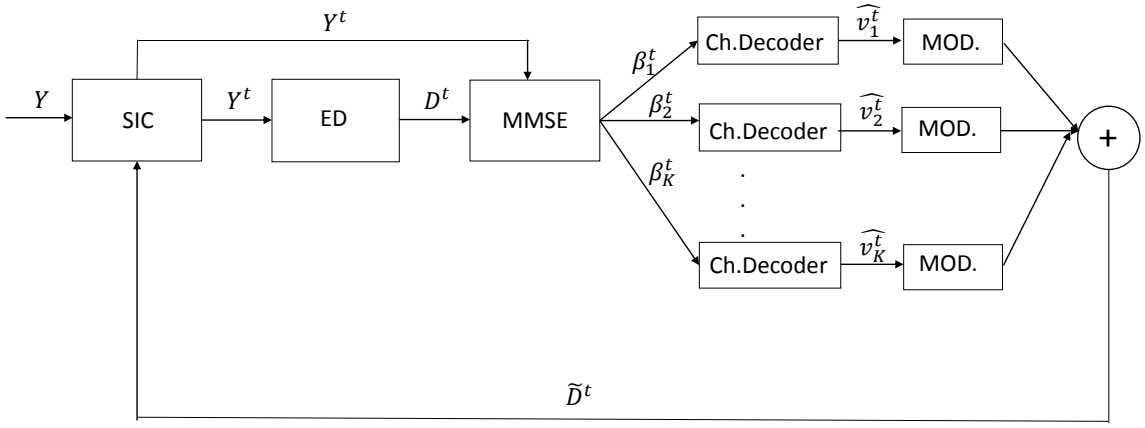

Figure 2.6: A typical decoding structure in random spreading.

As an alternative approach, in some works, the active users distribute their codeword bits onto the transmission frame where each user occupies a small fraction of it. For instance, in [41], LDPC encoded bits are sparsified along the channel frame through zero-padding and interleaving. At the receiver side, a single joint Tanner graph is utilized for message recovery. Furthermore, the authors of [42] propose to encode the data as the Kronecker product of the two component codewords, and employ an iterative decoder based on bilinear generalized approximate message passing to decompose the Kronecker product and a soft-in-soft-out decoder for the individual components.

In another recent work [44], the authors employ on-off division multiple access (ODMA) [43] in the context of URA, where the active users exploit a small part of the transmission frame to transmit their data encoded by a repeat-accumulate (RA) code according to a transmission pattern, while the remaining time instances are idle. At the receiver side, the BS first recovers the transmission patterns and then employs multiuser detection and decoding over a sparse factor graph. Among the proposed schemes for URA over a Gaussian MAC, [40] exhibits the best performance for $K_a \leq 225$ with the cost of high computational complexity, while the low complexity approach of [44] is superior to the other schemes for $K_a > 225$.

The conventional assumption in URA is that all the users have the same number of message bits, which holds in most of the existing works in the literature. However, this may be impractical in the situations where the devices have different payload requirements, power budgets etc. To address this problem, the scenario that two different user groups have different number of message bits is studied in [45, 46] in the coded compressed sensing framework, called multi-class URA. In [45], SPARCs are utilized as inner codes and LDPC codes are employed as outer codes. On the other hand, in [46], the compressed sensing techniques and tree codes are exploited as the inner and outer codes, respectively.

2.3 URA over Fading MAC

2.3.1 Channel Model

Gaussian MAC is an idealized model as the users' transmissions have equal received powers. A more realistic model is quasi-static fading MAC, where the codewords are attenuated by random fading coefficients. For the single-antenna quasi-static fading MAC, the model becomes [49]:

$$\mathbf{y}^n = \sum_{i=1}^{K_a} h_i \mathbf{x}_i + \mathbf{z}, \tag{2.9}$$

where $\mathbf{x}_i$ is the transmitted signal of the i-th user, $h_i \sim \mathcal{CN}(0,1)$ is the channel coefficient of the i-th user which are i.i.d. among the users and $\mathbf{z} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_n)$ is AWGN. Assuming that $\|\mathbf{x}_i\|^2 \leq nP$, the energy-per-bit is defined as

$$\frac{E_b}{N_0} = \frac{nP}{k}, \tag{2.10}$$

where k is the number of message bits. The authors of [49] also present a converse bound based on the results from [47] and meta-converse from [48], stated below.

Theorem 2 (**Converse bound [49]**). *Let*

$$L_n = n\ \log(1+PG) + \sum_{i=1}^{n} \left(1 - \left|\sqrt{PG}Z_i - \sqrt{1+PG}\right|^2\right), \tag{2.11}$$

$$S_n = n\ \log(1+PG) \sum_{i=1}^{n} \left(1 - \frac{\left|\sqrt{PG}Z_i - 1\right|^2}{1+PG}\right), \tag{2.12}$$

where $G = \|h\|^2$ *and* $Z_i \sim \mathcal{CN}(0,1)$. *Then, for every* n *and* $0 < \epsilon < 1$, *any* $(M, n-1, \epsilon)$ *code for the quasi-static* K_a *MAC satisfies*

$$\log(M) \leq \log(K_a) + \log \frac{1}{\mathbb{P}(L_n \geq n\gamma_n)}, \tag{2.13}$$

where γ_n *is the solution of* $\mathbb{P}\left[S_n \leq n\gamma_n\right] = \epsilon$. ■

The problem of designing of low complexity coding schemes approaching the information theoretic limits remains as an important problem. Towards this goal, the T-fold slotted ALOHA protocol mentioned in the previous section is a good candidate for the fading scenario as well since it can provide a good performance-complexity trade-off between the classical slotted ALOHA and joint decoding of all users. Therefore, the authors of [49] derive an achievability bound for T-fold slotted ALOHA and propose a practical coding scheme also employing LDPC coding, that is explained in the next subsection in more detail. The energy efficiencies of the achievability bound of T-fold slotted ALOHA and the practical LDPC coding based solution along with the converse bound are plotted in Figure 2.7 for $T = 4$, which demonstrates that there is a considerable performance gap between the proposed solution in [49] and the converse bound. Therefore, it is argued that it may be possible to develop practical transmission schemes with a significantly higher energy efficiency, some of which are reviewed in the next subsection.

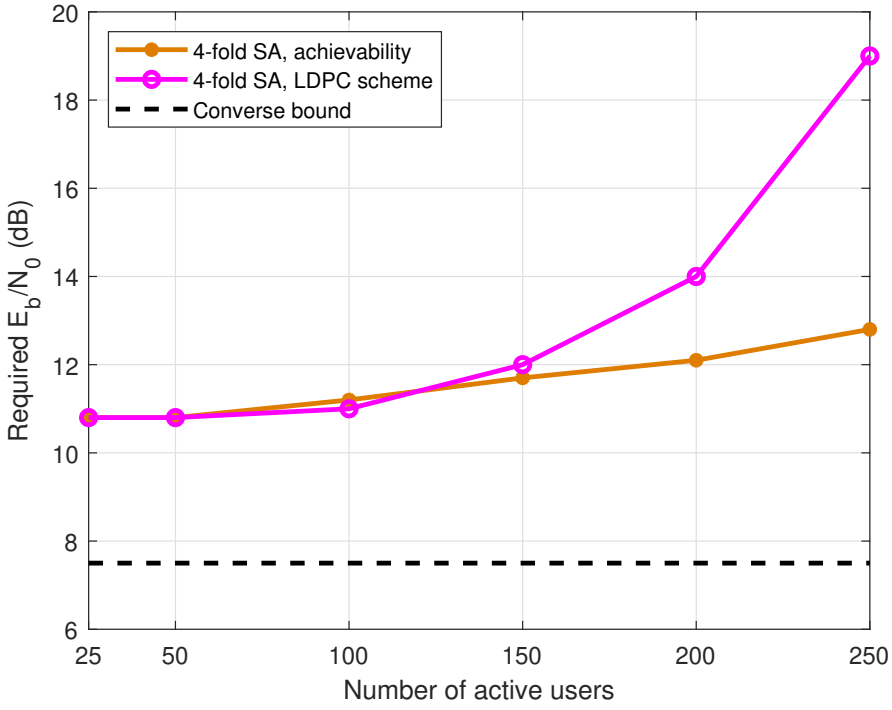

Figure 2.7: Required E_b/N_0 versus number of active users of the information theoretic limits for the fading case and the practical scheme in [49].

2.3.2 Review of Existing Works on URA over Fading MAC

The first practical coding scheme for URA over quasi-static fading MAC is based on a T-fold ALOHA structure in conjunction with LDPC coding [49]. In this work, the authors propose an alternating belief propagation (BP) decoder to estimate the channel coefficients as a part of joint BP decoding by adding the fading nodes to the joint BP decoder consisting of function nodes, and LDPC variable and check nodes.

In another work, a practical polar coding based scheme using T-fold ALOHA protocol combined with TIN-SIC decoding at the receiver side is proposed [51]. Namely, the users are decoded one-by-one given the channel coefficient estimates obtained by clustering techniques while the others are treated as noise and the effects of the successfully decoded users are subtracted at the end of each iteration. This scheme is particularly suitable for fading scenarios as the users with stronger channel coefficients can be decoded first and once their effects are subtracted

from the received signal, the remaining users can be decoded more easily. The authors also provide some achievability bounds for their proposed scheme, and demonstrate that using polar codes in conjunction with TIN-SIC detection can provide a considerable performance improvement over LDPC coding combined with joint BP decoding approach of [49].

As another solution, an approach where the active users utilize the entire transmission frame to transmit their data is considered in [52]. Specifically, the channel frame is divided into pilot and data parts. The active users pick a pilot sequence from a common codebook for transmission in the pilot part, and they repeat their codeword bits along the rest of the transmission frame with permutation and scrambling in the data part. At the receiver side, approximate message passing (AMP) is employed for joint pilot detection and channel estimation, and the data part is recovered by iterative data estimation and interference cancellation using the repeated codeword bits. Numerical examples show that this scheme offers the best performance in terms of energy efficiency in the context of URA over quasi-static fading MAC with a single antenna receiver. Furthermore, the idea of coded compressed sensing is applied for this setup in [53] by employing Reed-Solomon codes and tree codes modified to correct up to t errors.

The works in [49] - [53] consider a completely synchronous scenario, i.e., the transmissions of all active users are aligned in the time domain. With the motivation that this assumption may not be completely feasible in practical implementations, the authors of [54] consider an asynchronous scenario where the active users have time delays of less than $1/4$ of the slot length in a T-fold slotted ALOHA scheme. They propose an OFDM-based transmission scheme operating in the frequency domain to convert the time synchronization problem into a phase shift estimation problem, which can be addressed through a cyclic prefix (CP) extension to the OFDM signal. They employ LDPC codes as a channel code, and TIN-SIC decoding in the frequency domain at the receiver side where the problem of fading coefficient estimation is tackled as in [50]. Via numerical examples, they show that the performance of the proposed scheme in the asynchronous setup is competitive with its synchronous counterpart.

2.4 URA with a Massive MIMO Receiver

Massive MIMO systems play an important role in a wide class of next generation communication systems. In particular, in the context of URA, they have the potential to overcome some bottlenecks. With this motivation, several recent studies on unsourced random access consider a massive number of antennas at the BS. Note that in most of the existing works, a quasi-static Rayleigh fading channel model is assumed.

In [55], the idea of coded compressed sensing is utilized. Namely, the message payload is divided into sub-blocks and each sub-block is encoded individually to one column of a common codeword matrix to be transmitted in a particular slot. At the receiver side, a covariance based approach is employed to recover the inner sub-block codewords via maximum likelihood (ML) decoding or non-negative least squares (NNLS) based decoding. The recovered segments are pieced together by an outer tree code. Numerical examples show that the scheme is particularly suitable for fast fading scenarios as short blocklengths are employed for user transmissions.

The authors of [56] introduce the concept of tensor-based modulation in URA. More specifically, they propose to use rank-1 tensors constructed from Grassmannian sub-constellations as transmitted signals, and a two-step decoder involving multi-user separation exploiting the tensor structure and single-user demapping. The tensor structure allows multi-user separation without the pilot sequences. It exhibits good performance up to around 600 active users, and can be applied to the single-antenna fading and Gaussian MAC scenarios as well. This scheme is extended to the presence of timing offsets in [57] by employimg OFDM. The tensor-based approach is also considered in [58], where the authors propose a block-term decomposition for decomposition of the 3-dimensional tensors, where two low-rank factor matrices and a factor vector is utilized in each block term. This solution provides some performance improvements over the scheme in [56].

In several recent works on URA with a massive MIMO receiver, a separate

pilot transmission is adopted. For instance, in [59], the transmission frame is divided into pilot and data parts, where non-orthogonal pilots are utilized in the pilot part based on part of the message bits and the remaining bits are encoded by a polar code to be transmitted in the data part. At the receiver side, a multiple measurement vector approximate message passing (MMV-AMP) algorithm is used to recover the active pilot sequences followed by LMMSE channel estimation, MRC data detection, and single-user polar decoding. The authors of [60] propose to use multiple stages of orthogonal pilots instead of a single non-orthogonal pilot followed by a polar codeword in a slotted structure. They detect the active pilots using Neyman-Pearson hypothesis testing and utilize an iterative decoder to recover the data part. In [61], the idea of utilizing the diversity of the repeated bits in [52] is extended to the massive MIMO setup. That is, each user picks a pilot from a common codebook which is recovered by MMV-AMP at the receiver, and the data part of the signal is transmitted after repetition, permutation and scrambling, which is then decoded at the receiver employing multiuser detection techniques. Moreover, the idea of spreading the random sequences with channel coded bits through the transmission frame is applied to the MIMO case in [62], where the selected pilot and spreading sequences are identified by an energy detector, and a two-stage iterative decoding is employed for the data part. Namely, the estimates of the channel vectors and the transmitted symbols of the users are obtained by LMMSE filtering followed by MRC and single-user polar decoding in the first stage. Then, the channel vectors are re-estimated using the pilots and temporary coded decisions, and the subsequent operations are repeated with the re-estimated channel vectors with SIC at the end of each iteration. Furthermore, the authors of [63] extend the idea of encoding the data as a sparse Kronecker-product of two component codewords introduced in [42] to the MIMO case, which provides the best performance for URA over i.i.d. Rayleigh fading channels for a receiver equipped for a massive number of antennas.

While most of the works consider the transmission over i.i.d. Rayleigh fading channels, the more practical Rician fading case is considered in [64] through an application of coded compressed sensing. A correlated fading scenario is tackled in [65]. Moreover, the authors of [66] study the asynchronous scenario and propose

a slotted concatenated coding scheme based on the SPARCs and tree codes in an OFDM system.

The fundamental information theoretic limits on the performance of URA with a massive MIMO receiver are studied in [67], where the authors derive achievability and converse bounds for the required E_b/N_0 for a given PUPE for URA over quasi-static Rayleigh fading channels. The converse bound is stated below.

Theorem 3 (**Converse bound [67]**). *The minimum energy-per-bit for the URA model described in Section II of [67] can be lower bounded as*

$$E_b^*(n, M, \epsilon) \geq \inf_{P>0} \frac{nP}{J}, \tag{2.14}$$

where $J = \log_2 M$ is the number of message bits, M is the cardinality of the message set. The inf is taken over all $P > 0$ satisfying

$$J - \log_2 K_a \leq -\log_2 \mathbb{P}\left[\tilde{\chi}^2(2L) \geq (1 + (n+1)P)r\right], \tag{2.15}$$

where L is the number of receive antennas and r is the solution of $\mathbb{P}\left[\tilde{\chi}^2(2L) \leq r\right] = \epsilon$. ■

An achievability bound has also obtained in [67, Theorem 1]. The idea is to obtain an upper bound on the minimum required energy-per-bit by utilizing the random coding and ML decoding, and applying Fano's "good region" technique. The details of the theorem and its proof can be found in [67].

We plot the required E_b/N_0 with respect to the number of active users K_a for the achievability and converse bounds presented in [67] and SKP coding based scheme in [63] since it offers the best performance among the proposed practical approaches for this setup. The results in Figure 2.8 show that SKP coding based scheme performs very close to the achievability bound; however, there is still some room for improvement for the case of higher user loads. Moreover, the gap between achievability and converse bounds is less than 5 dB for $K_a \leq 1000$.

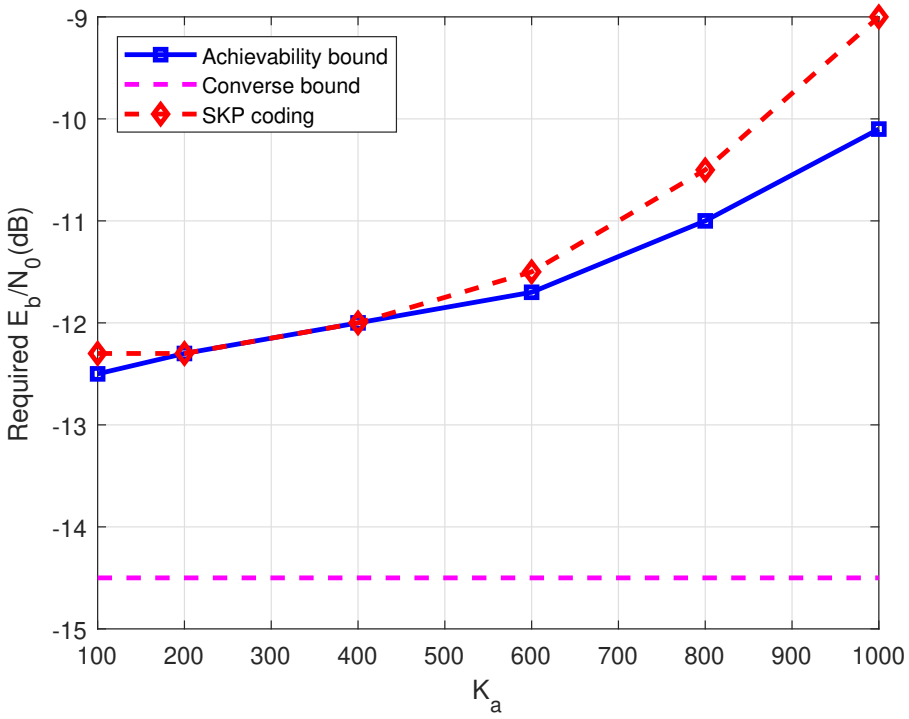

Figure 2.8: The information theoretic limits and the performance of the scheme in [63] for 50 receive antennas and $\epsilon = 0.05$.

2.5 Chapter Summary

In this chapter, a review on the existing literature on the multiple access and unsourced random access is presented. We first briefly reviewed the coordinated, grant-free and uncoordinated multiple access techniques. We then summarized the information theoretic performance limits and existing practical coding solutions on URA over Gaussian and fading MACs. Finally, we concluded the chapter with a review of URA with a massive MIMO receiver. In the following chapters, we propose new solutions for different practical scenarios in the context of URA.

Chapter 3

Unsourced Random Access over Frequency-Selective Channels

In this chapter, we propose and study URA over frequency-selective channels, which constitute more practical models for certain environments, e.g., for urban areas. We assume that the users experience independent multipath fading with arbitrary path delays and path gains, and propose to employ a slotted transmission scheme using OFDM. We develop a practical orthogonal matching pursuit (OMP) based joint activity detection and channel state information (CSI) estimation method utilizing the sparsity of the multipath channel via compressed sensing and use a TIN-SIC based detection at the receiver side. We also consider grant-based FDMA for comparison, and derive achievability bounds for multipath fading using normal approximations. Numerical examples show that the proposed URA scheme has a competitive performance with the ultimate case, namely, the grant-based FDMA; and that, there is a relatively small gap with its finite blocklength (FBL) performance bounds. Also, the performance loss is less than 2 dB with the estimated CSI at the receiver when the number of active users is not very large.

The rest of the chapter is organized as follows. In Section 3.1, we introduce the

system model. The proposed scheme is described in Section 3.2, and the grant-based FDMA is presented in Section 3.3. Section 3.4 consists of some numerical results, and Section 3.5 concludes the chapter.

3.1 System Model

We consider a massive random access scheme with K_a active users out of K_{tot} users ($K_a \ll K_{tot}$), each transmitting B bits of information through n uses of the channel. We assume that the BS is equipped with a single antenna, and the users encode their messages with the same channel code. The encoded signals are transmitted over a multipath fading channel, where the channel taps are zero mean complex Gaussian, i.e., Rayleigh fading. Each user has K channel taps, and the gain and delay of its i-th tap are denoted as h_i and τ_i, respectively. We assume that the channel taps do not change over each frame, i.e., slow fading scenario, and n channel resources (in the frequency domain) are divided into S slots of length n_1, and each active user independently transmits its signal in a randomly selected (frequency) slot.

In order to address the effects of multipath fading, we employ OFDM, i.e., different channel uses correspond to different subchannels in the frequency domain. Assuming that the cyclic prefix (CP) length is longer than the maximum delay spread, there is no interference between consecutive OFDM words. A high-level illustration is depicted in Figure 3.1. The transmitted time-domain signal of a user (without the CP) is given by [54]

$$x(t) = \sum_{l=1}^{n_1} X_l e^{2\pi j \Delta f (l - \frac{n_1}{2})t}, \quad t \in \left[0, \frac{1}{\Delta f}\right], \tag{3.1}$$

where X_l is the l-th element of the transmitted codeword of the user, and Δf is the subcarrier spacing, i.e., the symbol duration is $1/\Delta f$. The received signal in the frequency domain $\mathbf{Y} \in \mathbb{C}^{n_1 \times 1}$ is then

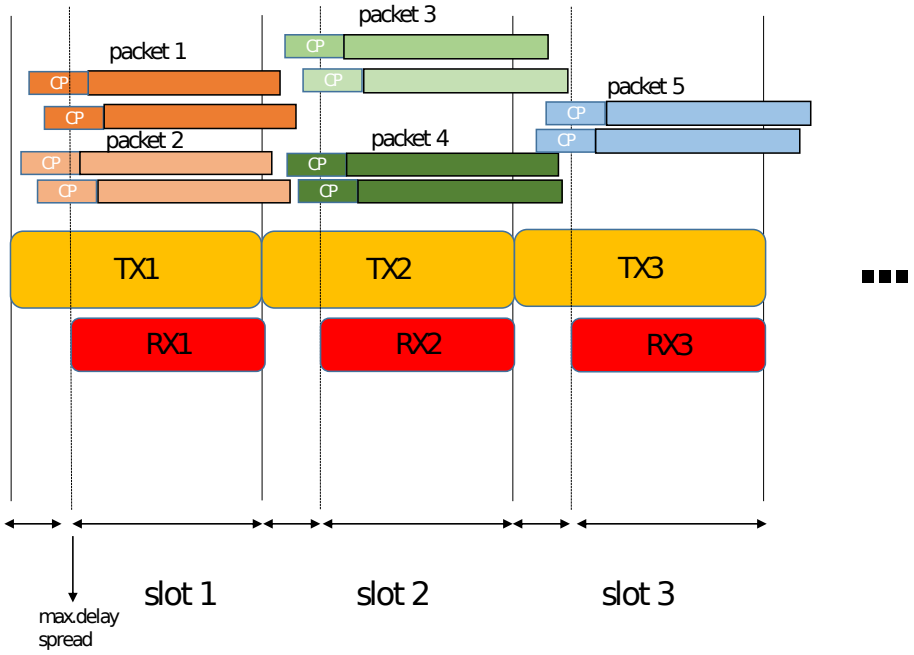

Figure 3.1: Transmission system model for two-tap multipath fading with a cyclic prefix longer than maximum delay spread.

$$\mathbf{Y} = \sum_{i=1}^{r} \mathbf{H}_i \odot \mathbf{X}_i + \mathbf{Z}, \tag{3.2}$$

where r is the number of users transmitting in the slot under consideration, $\odot$ denotes the element-wise multiplication, $\mathbf{X}_i \in \mathbb{C}^{n_1 \times 1}$ is the transmitted signal of the i-th user, $\mathbf{H}_i \in \mathbb{C}^{n_1 \times 1}$ is the vector of the subchannel gains, and $\mathbf{Z} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{n_1})$ is circularly symmetric complex additive white Gaussian noise (AWGN). Note that the slot index is dropped in the rest of the letter to simplify the notation. The elements of $\mathbf{X}_i$ have average power P; i.e., the total transmitted power is $(n_1 + L)P$, where L is the CP length in terms of the time-domain samples.

The decoder aims to recover the list of messages $\mathcal{L}(y) = \{\hat{\mathbf{w}}_1, \ldots, \hat{\mathbf{w}}_{K_a}\}$ up to a permutation. The PUPE defined as

$$P_e = \frac{1}{K_a} \sum_{i=1}^{K_a} Pr(\mathbf{w}_i \notin \mathcal{L}(y)) \tag{3.3}$$

is used as the performance metric. The energy efficiency of the scheme is measured by the required energy-per-bit, which can be calculated as

$$\frac{E_b}{N_0} = \frac{(n_1 + L)P}{B} \tag{3.4}$$

to obtain $P_e \leq \epsilon$, where ϵ is the target PUPE.

3.2 Coding Scheme

3.2.1 Encoding

In this section, we describe the encoding process in each slot. We assume that each active user picks a pilot signature from a common codebook $\mathbf{A} \in \mathbb{R}^{n_p \times M_s}$ based on the first J bits of its message $\mathbf{w}$ similar to [76], where $M_s = 2^J$ and n_p is the pilot length, and distributes it uniformly across the OFDM word. Namely, part of the subcarriers are allocated for pilot transmission, and the rest of them are used to transmit the data. The elements of $\mathbf{A}$ are drawn from a standard normal distribution and its columns are normalized as $\|\mathbf{a}_i\|^2 = n_p P, \quad i = 1, 2, \ldots M_s$. Note that in URA, it is impossible to assign dedicated pilots to the users since there are a massive number of them, hence there will be collisions when the pilot bits of the different users are selected identically. Denoting the pilot locations by a vector $\mathbf{p}$, the received signal at the pilot locations in a slot becomes

$$\mathbf{Y_p} = \sum_{k=1}^{r} diag(\mathbf{A}_k)\mathbf{H}_{\mathbf{p},k} + \mathbf{z_p} = \sum_{k=1}^{r} diag(\mathbf{A}_k)\mathbf{F}_{[\mathbf{p},1:L]}\mathbf{h}_k + \mathbf{z_p}, \tag{3.5}$$

where $\mathbf{A}_k$ is the pilot selected by the k-th user, $\mathbf{H}_{\mathbf{p},k}$ is the vector of the subchannel gains of the k-th user specified by the vector $\mathbf{p}$, $\mathbf{z_p}$ is the AWGN vector, $\mathbf{F}_{[\mathbf{p},1:L]}$ is the submatrix obtained by taking the first L columns and rows of the n_1-point discrete Fourier transform (DFT) matrix $\mathbf{F}$ corresponding to the pilot locations, and $\mathbf{h}_k$ is the channel vector of the k-th user in the time domain.

The remaining $k-J$ message bits are encoded by an $(n_c, k-J+c)$ polar code and modulated using BPSK to obtain the transmitted symbol vector $\mathbf{s}$, where c is the number of cyclic redundancy check (CRC) bits. Denoting the data locations by a vector $\mathbf{d}$, the received signal at the data subcarriers in a slot can be written as

$$\mathbf{Y_d} = \sum_{k=1}^{r} diag(\mathbf{s}_k)\mathbf{H}_{\mathbf{d},k} + \mathbf{z_d}, \tag{3.6}$$

where $\mathbf{s}_k$ is the symbol vector of the k-th user, $\mathbf{H}_{\mathbf{d},k}$ is the vector of the subchannel gains specified by $\mathbf{d}$, and $\mathbf{z_d}$ is the AWGN vector.

3.2.2 Activity Detection and CSI Estimation

At the receiver side, we first employ a compressed sensing-based joint activity detection and CSI estimation algorithm to obtain the channel taps of the active users along with the selected signatures. Note that estimating the channel coefficients on the pilot subcarriers and obtaining the remaining ones by linear interpolation is also possible; however, its performance would be inferior due to the multiuser interference. An example of the subchannel gains in the frequency domain due to the multipath channel is depicted in Figure 3.2.

We estimate the channel taps of the users from the most significant to the least significant one by employing an OMP-type solution which is similar to the one used in [68]. For this purpose, we first extend the pilot codebook by taking an elementwise multiplication of each pilot signature by the columns of the submatrix $\mathbf{F}_{[\mathbf{p},1:L]}$ to get an observation matrix $\mathbf{A}_x = [\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_L]$, where $\mathbf{A}_1 = [\mathbf{a}_1 \odot \mathbf{F}_{[\mathbf{p},1]}, \mathbf{a}_2 \odot \mathbf{F}_{[\mathbf{p},1]}, \ldots, \mathbf{a}_{M_s} \odot \mathbf{F}_{[\mathbf{p},1]}]$.

We then perform a joint iterative estimation of the selected pilot sequences and multipath channel vectors in the time domain. At each iteration, we first multiply the observation matrix with the residual pilot signal as in [68]

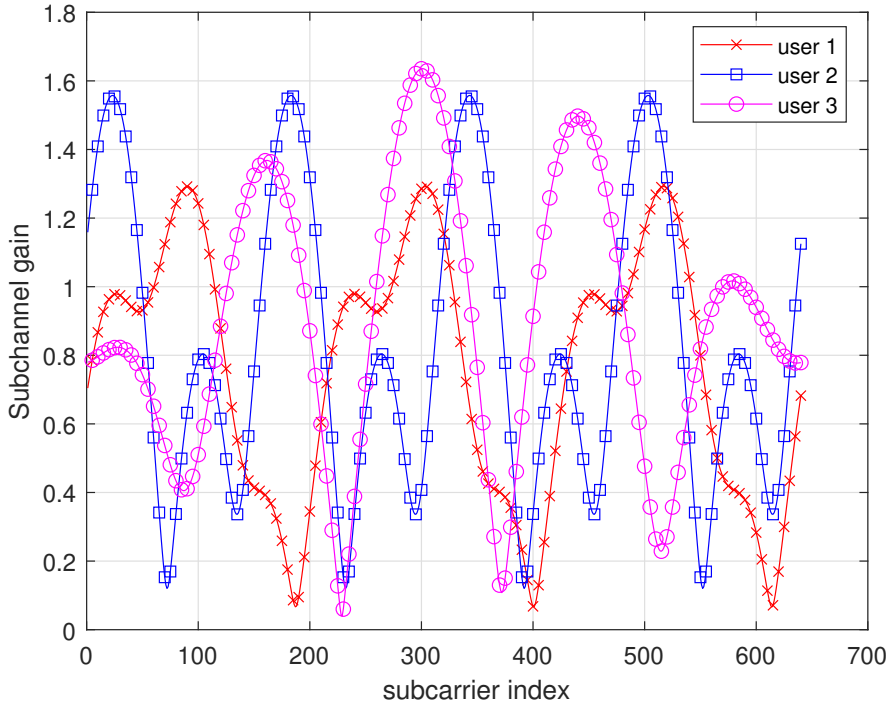

Figure 3.2: Frequency domain subchannel gain magnitudes of the multipath channels of three different users with 3 time-domain channel taps.

$$\mathbf{R} = \left| \mathbf{A}_x^H \mathbf{Y_p}^{(j)} \right|, \tag{3.7}$$

to obtain an $M_s L \times 1$ vector $\mathbf{R}$ where $|\mathbf{x}|$ denotes the elementwise absolute values of the vector $\mathbf{x}$, and $\mathbf{Y_p}^{(j)}$ denotes the residual signal at the j-th step which is initialized as $\mathbf{Y_p}^{(1)} = \mathbf{Y_p}$. Then, the index of the maximum element of $\mathbf{R}$, I_t gives the selected signature and delay of the most significant tap. Namely, the index of the pilot signature estimate can be obtained as $\hat{a}_t \equiv I_t \mod M_s$, and the delay estimate becomes $\hat{d}_t = \lceil I_t / M_s \rceil$, where $\mod$ and $\lceil . \rceil$ denote the modulo and ceiling operations. Adding I_t to the set of detected indices $\hat{I}$, the estimates of path gains of the channel taps at the j-th step are obtained as

$$\hat{\mathbf{h}}_j = ((\mathbf{A}_x)_{[:,\hat{I}]})^+ \mathbf{Y}_p, \tag{3.8}$$

where $(\mathbf{A}_x)_{[:,\hat{I}]}$ denotes the set of collections of columns of $\mathbf{A}_x$ in $\hat{I}$ and $\mathbf{A}^+$ is the pseudo-inverse of $\mathbf{A}$. We then subtract the effect of the current set from the residual as

$$\mathbf{Y_p}^{(j+1)} = \mathbf{Y_p} - (\mathbf{A}_x)_{[:,\hat{I}]} \hat{\mathbf{h}}_j, \tag{3.9}$$

and continue the iterations until $\min_{\forall i \in \left|\hat{I}\right|} \left|\hat{h}_j(i)\right| < \delta$ where $\left|\hat{I}\right|$ is the cardinality of the set $\hat{I}$ as the receiver does not know the actual number of taps and δ is a threshold on the absolute values of the channel taps. We then collect the channel taps corresponding to the same signature to obtain the channel vectors of the users in time-domain, and use their DFT in the decoding algorithm which is described in the next subsection.

3.2.3 TIN-SIC Decoding Algorithm

We adopt the TIN-SIC coupled with a polar successive cancellation list decoder (SCLD) with CRC as the decoding approach. This scheme is particularly suitable for fading scenarios as it allows the receiver to decode the messages of users with higher Euclidean-norm channel vectors, subtract their effects and decode the remaining users in subsequent iterations.

In order to decode the message of a particular user for a given slot, we first eliminate the multipath fading effect using the CSI estimate of the users. Let $\hat{\mathbf{H}}$ denote the channel vector estimate of the user to be decoded and $\mathbf{Y_d}$ be the received signal at the data locations in the frequency domain, then the l-th element of the received signal of the data part can be written as

$$Y^l_{\mathbf{d}} = \hat{H}^l_{\mathbf{d}} s^l + \sum_{i=2}^{r} \hat{H}^l_{\mathbf{d},i} s^l_i + Z^l, \tag{3.10}$$

where $\hat{H}^l_{\mathbf{d},i}$ is the estimate of the l-th channel coefficient of the data part of the channel vector of the i-th user in the frequency domain and s^l_i is the l-th element of the symbol vector of the i-th user. Rewriting $\hat{H}^l_{\mathbf{d}} = \left|\hat{H}^l_{\mathbf{d}}\right| e^{j\angle \hat{H}^l_{\mathbf{d}}}$, de-phasing $Y^l_{\mathbf{d}}$ with $e^{-j\angle \hat{H}^l_{\mathbf{d}}}$ and taking the real part, (3.10) becomes

$$Y^l_{\mathbf{d}} = \left|\hat{H}^l_{\mathbf{d}}\right| s^l + \sum_{i=2}^{r} \left|\hat{H}^l_{\mathbf{d},i}\right| \cos\left(\angle \hat{H}^l_{\mathbf{d},i} - \angle \hat{H}^l_{\mathbf{d}}\right) s^l_i + Z^l. \tag{3.11}$$

We can then compute the log-likelihood ratio (LLR) of s^l by treating the interference as noise. Once the LLRs are calculated, a single-user polar decoder is employed. If the decoded sequence satisfies the CRC check, we conclude that it is successfully decoded and add it to the output list $\hat{\mathcal{D}}^{(j)}$. At the end of a decoding attempt, the effect of the transmitted signals of the successfully decoded users are subtracted from the received signal as follows

$$\mathbf{Y}^{(j+1)} = \mathbf{Y}^{(j)} - \sum_{k \in \hat{\mathcal{D}}^{(j)}} \hat{\mathbf{H}}_k \odot \hat{\mathbf{X}}_k, \tag{3.12}$$

where $\hat{\mathbf{X}}_k$ is the re-constructed transmitted signal of the k-th user, $\hat{\mathbf{H}}_k$ is the DFT of the channel vector estimate of the k-th user, and $\mathbf{Y}^{(j)}$ is the residual signal in the j-th decoding attempt. We set $\mathbf{Y}^{(1)} = \mathbf{Y}$. The decoding attempts continue until there is no improvement between two consecutive attempts or the maximum number of iterations $n_{\max}$ is reached. A pseudo-code of the receiver operation is provided in Algorithm 1.

3.2.4 Complexity Analysis

We can assess computational complexity of the proposed scheme by computing the average number of multiplications for decoding of messages of active users in each slot. The complexity of the joint activity detection and CSI estimation algorithm is dominated by the correlation step in Eq. (3.7), with a complexity of $\mathcal{O}(M_s L n_p)$. The complexity of single-user polar decoding is $\mathcal{O}(r\, n_L n_d \log n_d)$ and the elementwise multiplication in Eq. (3.12) has a complexity of $\mathcal{O}\left(\left|\hat{\mathcal{D}}^{(j)}\right| n_1\right)$. Therefore, the overall complexity of the proposed scheme is $\mathcal{O}(M_s L n_p)$, as it is dominated by the complexity of the correlation step of the joint activity detection and CSI estimation method.

Algorithm 1 Receiver operation of the proposed scheme.

Input: $\mathbf{Y}_p$, $\mathbf{Y}_d$, $\mathbf{A}_x$, δ, $n_{\max}$
Joint activity detection and CSI estimation:
Initialize $\hat{I} = \emptyset$
for $j = 1, 2, \ldots$ **do**
    $\mathbf{R} = \left| \mathbf{A}_x^H \mathbf{Y_p}^{(j)} \right|$.
    Set the index of the maximum element of $\mathbf{R}$ to I_t and add it to $\hat{I}$.
    $\hat{\mathbf{h}}_j = ((\mathbf{A}_x)_{[:,\hat{I}]})^+ \mathbf{Y}_p$.
    If the termination condition is satisfied, go Step 11.
    Otherwise, $\mathbf{Y_p}^{(j+1)} = \mathbf{Y_p} - (\mathbf{A}_x)_{[:,\hat{I}]} \hat{\mathbf{h}}_j$
end for
Collect the channel taps corresponding to the same signature to obtain the channel vectors.
TIN-SIC Decoding:
for $j = 1, 2, \ldots n_{\max}$ **do**
    Compute the LLRs by TIN.
    Decode the users one-by-one.
    if $\left| \hat{\mathcal{D}}^{(j)} \right| = 0$ **then**
        Terminate the algorithm.
    end if
    Perform SIC by (3.12).
end for
Output: List of messages

3.3 Comparison with Grant-Based FDMA

In order to assess the performance of the proposed URA scheme, we compare it with that of a grant-based FDMA approach. In the grant-based mechanism, a prior handshaking process is performed between the users and the BS, which enables the assignment of fixed transmission resources for each user. This kind of mechanism is not feasible for massive random access since it would lead to an excessive delay and signaling overhead, which may motivate grant-free random access such as the one in [69]. In the grant-based FDMA, the subcarriers are divided into non-overlapping channels and each user transmits its data in its dedicated set of subcarriers through proper user scheduling. Despite being infeasible for unsourced MAC, we employ the grant-based FDMA for comparison as it represents the best-case scenario.

We implement the FDMA scheme by dividing the available number of subchannels to the number of active users and utilizing a practical channel code matching the slot length. That is, in each slot, we consider

$$\mathbf{Y} = \mathbf{H} \odot \mathbf{X} + \mathbf{Z}, \tag{3.13}$$

where $\mathbf{H} = [H_1, H_2, \ldots, H_{n_1}]$ is the channel vector with n_1 being the number of channel resources in each slot, $\mathbf{X}$ is the transmitted signal, and $\mathbf{Z}$ is the Gaussian noise. Conditioned on the channel gains (i.e., the gains of different subchannels in our set-up), using the formulation for parallel Gaussian channels, the codeword error probability for a given information rate using normal approximation is given by

$$P_e \approx Q\left(\frac{C - R}{\sqrt{V/n_c}}\right) \tag{3.14}$$

where R is the code rate, C is the channel capacity, V is the channel dispersion, P_e is the block error rate and n_c is the code blocklength [70]. The channel capacity and dispersion for parallel AWGN channels are given in [71] as follows

$$C = \frac{1}{n_1}\sum_{k=1}^{n_1}\log(1+\rho_k), \quad V = \frac{1}{n_1}\sum_{k=1}^{n_1}\frac{\rho_k(\rho_k+2)}{(\rho_k+1)^2}\log^2(e), \tag{3.15}$$

where $\rho_k = \rho|H_k|^2$ and $\rho = \frac{E_b}{N_0}$. Clearly, the codeword error probability will be different for different realizations of the channel. To determine the block error probability as a function of the transmission rate, we average this quantity over the channel realizations. That is, we compute

$$P_{e,avg} \approx E_H\left[Q\left(\frac{C - R}{\sqrt{V/n_c}}\right)\right] \tag{3.16}$$

where the expectation is taken over the random channel taps in time-domain. We evaluate this expectation through Monte Carlo simulations by generating

many realizations of the channel taps according to the multipath fading statistics, determining the corresponding subchannel gains in the frequency domain, and taking the average of the resulting block error probability realizations.

3.4 Numerical Results

To exemplify the proposed URA scheme over frequency-selective channels, we consider a transmission bandwidth of 10 MHz and assume that the multipath delay spread is 2 μs [72]. We employ a CP length of 2 μs to make sure that there is no ISI between the OFDM words of consecutive slots. We choose $n = 30000$ and $B = 100$ as in [2]. We consider two scenarios: (i) all the users' channels have two taps; (ii) the number of channel taps is taken uniformly and randomly between 2 and 5. We consider a slot length of 66 μs for scenario (i) and 78.8 μs for scenario (ii), corresponding to 640 and 768 subchannels in frequency domain, respectively. We take $M_s = 2^{10}$, utilize a polar code with a length of 512, and set the list size to 128 and the CRC length to 11. The path gains are distributed as $h_i \sim \mathcal{CN}(0, 1/K)$, and the channel taps are located randomly between 0 and L (in terms of the time-domain samples).

We run the CSI estimation algorithm for both scenarios (i) and (ii) with $\delta = 0.23$, selected to minimize the probability of missing channel taps without substantially increasing the complexity. Figure 3.3 demonstrates that doubling the pilot length improves the MSE by up to 8 dB for scenario (i) and up to 10 dB for scenario (ii). Based on this result, for the rest of the examples, we take $n_p = 128$ for scenario (i) and $n_p = 256$ for scenario (ii) for channel estimation as they provide a good CSI estimation performance with an acceptable amount of overhead.

We measure the energy efficiency of the proposed scheme by the minimum required energy to serve K_a active users with a target PUPE of 0.1. We assume that the decoder tries to resolve simultaneous transmissions in a slot regardless of the number of them, and we compare the performance of the proposed scheme

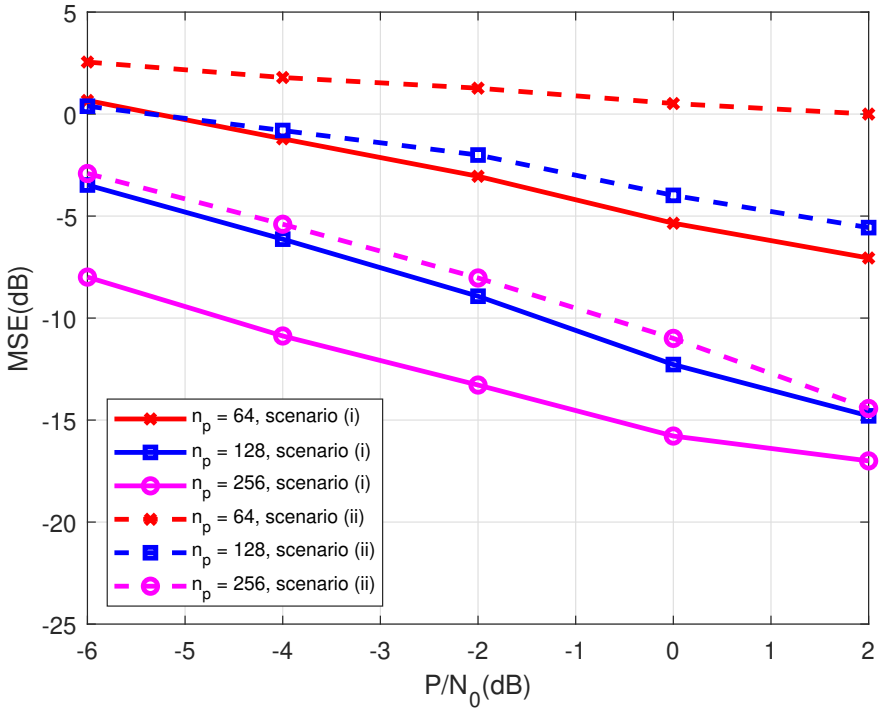

Figure 3.3: The MSE performance of the channel estimation algorithm for different pilot lengths.

with that of the grant-based FDMA and its performance limits. For FDMA, the polar code length and modulation are chosen based on the number of available subcarriers for each user.

The results in Figure 3.4 and Figure 3.5 illustrate that the proposed unsourced MAC with TIN-SIC utilizing the polar codes offers a competitive performance with the grant-based FDMA for the perfect CSI case in both scenarios, even though there is coordination among the users in the latter. While the grant-based FDMA is infeasible for URA, we use it for comparison purposes since there is no other URA scheme over frequency-selective channels for direct comparison. The performance gap with the ultimate performance offered by FDMA varies between 2-6 dB for both scenarios. For instance, for $K_a = 250$, the difference between the unsourced MAC performance and the ultimate limit of FDMA is only around 2 dB. We also study the performance of LDPC codes, and observe that polar codes outperform them by 2-6 dB in scenario (i), and 2-8 dB in scenario (ii). The results also show that with practical CSI estimation at the receiver, the extra required energy is about 2 dB or less for $K_a \leq 200$ for scenario (i), and 2-3 dB for scenario (ii).

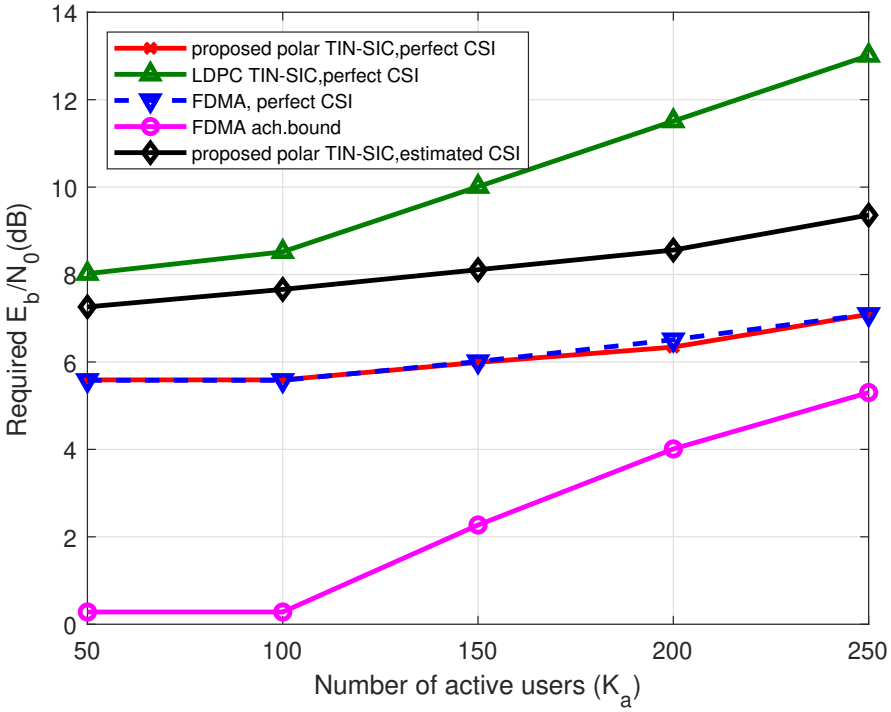

Figure 3.4: Minimum required E_b/N_0 for $\epsilon \leq 0.1$, $n = 30000$, $B = 100$ with different schemes for scenario (i).

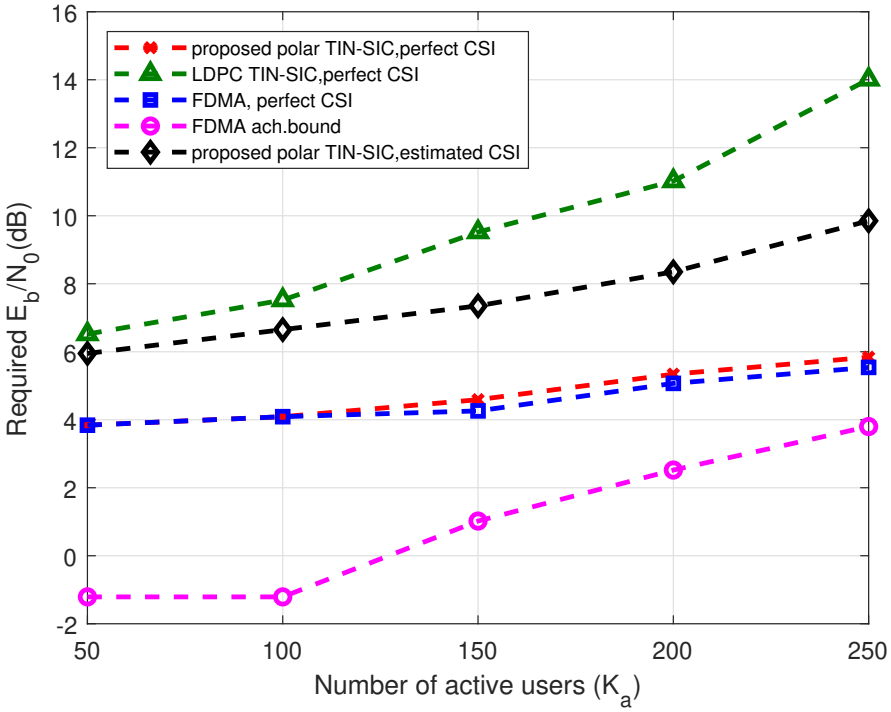

Figure 3.5: Minimum required E_b/N_0 for $\epsilon \leq 0.1$, $n = 30000$, $B = 100$ with different schemes for scenario (ii).

3.5 Chapter Summary

We study URA over frequency-selective channels and propose to employ OFDM to overcome the deleterious effects of multipath fading. The receiver utilizes a compressed sensing-based joint activity detection and CSI estimation method which exploits the sparsity of the multipath channel followed by TIN-SIC, and polar coding is employed for channel coding. Our results demonstrate that the performance of the proposed scheme with TIN-SIC is highly competitive even with that of the grant-based FDMA, which requires coordination.

Chapter 4

Unsourced Random Access with a Massive MIMO Receiver

In this chapter, we consider URA over quasi-static Rayleigh fading channels, and develop a low complexity and high performing solution based on slotted transmissions to reduce the multiuser interference, called slotted non-orthogonal pilot-based unsourced random access (SNOP-URA). We employ generalized orthogonal matching pursuit (gOMP) [73] for AD assuming a massive number of antennas at the BS. The slot length is smaller than the channel coherence time. Each user transmits a pilot sequence from a common non-orthogonal codebook followed by its polar encoded and modulated codeword. At the receiver, we first detect the active pilots by gOMP, and then estimate the channel vectors by employing LMMSE in each slot. Next, we perform iterative decoding by estimating the user symbols with MRC, passing these estimates to a polar decoder, and subtracting the effects of the successfully decoded messages by SIC. Note that in the SIC step, instead of using the initially estimated channel vectors, we re-estimate the channel vectors for the successfully decoded users employing both the pilots and the decisions on data bits by an LMMSE filter, and use the newly estimated channel vectors for SIC. We also characterize the performance of the proposed scheme analytically using normal approximations and provide a detailed complexity analysis. Numerical examples demonstrate that our newly proposed

scheme either outperforms the existing works or offers a competitive performance with a lower complexity. Besides, it is suitable for fast fading scenarios due to its excellent performance in the short blocklength regime.

The rest of the chapter is organized as follows. We describe the system model in Section 4.1, and the proposed scheme in Section 4.2. We present a performance analysis in Section 4.3, a set of numerical results in Section 4.4, and conclude the chapter in Section 4.5.

4.1 System Model

We consider a massive random access scheme where K_a out of K_{tot} active users ($K_a \ll K_{tot}$) transmit B bits of information over a quasi-static fading channel to a common BS equipped with M antennas. We assume that K_a is not known at the receiver. We divide the length-n channel frame into V slots of length L, and assume that the channel vectors of the users remain constant over each slot. Each user transmits its signal in a randomly selected slot. Then, the received signal in the s-th slot can be written as

$$\mathbf{Y}_s = \sum_{k \in \mathcal{K}_s} \mathbf{x}(\mathbf{m}_k)\mathbf{h}_k + \mathbf{Z}_s, \tag{4.1}$$

where $\mathbf{Y}_s \in \mathbb{C}^{L \times M}$, $\mathcal{K}_s$ is the set of active users transmitting in the s-th slot, $\mathbf{m}_k$ is the message of the k-th user, $\mathbf{x}(\mathbf{m}_k) \in \mathbb{C}^{L \times 1}$ is the transmitted signal of the k-th user corresponding to the message $\mathbf{m}_k$, $\mathbf{h}_k \in \mathbb{C}^{1 \times M}$ is the channel vector of the k-th user with independent and identically distributed (i.i.d.) elements $h_{k,i}$ with zero mean and unit variance, i.e., $h_{k,i} \sim \mathcal{CN}(0,1)$, and $\mathbf{Z}_s$ is the circularly symmetric complex AWGN with i.i.d. zero mean elements with variance N_0. Assuming that the elements of $\mathbf{x}(\mathbf{m}_k)$ have an average power of P, we have

$$\frac{E_b}{N_0} = \frac{LP}{BN_0}, \tag{4.2}$$

where E_b is the energy-per-bit. The receiver produces a list of the decoded messages $\mathcal{L}$. Let $n_{\text{md}} = \sum_{k \in \mathcal{K}_a} \mathbb{1}_{\{\mathbf{m}_k \notin \mathcal{L}\}}$ be the number of misdetections and $n_{\text{fa}} = |\mathcal{L} \setminus \{\mathbf{m}_k : k \in \mathcal{K}_a\}|$ be the number of false alarms, where $\mathcal{K}_a$ is the set of active users, $|.|$ denotes the cardinality of a set and $\mathbb{1}_{\{.\}}$ is the indicator function. Then, the misdetection probability P_{md} and the false alarm probability P_{fa} are defined as

$$P_{\text{md}} = \frac{\mathbb{E}[n_{\text{md}}]}{K_a}, \;\; P_{\text{fa}} = \mathbb{E}\Big[\frac{n_{\text{fa}}}{|\mathcal{L}|}\Big], \tag{4.3}$$

where $|\mathcal{L}| = K_a - n_{\text{md}} + n_{\text{fa}}$ is the size of the output list $\mathcal{L}$. The PUPE of the system P_e is defined as the sum of misdetection and false alarm probabilities, i.e., $P_e = P_{\text{md}} + P_{\text{fa}}$. Our objective is to design a URA scheme minimizing the required $\frac{E_b}{N_0}$ with $P_e \leq \epsilon$, where ϵ is the target PUPE of the system.

4.2 Proposed Scheme

4.2.1 Encoding

We first divide the transmission slot into pilot and data parts of lengths n_p and n_d, respectively. In the pilot phase, each user picks one of the non-orthogonal pilot signatures from a common codebook $\mathbf{A} \in \mathbb{C}^{n_p \times N}$ based on the first J message bits $\mathbf{m}_p$ with the corresponding pilot index $b(\mathbf{m}_p)$, where $N = 2^J \gg K_a$. The columns of $\mathbf{A}$ are normalized as $\|\mathbf{A}_i\|^2 = n_p P_p$, where P_p is the average symbol power of the pilot sequence. The received signal of the pilot part in the s-th slot can be written as

$$\mathbf{Y}_{s,p} = \mathbf{A_s H} + \mathbf{Z}_{s,p}, \tag{4.4}$$

where $\mathbf{Y}_{s,p} \in \mathbb{C}^{n_p \times M}$ corresponds to the first n_p rows of $\mathbf{Y}_s$, $\mathbf{A}_s \in \mathbb{C}^{n_p \times K_s}$ is a submatrix of $\mathbf{A}$ including the selected pilot sequences in its columns, $\mathbf{H} =$

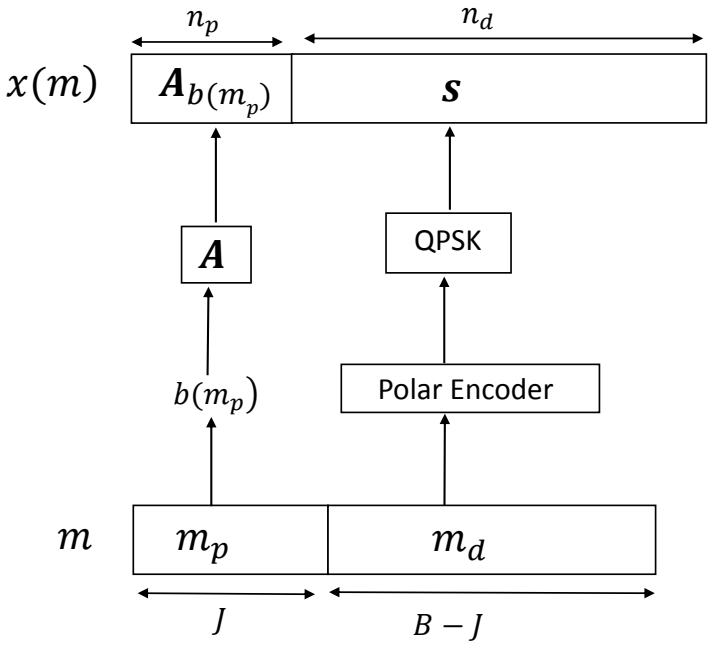

Figure 4.1: Encoding process of the proposed scheme for a user.

$[\mathbf{h}_1^T, \mathbf{h}_2^T, \ldots, \mathbf{h}_{K_s}^T]^T \in \mathbb{C}^{K_s \times M}$ is the channel matrix including channel vectors of the users transmitting in the s-th slot, K_s is the number of transmitting users in the s-th slot, and $\mathbf{Z}_{s,p}$ is the first n_p rows of $\mathbf{Z}_s$.

The remaining $B - J$ message bits $\mathbf{m}_d$ are encoded using a $(2n_d, B - J + r)$ polar code and modulated using quadrature phase shift keying (QPSK), where r is the number of cyclic redundancy check (CRC) bits. Then, the received signal of the data part in the s-th slot can be written as follows:

$$\mathbf{Y}_{s,d} = \sum_{k \in \mathcal{K}_s} \mathbf{s}_k \mathbf{h}_k + \mathbf{Z}_{s,d}, \tag{4.5}$$

where $\mathbf{Y}_{s,d} \in \mathbb{C}^{n_d \times M}$ corresponds to the last n_d rows of $\mathbf{Y}_s$, $\mathbf{s}_k$ is the $n_d \times 1$ vector with elements in $\{\sqrt{P_d/2}(\pm 1 \pm j)\}$ denoting the transmitted symbols of the k-th user, P_d is the average symbol power of the data part, and $\mathbf{Z}_{d,s}$ is the last n_d rows of $\mathbf{Z}_s$. The encoding process is depicted in Figure 4.1, where the user index is dropped for ease of presentation.

4.2.2 Activity Detection and Decoding

At the receiver side, we perform an iterative decoding by first detecting the set of the selected pilot sequences $\mathcal{I}$ and estimating the corresponding channel vectors followed by the decoding of the data part in each slot, and SIC at the end of each iteration. In the following, we explain these processes in detail. Note that we drop the slot index s for the rest of the subsection to simplify the notation.

We first need to detect the selected pilots in each slot, which is an activity detection problem to identify the selected columns of $\mathbf{A}$. Since $K_a \ll N$, this problem can be regarded as a compressed sensing problem with $\mathbf{A}$ as the measurement matrix. We propose to employ the gOMP algorithm [73] to solve this problem, which is a lower complexity generalization of orthogonal matching pursuit (OMP) [74]. OMP is a greedy iterative algorithm to solve sparse recovery problems. The OMP algorithm selects the column of the measurement matrix having the maximum correlation with the current residual signal at each iteration and adds its index to the output list. On the other hand, in gOMP, multiple indices can be added to the list of the detected indices at each iteration.

In order to implement the gOMP algorithm, we first calculate the correlation between the columns of $\mathbf{A}$ and the current residual of the received pilot signal at each iteration as follows:

$$\mathbf{R}^{(t)} = \mathbf{A}^H \mathbf{Y}_p^{(t)}, \tag{4.6}$$

where $\mathbf{R}^{(t)} \in \mathbb{C}^{N \times M}$, and $\mathbf{Y}_p^{(t)}$ is the residual of received pilot signal at the t-th gOMP iteration. We initialize $\mathbf{Y}_p^{(0)} = \mathbf{Y}_p$, then, we calculate the Euclidean norm of each row of $\mathbf{R}^{(t)}$ to obtain an $N \times 1$ vector $\mathbf{R}_c^{(t)}$, where we exploit receive (antenna) diversity. We then take the $i_{\text{OMP}} = \left\lceil (\frac{\hat{K}_a}{V} + \Delta)/n_{\text{OMP}} \right\rceil$ largest elements of $\mathbf{R}_c^{(t)}$, where $\hat{K}_a$ is the estimate of the total number of active users, $\frac{\hat{K}_a}{V}$ is the estimated average active user load in a slot, Δ is an integer which may be taken sufficiently high (e.g., half of the estimated average active user load) to make the probability of undercounting the number of the transmitting users in the slot

negligible, and n_{OMP} is the number of gOMP iterations. Using the result in [30], we estimate the total number of active users as follows

$$\hat{K}_a = \frac{1}{M} \sum_{s=1}^{V} \sum_{i=1}^{M} \left\lfloor \frac{\|\mathbf{Y}_{s,i}\|^2 - LN_0}{n_p P_p + n_d P_d} \right\rceil, \tag{4.7}$$

where $\mathbf{Y}_{s,i}$ is the received signal of the i-th receive antenna in the s-th slot.

We then add the selected indices to the set of detected signature indices $\hat{\mathcal{I}}$ and subtract their effect from the residual as follows (unless the termination condition is satisfied as explained later):

$$\mathbf{Y}_p^{(t+1)} = \mathbf{Y}_p - \mathbf{A}_{\hat{\mathcal{I}}^{(t)}} (\mathbf{A}_{\hat{\mathcal{I}}^{(t)}}^H \mathbf{A}_{\hat{\mathcal{I}}^{(t)}} + N_0 \mathbf{I}_{\left|\hat{\mathcal{I}}^{(t)}\right|})^{-1} \mathbf{A}_{\hat{\mathcal{I}}^{(t)}}^H \mathbf{Y}_p, \tag{4.8}$$

where $\mathbf{A}_{\hat{\mathcal{I}}^{(t)}}$ denotes the submatrix of $\mathbf{A}$ obtained by retaining the columns in $\hat{\mathcal{I}}^{(t)}$, which is the set of detected signatures after the t-th gOMP iteration. In this approach, setting i_{OMP} to a high value by taking a large Δ may lead to a high number of overcounts. In order to reduce them, we check the $l1$-norm of each row of $(\mathbf{A}_{\hat{\mathcal{I}}^{(t)}}^H \mathbf{A}_{\hat{\mathcal{I}}^{(t)}} + N_0 \mathbf{I}_{\left|\hat{\mathcal{I}}^{(t)}\right|})^{-1} \mathbf{A}_{\hat{\mathcal{I}}^{(t)}}^H \mathbf{Y}_p^{(t)}$ at each iteration, and terminate the gOMP iterations if minimum of them divided by the vector length becomes less than a threshold δ to obtain the finalized set of detected signature indices $\hat{\mathcal{I}}$. Our extensive simulations show that setting δ to a proper value according to the active user load while taking Δ sufficiently high gives an accurate estimate of the number of active users in each slot, however, one can set δ to a very low value to minimize the misdetections without empirical simulations with a slight increase in complexity. Note that even if the users are overcounted, the corresponding decoded message in the data part is unlikely to satisfy the CRC check, which ensures that false alarms will not be a problem, while undercounting the users will lead to a significant number of misdetections. Also, note that we have changed the least squares (LS) criterion in the conventional OMP to MMSE in (4.8) to improve its performance. Note also that the AD can be performed at each decoding iteration, however, our extensive simulations demonstrate that this approach provides only a slight performance improvement while increasing the complexity.

Once the set of the detected indices is available, the matrix of the channel vectors of the active users can be estimated via LMMSE as

$$\hat{\mathbf{H}} = (\mathbf{A}_{\hat{\mathcal{I}}}^H \mathbf{A}_{\hat{\mathcal{I}}} + N_0 \mathbf{I}_{\hat{K}_s})^{-1} \mathbf{A}_{\hat{\mathcal{I}}}^H \mathbf{Y}_p^{(j)}, \tag{4.9}$$

where $\mathbf{Y}_p^{(j)}$ is the received pilot signal at the j-th decoding iteration which is initialized as $\mathbf{Y}_p^{(0)} = \mathbf{Y}_p$, and $\hat{K}_s$ is the cardinality of the (estimated) active user set $\hat{\mathcal{I}}$.

The decoding of the data part is done by first performing MRC to separate the users, and then passing the symbol estimates to a single-user polar decoder. So, the symbol estimates of the k-th user are obtained as follows

$$\hat{\mathbf{s}}_k = \mathbf{Y}_d^{(j)} \hat{\mathbf{h}}_k^H, \tag{4.10}$$

where $\mathbf{Y}_d^{(j)}$ is the data part of the received signal at the j-th decoding iteration, and $\hat{\mathbf{h}}_k$ is the channel vector estimate of the k-th user, which is the k-th row of $\hat{\mathbf{H}}$. We extract the bit-wise log-likelihood ratio (LLR) values $\hat{\beta}_k$ from the complex symbol estimate $\hat{\mathbf{s}}_k$ by treating it as the output of a virtual single-user channel with a signal-to-noise ratio (SNR) that is equal to the effective signal-to-interference-plus-noise ratio (SINR) of the k-th user, which can be well approximated by the following result.

Theorem 4. *The SINR corresponding to the k-th user at the input of the polar decoder in the first iteration γ_k' can be approximated as*

$$\gamma_k' \approx \frac{P_d \, c_1^{k,k}}{P_d \sum\limits_{i=1, i\neq k}^{K_s} c_1^{i,k} + N_0 c_2^k}, \tag{4.11}$$

where $c_1^{i,k} = \left|\mathbf{h}_i \mathbf{H}^H \mathbf{A}_{\mathcal{I}}^H \mathbf{C}_B^{-1} \mathbf{A}_{\mathcal{I}_k}\right|^2 + N_0 \|\mathbf{h}_i\|^2 \left\|\mathbf{C}_B^{-1} \mathbf{A}_{\mathcal{I}_k}\right\|^2$,
$c_2^k = \left\|\mathbf{H}^H \mathbf{A}_{\mathcal{I}}^H \mathbf{C}_B^{-1} \mathbf{A}_{\mathcal{I}_k}\right\|^2 + M N_0 \left\|\mathbf{C}_B^{-1} \mathbf{A}_{\mathcal{I}_k}\right\|^2$, and $\mathbf{C}_B = \mathbf{A}_{\mathcal{I}} \mathbf{A}_{\mathcal{I}}^H + N_0 \mathbf{I}_{n_p}$.

Proof. Plugging (4.5) into (4.10), the symbol vector estimate of the k-th user at the output of MRC can be written as follows

$$\hat{\mathbf{s}}_k = \mathbf{s}_k \mathbf{h}_k \hat{\mathbf{h}}_k^H + \sum_{i=1, i\neq k}^{K_s} \mathbf{s}_i \mathbf{h}_i \hat{\mathbf{h}}_k^H + \mathbf{Z}_d \hat{\mathbf{h}}_k^H, \tag{4.12}$$

where the first term is the desired signal, the second term is the multiuser interference, and the last one is the noise term. Then, the power of each term in (4.12) can be calculated as follows

$$\begin{aligned}
P'_S &= \frac{1}{n_d}\mathbb{E}\left[(\mathbf{s}_k \mathbf{h}_k \hat{\mathbf{h}}_k^H)^H \mathbf{s}_k \mathbf{h}_k \hat{\mathbf{h}}_k^H\right] = P_d\, c_1^{k,k} \\
P'_I &= \frac{1}{n_d} \sum_{\substack{i=1,\\ i\neq k}}^{K_s} \sum_{\substack{j=1,\\ j\neq k}}^{K_s} \mathbb{E}\left[\left(\mathbf{s}_i \mathbf{h}_i \hat{\mathbf{h}}_k^H\right)^H \mathbf{s}_j \mathbf{h}_j \hat{\mathbf{h}}_k^H\right] = P_d \sum_{\substack{i=1,\\ i\neq k}}^{K_s} c_1^{i,k} \\
P'_N &= \frac{1}{n_d}\mathbb{E}\left[\hat{\mathbf{h}}_k \mathbf{Z}_d^H \mathbf{Z}_d \hat{\mathbf{h}}_k^H\right] = N_0\, c_2^k
\end{aligned}$$

where $c_1^{i,k} = \mathbb{E}\left[\hat{\mathbf{h}}_k \mathbf{h}_i^H \mathbf{h}_i \hat{\mathbf{h}}_k^H\right]$, $c_2^k = \mathbb{E}\left[\hat{\mathbf{h}}_k \hat{\mathbf{h}}_k^H\right]$, and P'_S, P'_I, and P'_N are the powers of the desired signal, multiuser interference, and Gaussian noise after MRC, respectively.

In order to calculate $c_1^{i,k}$, we first need the following lemma.

Lemma 1. *Letting* $\mathbf{Q}$ *be an* $N \times M$ *random matrix with i.i.d. zero mean entries with variance* σ_q^2*, and* $\mathbf{B}$ *be an arbitrary matrix, we have* $\mathbb{E}\left[\mathbf{Q}^H \mathbf{B} \mathbf{Q}\right] = \sigma_q^2 \,\mathrm{Tr}\{\mathbf{B}\}\mathbf{I}_M$.

Proof. Letting $\mathbf{G} = \mathbb{E}\left[\mathbf{Q}^H \mathbf{B} \mathbf{Q}\right]$, we have

$$\begin{aligned}
\mathbf{G}_{i,j} &= \mathbb{E}\left[(\mathbf{Q}^H)_{(i,:)} \mathbf{B} (\mathbf{Q})_{(:,j)}\right] \\
&= \mathbb{E}\left[\sum_r \sum_s q_{i,r}^* \mathbf{B}_{rs} q_{s,j}\right] \\
&= \begin{cases} \sum_r \mathbb{E}\left[|q_{r,i}|^2\right] \mathbf{B}_{r,r} & \text{if } i = j \\ 0, & \text{otherwise} \end{cases} \\
&= \begin{cases} \sum_r \sigma_q^2 \mathbf{B}_{r,r} & \text{if } i = j \\ 0, & \text{otherwise} \end{cases}
\end{aligned} \tag{4.13}$$

■

Using Lemma 1, $c_1^{i,k}$ and c_2^k can be obtained as

$$\begin{aligned}
c_1^{i,k} &= \mathbb{E}\left[(\mathbf{A}_{\mathcal{I}_k}^H \mathbf{C}_B^{-1} (\mathbf{A}_{\mathcal{I}} \mathbf{H} + \mathbf{Z}_p)) \mathbf{h}_i^H \mathbf{h}_i ((\mathbf{H}^H \mathbf{A}_{\mathcal{I}}^H + \mathbf{Z}_p^H) \mathbf{C}_B^{-1} \mathbf{A}_{\mathcal{I}_k})\right] \\
&= \left|\mathbf{h}_i \mathbf{H}^H \mathbf{A}_{\mathcal{I}}^H \mathbf{C}_B^{-1} \mathbf{A}_{\mathcal{I}_k}\right|^2 + N_0 \|\mathbf{h}_i\|^2 \left\|\mathbf{C}_B^{-1} \mathbf{A}_{\mathcal{I}_k}\right\|^2,
\end{aligned} \tag{4.14}$$

$$\begin{aligned}
c_2^k &= \mathbb{E}\left[(\mathbf{A}_{\mathcal{I}_k}^H \mathbf{C}_B^{-1} (\mathbf{A}_{\mathcal{I}} \mathbf{H} + \mathbf{Z}_p))((\mathbf{H}^H \mathbf{A}_{\mathcal{I}}^H + \mathbf{Z}_p^H) \mathbf{C}_B^{-1} \mathbf{A}_{\mathcal{I}_k})\right] \\
&= \left\|\mathbf{H}^H \mathbf{A}_{\mathcal{I}}^H \mathbf{C}_B^{-1} \mathbf{A}_{\mathcal{I}_k}\right\|^2 + M N_0 \left\|\mathbf{C}_B^{-1} \mathbf{A}_{\mathcal{I}_k}\right\|^2,
\end{aligned} \tag{4.15}$$

where $\mathbf{C}_B = \mathbf{A}_{\mathcal{I}} \mathbf{A}_{\mathcal{I}}^H + N_0 \mathbf{I}_{n_p}$. This concludes the proof. ■

Note that the SINR approximation in (4.11) depends on the actual channel vectors, however, they are replaced with their estimates in the LLR calculation since they are not available at the receiver. We then feed the calculated LLR values to a single-user polar decoder, and if the decoded sequence satisfies the CRC check, it is assumed as successfully decoded and is added to the output list $\hat{\mathcal{D}}$. Once this step is completed, we know the successfully decoded messages and the corresponding pilot signatures as well as the channel vector estimates. Then, we can re-encode and modulate the successfully decoded messages to obtain the actual QPSK symbols unless the decoded signal is a false alarm, and construct the estimate of the transmitted signal of the k-th user in the output list as follows:

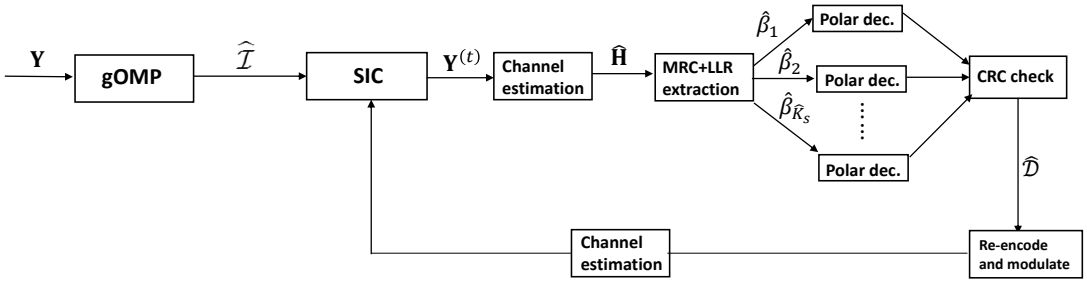

Figure 4.2: Decoding process of the proposed scheme (SNOP-URA) in a slot.

$$\hat{\mathbf{x}}_k = \begin{bmatrix} \mathbf{A}_{\hat{\mathcal{I}}_k}^T & \mathbf{s}_k'^T \end{bmatrix}^T, \tag{4.16}$$

where $\mathbf{A}_{\hat{\mathcal{I}}_k}$ is the estimated pilot signature, and $\mathbf{s}'_k$ is the re-constructed transmitted symbol for the k-th user.

Ignoring the interference of the non-decoded users, the channel vectors corresponding to the successfully decoded messages using an LMMSE filter can be re-estimated as

$$\hat{\mathbf{H}}_{\text{SIC}} = (\hat{\mathbf{X}}_{\hat{\mathcal{D}}}^H \hat{\mathbf{X}}_{\hat{\mathcal{D}}} + N_0 \mathbf{I}_{|\hat{\mathcal{D}}|})^{-1} \hat{\mathbf{X}}_{\hat{\mathcal{D}}}^H \mathbf{Y}^{(j)}, \tag{4.17}$$

where $\hat{\mathbf{X}}_{\hat{\mathcal{D}}}$ is the matrix of the re-constructed transmitted signals for the successfully decoded user messages and $\mathbf{Y}^{(j)}$ is the residual received signal at the j-th iteration. Using these re-estimated channel vectors, we perform SIC as

$$\mathbf{Y}^{(j+1)} = \mathbf{Y}^{(j)} - \hat{\mathbf{X}}_{\hat{\mathcal{D}}} \hat{\mathbf{H}}_{\text{SIC}}. \tag{4.18}$$

At this point, one full iteration is complete. To proceed, we pass the residual signals to the decoder and start another iteration, and continue this process until there is no improvement between two consecutive iterations, i.e., when no message satisfies the CRC check in the current iteration, or the maximum number of iterations is reached. An illustration of the decoding process in a slot is depicted in Figure 4.2, and a pseudo-code for the receiver operation is provided in Algorithm 1, where $n_{\max}$ is the maximum number of decoding iterations.

Algorithm 2 Receiver operation of the proposed scheme (SNOP-URA)

Input: $\mathbf{Y}_p$, $\mathbf{Y}_d$, $\mathbf{A}$, δ, Δ, $n_{\max}$, n_{OMP}
gOMP activity detection algorithm:
for $t = 1, 2, \ldots n_{\text{OMP}}$ **do**
    $\mathbf{R}^{(t)} = \mathbf{A}^H \mathbf{Y}_p^{(t)}$.
    Calculate the Euclidean norm of each row of $\mathbf{R}^{(t)}$ to get $\mathbf{C}^{(t)}$, take the $i_{\text{OMP}} = \left\lceil (\frac{K_a}{V} + \Delta)/n_{\text{OMP}} \right\rceil$ maximum elements of $\mathbf{C}^{(t)}$ and add them to the set $\hat{\mathcal{I}}$.
    If the termination condition is satisfied, go to Step 10.
    Otherwise, $\mathbf{Y}_p^{(t+1)} = \mathbf{Y}_p - \mathbf{A}_{\hat{\mathcal{I}}^{(t)}} (\mathbf{A}_{\hat{\mathcal{I}}^{(t)}}^H \mathbf{A}_{\hat{\mathcal{I}}^{(t)}} + N_0 \mathbf{I}_{\left|\hat{\mathcal{I}}^{(t)}\right|})^{-1} \mathbf{A}_{\hat{\mathcal{I}}^{(t)}}^H \mathbf{Y}_p$.
end for
Channel estimation and decoding:
for $j = 1, 2, \ldots n_{\max}$ **do**
    Initialize $\hat{\mathcal{D}} = \emptyset$, and set $\hat{K}_s = \left|\hat{\mathcal{I}}\right|$.
    Estimate the channel matrix $\hat{\mathbf{H}}$ using (4.9).
    for $i = 1, 2, \ldots, \hat{K}_s$ **do**
        MRC: Estimate $\hat{\mathbf{s}}_i$ by (4.10).
        LLR extraction + polar decoding $\rightarrow \hat{\mathbf{m}}_i$
        If CRC check is satisfied, $\hat{\mathcal{D}} = \hat{\mathcal{D}} \cup i$
    end for
    if $\left|\hat{\mathcal{D}}\right| = 0$ **then**
        Terminate the algorithm.
    end if
    SIC:
    for k in $\hat{\mathcal{D}}$ **do**
        Re-encode and modulate $\hat{\mathbf{m}}_k \rightarrow \mathbf{s}'_k$
    end for
    Re-estimation and interference cancellation by (4.17) and (4.18)
end for
Output: List of messages

4.3 Performance Analysis

4.3.1 Error Probability

In this subsection, we present an approximate expression for the error probability of the proposed scheme in the finite blocklength (FBL) regime. We consider the operation in a single slot since the slot operations are identical and independent. In order to simplify the analysis, we assume that the active pilot sequences are detected without error, i.e., $\hat{\mathcal{I}} = \mathcal{I}$. This is justified based on our extensive simulations since the error probability of AD is much less than the target PUPE for the considered set of parameters. Moreover, we calculate the approximate error probability due to the collisions in Sec. 4.3.2 and subtract it from the target PUPE in order to characterize the performance of the proposed scheme more precisely.

In order to characterize the error probability analytically, we first calculate the effective signal-to-interference-plus-noise ratio (SINR) for each user at the output of MRC as the users are separated at this step. Then, symbol estimates of the users become outputs of a virtual single-user channel with an SNR equal to their effective SINR by treating the multiuser interference as Gaussian due to the high number of interfering terms. Therefore, the finite blocklength analysis based on normal approximations can be applied.

Assuming that the user channel vectors are perfectly known, the symbol estimate of the k-th user at the output of MRC can be written as

$$\begin{aligned}\hat{\mathbf{s}}_k &= \mathbf{s}_k \mathbf{h}_k \mathbf{h}_k^H + \sum_{i=1, i\neq k}^{K_s} \mathbf{s}_i \mathbf{h}_i \mathbf{h}_k^H + \mathbf{Z}_d \mathbf{h}_k^H \\ &= \|\mathbf{h}_k\|^2 \mathbf{s}_k + \sum_{i=1, i\neq k}^{K_s} \mathbf{s}_i \mathbf{h}_i \mathbf{h}_k^H + \mathbf{Z}_d \mathbf{h}_k^H, \end{aligned} \tag{4.19}$$

where the first term is the desired signal, the second term is the multiuser interference, and the last one is the noise term. Then, the SINR of the k-th user at

the output of maximal ratio combiner, i.e., the input of the polar decoder, in the first iteration γ_k can be calculated as

$$\gamma_k = \frac{P_S}{P_I + P_N}, \tag{4.20}$$

where P_S is the power of the desired signal, P_I is the power of the multiuser interference, and P_N is the power of the Gaussian noise after MRC. Assuming that for $n_d \to \infty$, the transmitted polar codewords consist of uncorrelated and equally likely QPSK symbols, i.e., $\mathbb{E}\left[\mathbf{s}_k \mathbf{s}_k^H\right] = P_d \mathbf{I}_{n_d}\ \forall k \in \mathcal{K}_s$, and that the polar codewords of the two different users are uncorrelated under the assumption that the user messages are independent, each term in (4.20) can be calculated as

$$P_S = \frac{1}{n_d}\mathbb{E}\left[(\mathbf{s}_k \mathbf{h}_k \mathbf{h}_k^H)^H \mathbf{s}_k \mathbf{h}_k \mathbf{h}_k^H\right] = P_d \left\|\mathbf{h}_k\right\|^4, \tag{4.21}$$

$$\begin{aligned} P_I &= \frac{1}{n_d} \sum_{i=1, i\neq k}^{K_s} \sum_{j=1, j\neq k}^{K_s} \mathbb{E}\left[\left(\mathbf{s}_i \mathbf{h}_i \mathbf{h}_k^H\right)^H \mathbf{s}_j \mathbf{h}_j \mathbf{h}_k^H\right] \\ &= P_d \sum_{i=1, i\neq k}^{K_s} \left|\mathbf{h}_i \mathbf{h}_k^H\right|^2, \end{aligned} \tag{4.22}$$

$$P_N = \frac{1}{n_d}\mathbb{E}\left[\mathbf{h}_k \mathbf{Z}_d^H \mathbf{Z}_d \mathbf{h}_k^H\right] = N_0 \left\|\mathbf{h}_k\right\|^2, \tag{4.23}$$

where the expectation is taken over the transmitted symbols and the noise. Strictly speaking, the assumption of uncorrelated QPSK symbols is not accurate for coded systems, however, it is useful to obtain a good SINR approximation.

For a channel code with blocklength n_c, the approximate achievable rate for an error probability of p_e on an AWGN channel is given by the normal approximation [27]

$$R \approx C - \sqrt{\frac{V_{dis}}{n_c}} Q^{-1}(p_e), \tag{4.24}$$

where $Q(.)$ is the standard Q-function, R is the code rate, C is the channel capacity, and V_{dis} is the channel dispersion. It is also known that a quasi-static

channel is conditionally ergodic given the channel gain [75]. Therefore, for a given set of channel coefficients ($\mathbf{h}_i$'s), the SINR can be treated as known and the channel capacity and dispersion can be calculated as

$$C = \frac{1}{2}\log_2(1+\gamma_k), \;\; V_{dis} = \frac{\gamma_k}{2}\frac{\gamma_k+2}{(\gamma_k+1)^2}\log_2^2(e). \tag{4.25}$$

Using channel capacity and dispersion in (4.25), the block error probability conditioned on the channel gains for a given code rate can be calculated as

$$p_e \approx Q\left(\frac{C-R}{\sqrt{V_{dis}/n_c}}\right). \tag{4.26}$$

Therefore, to determine the average error probability across all the users, we average this quantity over the channel realizations and obtain

$$p_{e,avg} \approx \mathbb{E}_H\left[Q\left(\frac{C-R}{\sqrt{V_{dis}/n_c}}\right)\right], \tag{4.27}$$

where the expectation is taken over the randomness of the channel coefficients. We can evaluate this expectation through Monte Carlo simulations to obtain the required input power to obtain a target PUPE.

4.3.2 Collisions

In this subsection, a brief analysis of the effects of the pilot collisions on the system performance is provided. In URA, it is impossible to assign fixed pilot sequences to the users due to their massive numbers. Therefore, when the pilot bits of multiple users are the same, they pick the same pilot sequence, and a collision happens. The operations in the slots are identical yet independent, hence, we can consider the collision probability in one slot. Note that this is analogous to the generalization of the birthday problem where the number of days is the number

of pilot sequences and the number of users is the number of people. Therefore, we can calculate the probability of a collision as follows

$$P_{co} = 1 - \prod_{i=1}^{K_s-1} \left(1 - \frac{i}{N}\right). \tag{4.28}$$

In order to obtain the effect of the pilot collisions on the system performance we need to calculate the number of users in a collision. For this purpose, first of all, the probability of a specific user being in a collision can be calculated as follows

$$P_s = 1 - \left(\frac{N-1}{N}\right)^{K_s-1}. \tag{4.29}$$

Then, the expected number of users in a collision for a given K_s can be obtained as

$$\mathbb{E}[S|K_s] = K_s \left(1 - \left(\frac{N-1}{N}\right)^{K_s-1}\right). \tag{4.30}$$

Note that K_s is a random variable with a Binomial distribution as it models the number of successes in K_a independent experiments with probability $\frac{1}{V}$. Using this fact, the expected number of users in a collision is given by

$$\mathbb{E}[S] = \sum_{k_s=1}^{K_a} k_s \left(1 - \left(\frac{N-1}{N}\right)^{k_s-1}\right) \binom{K_a}{k_s} \left(\frac{1}{V}\right)^{k_s} \left(\frac{V-1}{V}\right)^{K_a-k_s}. \tag{4.31}$$

We know that the first-order approximation $(1+x)^n \approx 1+nx$ is tight when $x \ll 1$. In addition, in the massive random access, since K_a is large, it is reasonable to assume that K_s has a normal distribution with mean $\frac{K_a}{V}$ and variance $\frac{K_a}{V}\frac{V-1}{V}$, i.e., $K_s \sim \mathcal{N}\left(\frac{K_a}{V}, \frac{K_a(V-1)}{V^2}\right)$ when $\frac{1}{V}$ is not too small [77]. Therefore, using the first order approximation for (4.30), the expected number of users in a collision for a given K_s can be approximated as

$$\mathbb{E}[S|K_s] \approx \frac{K_s(K_s-1)}{N}. \tag{4.32}$$

Then, the expected number of users in collision can be approximated as

$$\mathbb{E}[S] \approx \frac{\mathbb{E}[K_s^2]-\mathbb{E}[K_s]}{N} = \frac{K_a^2-K_a}{V^2N}. \tag{4.33}$$

As an example, for $K_a = 500$, $V = 4$ and $N = 2^{13}$ (13 pilot bits), the expected number of users involved in a collision is about 1.90, which leads to an error probability of 0.0152 if none of the users involved in the collision can be recovered. On the other hand, if a suitably designed polar code of rate less than $\frac{1}{2}$ is used over a Gaussian MAC, then both of the colliding users may be recovered [38]. For our case, if the channel vector of one of the two users has a considerably higher Euclidean norm, that user will be recovered with a high probability. So, if we assume that one of the colliding users is non-decodable, the overall error probability becomes 0.0076.

Note that (4.33) is a decreasing function of V, hence the collision probability can be decreased by increasing V for a fixed number of pilot sequences; however, it also decreases the packet lengths which can degrade the system performance. There is also a trade-off between the number of pilot sequences N and complexity, i.e., increasing the number of pilot sequences decreases the collision probability while increasing the system complexity.

4.3.3 Complexity

In this subsection, we provide a computational complexity analysis of the proposed scheme considering the average number of multiplications in each slot. The total complexity of the scheme can be calculated by summing this quantity over V slots. The complexity of each iteration of gOMP is dictated by the correlation step in (4.6), with a complexity of $\mathcal{O}(Nn_pM)$, where $\mathcal{O}(.)$ is used for the standard big-O notation. This can be further reduced to $\mathcal{O}(MN\log N)$ by employing a sub-sampled DFT matrix as the pilot codebook and converting matrix

multiplications in the gOMP algorithm to FFT operations [76]. The complexity of LMMSE channel estimation is dominated by the complexity of the matrix multiplication, which is $\mathcal{O}((\hat{K}_s^{(j)})^2 n_p)$, where $\hat{K}_s^{(j)}$ is the number of remaining users to be decoded in the j-th decoding iteration. The complexity of the matrix multiplication in the MRC step is $\mathcal{O}(\hat{K}_s^{(j)} M n_d)$, and single-user decoding has a complexity of $\mathcal{O}(\hat{K}_s^{(j)} n_L n_d \log n_d)$. The complexity of the re-estimation of the channel vectors for SIC is $\mathcal{O}((\hat{K}_{s,d}^{(j)})^2 L)$, where $\hat{K}_{s,d}^{(j)}$ is the number of successfully decoded users in the j-th decoding iteration. Therefore, by normalizing over the decoding iterations, the per-iteration complexity of the proposed scheme becomes $\mathcal{O}\left(\frac{n_{\text{OMP}}}{n_{\text{dec}}} MN \log N + (\hat{K}_s^{(j)})^2 n_p + (\hat{K}_{s,d}^{(j)})^2 L\right)$. This quantity is either dominated by the complexity of the gOMP step, the LMMSE channel estimation process or the re-estimation of the channel vectors for SIC, depending on the system parameters, where n_{dec} is the number of decoding iterations. However, based on our extensive simulations, in the first decoding iteration, $\hat{K}_s^{(j)} \approx \frac{K_a}{V}$ and $\hat{K}_{s,d}^{(j)} \ll \frac{K_a}{V}$, and in the last decoding iteration, $\hat{K}_s^{(j)} \approx 0$. Assuming that the complexity of the intermediate decoding iterations is in the same order as the first and last iterations, the complexity of SNOP-URA normalized over the decoding iterations can be bounded as follows

$$\mathcal{C}_{SNOP-URA} \leq \mathcal{O}\left(\frac{n_{\text{OMP}}}{n_{\text{dec}}} MN \log N + \max\left(\left(\frac{K_a}{V}\right)^2 n_p, \max_j (\hat{K}_{s,d}^{(j)})^2 L\right)\right). \tag{4.34}$$

For comparison, the complexity of the scheme in [62] is dominated by the energy detector which has a complexity of $\mathcal{O}(N(n_d L_{ss} + n_p)M)$ at each iteration, where L_{ss} is the length of the employed spreading sequence. This complexity is much higher than the complexity of the proposed scheme in this paper for practical system parameters.

4.4 Numerical Results

We now present a performance evaluation of the proposed URA scheme (SNOP-URA). We choose $n = 3200$ for a fair comparison with the available results in the literature, $B = 100$, $J = 13$, $M = 50$, $\epsilon = 0.05$, and assume i.i.d. Rayleigh fading channels unless otherwise specified. Our extensive simulations as well as the approximate analytical results demonstrate that setting the number of slots to 4 outperforms the other choices for a wide range of active user loads (as will be clarified later). Therefore, we take the number of slots as 4 which corresponds to a slot size of 800 and we take $n_p = 288$ and $n_d = 512$. We employ 5G polar codes with a length of 1024, set the CRC length to 16, and employ successive cancellation list decoding (SCLD) with a list size of 128. Furthermore, we use a randomly sub-sampled DFT matrix of size $N = 2^J$ for the pilot codebook.

4.4.1 Activity Detection

We first examine the effect of the number of gOMP iterations on the performance of AD for $K_a = 300$ and 700. We also consider the case of standard OMP, where the number of iterations is equal to the estimated number of users in the slot since only one index is added to the output list at each iteration. This is denoted by $\hat{K}_s$ *iter* in Figure 4.3. Figure 4.3 shows that implementing gOMP with 4 iterations offers competitive AD performance with the standard OMP while having a lower complexity. With this observation, we take the number of gOMP iterations as 4 in the rest of the simulations.

We next evaluate the performance of gOMP with 4 iterations for different Δ values with respect to the average active user load along with the case that the number of active users in each slot is exactly known. The results in Figure 4.4 show that taking Δ as approximately 30% of the estimated average active user load or higher leads to a good AD performance. Also, it performs even better than knowing the exact number of users in the slots since some users with weak channel vectors may not be in the list of detected users; however, they can be

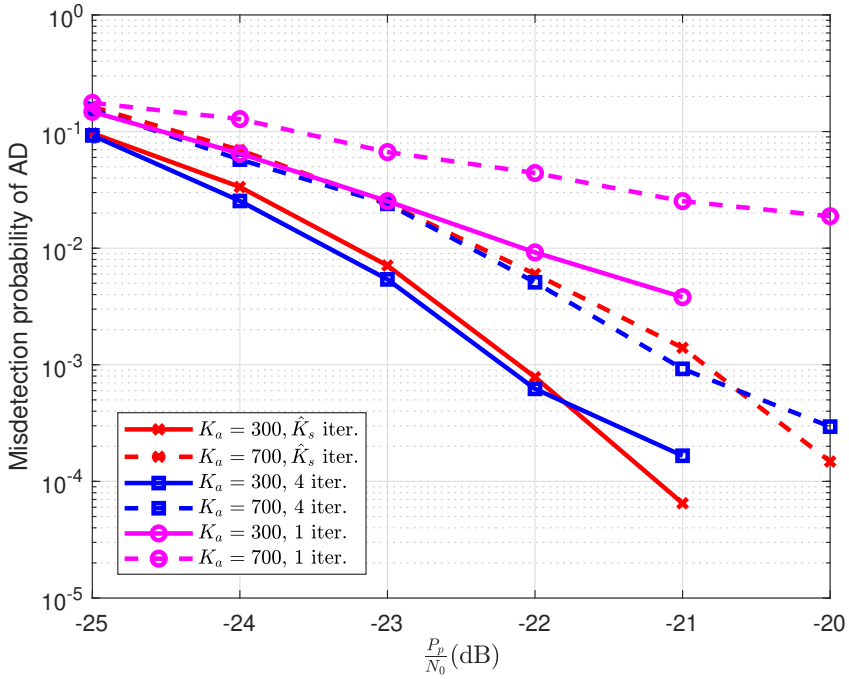

Figure 4.3: Misdetection probability of gOMP for different number of iterations for $K_a = 300, 700$.

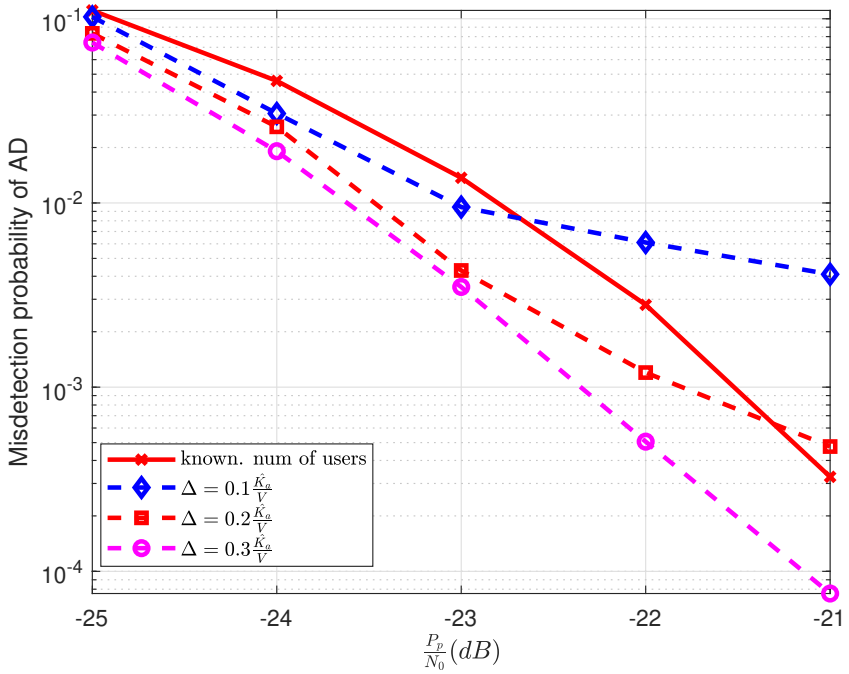

Figure 4.4: Misdetection probability of activity detection with different Δ values for $K_a = 300$.

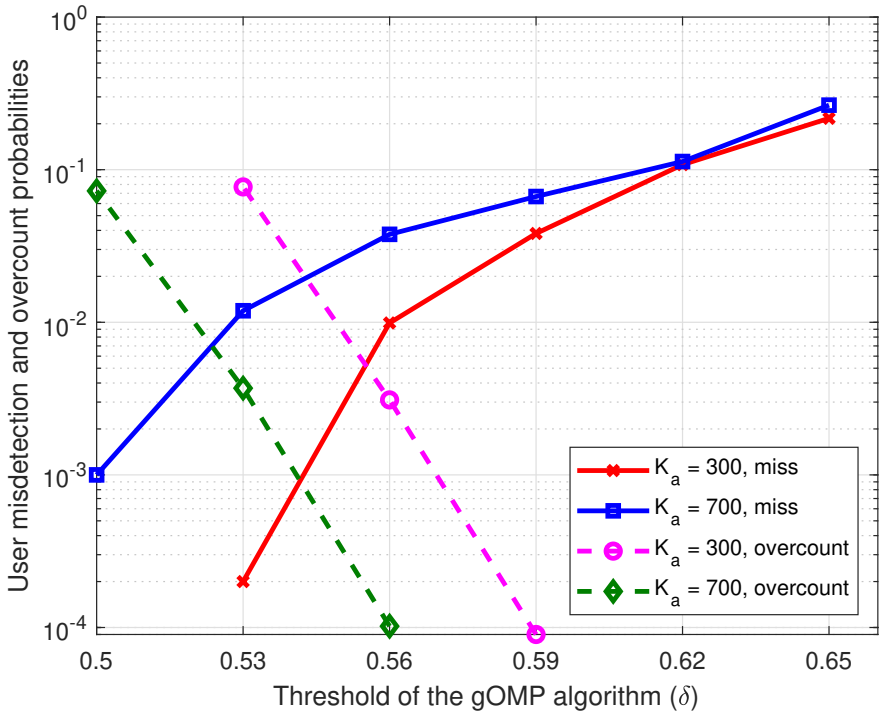

Figure 4.5: Misdetection and overcount probabilities of gOMP with respect to δ for $K_a = 300, 700$.

detected when the indices are selected with some overhead on the average active user load. In the rest of the simulations, we set $\Delta = \lfloor 0.3\frac{\hat{K}_a}{V} \rceil$.

We investigate the effect of the choice of threshold δ in the gOMP algorithm on the system performance by calculating the misdetection (undercount) and overcount probabilities for $K_a = 300, 700$, and $\frac{P_p}{N_0} = -22$ dB in Figure 4.5. The results illustrate that taking the threshold as 0.53 for $K_a = 300$ and 0.5 for $K_a = 700$ leads to a misdetection probability of 10^{-3} or less while the overcount is less than 10 % of the estimated active user load. Considering that these active user loads are representative for low and high multiuser interference regimes, we take $\delta = 0.53$ if $K_a \leq 400$, and $\delta = 0.5$ otherwise, for the rest of the simulations.

4.4.2 Overall Performance

We assess the energy efficiency of SNOP-URA by calculating the minimum E_b/N_0 for an error probability of 0.05 in Figure 4.6, which demonstrates that the proposed SNOP-URA outperforms the one referred to as *multi stage. orth. pilot*

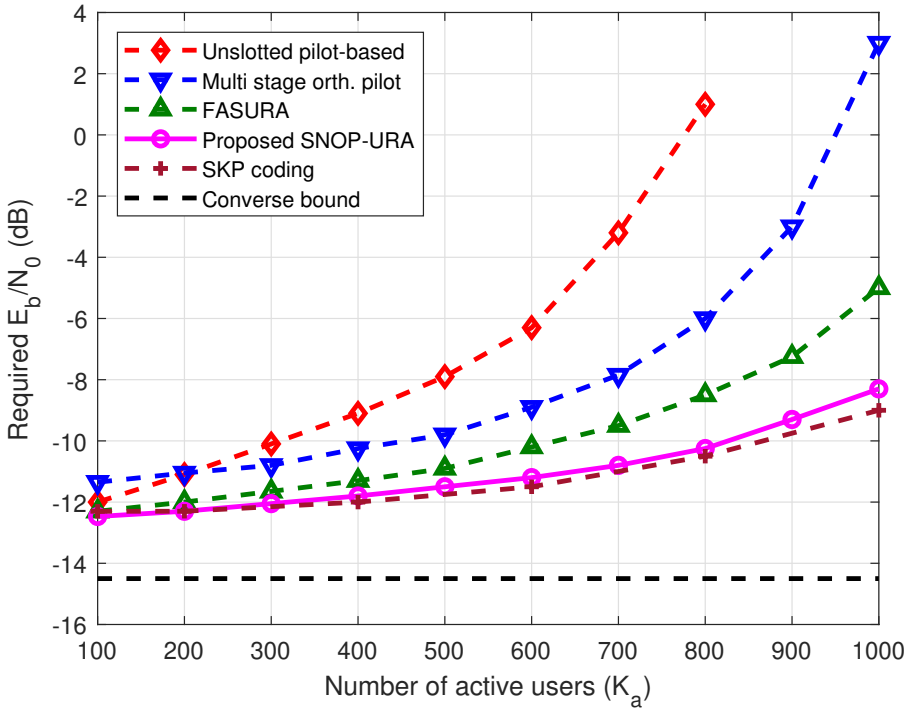

Figure 4.6: Required E_b/N_0 versus number of active users of the proposed scheme and the schemes in the literature for $M = 50$ and $\epsilon = 0.05$.

[60] by up to 11 dB for a user load of $K_a \leq 1000$. The proposed scheme is also superior to the pilot-based non-slotted scheme developed in [59] by up to 11 dB for $K_a \leq 800$, and it outperforms FASURA by up to 3 dB for $K_a \leq 1000$. Note that we take $J = 14$ for $K_a > 800$. For further comparison, we also calculate the complexity orders of the proposed scheme and FASURA for different active user loads using the parameters in Figure 4.6 with $M = 50$. As summarized in Table 4.1, the SNOP-URA provides a substantial complexity advantage.

Very recently, a scheme that adopts sparse Kronecker-product (SKP) coding in [42] for the MIMO scenario is proposed in [63], for which the newly proposed scheme has a better performance by 0.2 dB for $K_a = 100$, a similar performance for $K_a = 200$, and it is superior to the proposed SNOP-URA by at most 0.3 dB for $200 < K_a \leq 800$ as shown in Figure 4.6. However, the scheme in [63] has a complexity of $\mathcal{O}(MK_an)$, which is dominated by the outer matrix factorization in that work. This complexity is higher than that of the newly proposed SNOP-URA scheme. For instance, for $K_a = 600$, the number of the multiplications required for the solution in [63] is 96×10^6, while it is at most 34.6×10^6 for the SNOP-URA, which amounts to a saving of more than 60 %. Furthermore, [63] considers only the misdetections as an error metric, and there is no checking mechanism for false alarms, namely, the corresponding performance may be slightly worsened

Table 4.1: Comparison of the complexity orders ($\times 10^6$ multiplications)

URA scheme	$K_a = 300$	$K_a = 600$	$K_a = 900$
Prop. SNOP-URA	21.68	34.6	74.06
FASURA	10486	10486	10486

when considering both the misdetections and false alarms to calculate the PUPE.

In order to compare the performance of the proposed SNOP-URA scheme with the ultimate limits, we consider the converse bound derived in a recent work [67, Theorem 2]. We observe that our proposed scheme performs within 4.5 dB of the converse bound for $K_a \leq 800$ as shown in Figure 4.6, and the performance gap is less than 3 dB for $K_a \leq 500$. In addition, the performance gap of the proposed scheme and the achievability bound in [67, Theorem 1] is less than 2 dB for $K_a \leq 1000$. These aspects demonstrate that the proposed scheme not only has a superior or competitive performance with the existing works but also performs relatively close to the information-theoretic (achievability and converse) bounds.

The performance advantage of the proposed SNOP-URA with respect to FASURA demonstrates that slotting the transmission frame, combined with an optimization over the number of slots can outperform spreading over the transmission frame. In addition, we utilize an iterative gOMP algorithm for AD while FASURA employs a correlation-based energy detector in a non-iterative fashion to detect the spreading sequences. Moreover, we have more flexibility on the slot length selection compared to the scheme in [60], since we utilize non-orthogonal pilots rather than multiple stages of orthogonal ones. In other words, while offering some advantages, the use of orthogonal pilots brings additional constraints on the pilot lengths, hindering the optimization of the number of slots.

Let us also evaluate the performance of the proposed scheme for the short blocklength regime by comparing it with the existing schemes operating in this regime, i.e., for $L \approx 200$ and $M = 100$. The results in Figure 4.7 demonstrate that the proposed scheme outperforms the schemes in [60], referred to as *Multi. stage. orth. pilot* for $K_a \leq 600$, and [55], referred to as *Non-Bayesian AD*

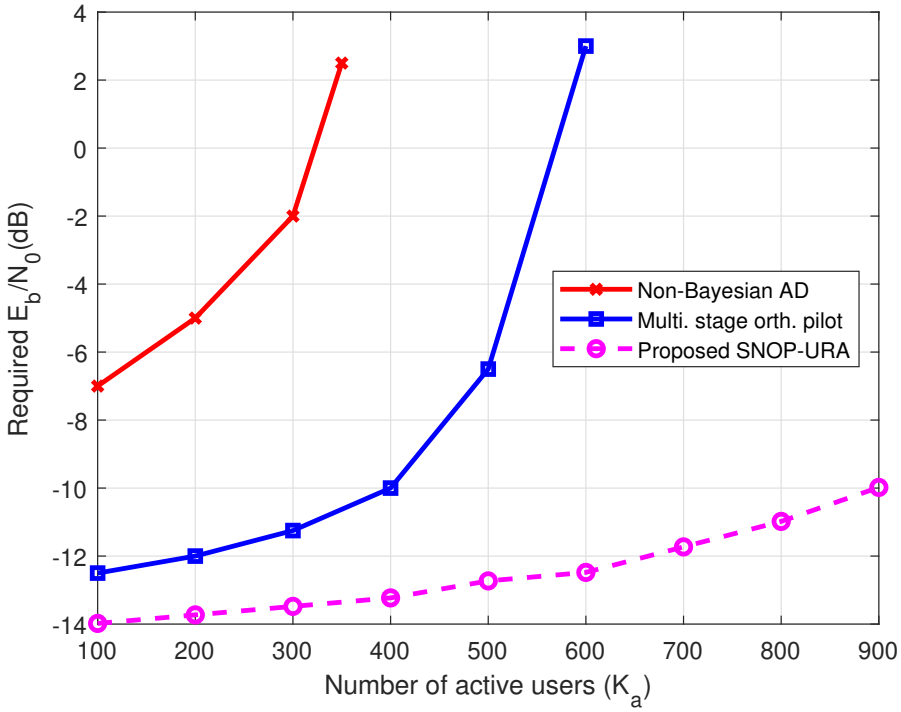

Figure 4.7: Required E_b/N_0 versus the number of active users for the short blocklength regime ($L \approx 200$) for $\epsilon = 0.05$.

for $K_a \leq 350$, by up to 15 dB. In addition, the proposed SNOP-URA solution supports an active user load of up to 900 in the short blocklength regime, while the schemes in [55] and [60] can support up to 350 and 600 users, respectively.

We also investigate the effect of the correlation among the antennas on the system performance. For this purpose, we multiply the received signal by a correlation matrix $\mathbf{C}_r \in \mathbf{C}^{M \times M}$ with elements $C_r(i,j) = \alpha^{|i-j|}$ where $\alpha \leq 1$ is the correlation constant. We evaluate the performance of our proposed scheme for $\alpha = 0.1, 0.25, 0.4$ and the parameters in Figure 4.6 with the assumption that the receiver is not aware of the correlation. The results in Figure 4.8 illustrate that the correlation among the antennas degrades the system performance by up to 0.5 dB for $\alpha = 0.1$, 2.5 dB for $\alpha = 0.25$ and 10 dB for $\alpha = 0.4$ for $K_a \leq 700$, on the other hand, the solution is still effective for a wide range of active user loads.

Finally, we evaluate the performance of the proposed scheme for the case of multiple but not massive number of antennas, namely, for $M = 8$. We compare the performance of SNOP-URA to those of SKP coding [63] and FASURA [62] for $M = 8$ and $\epsilon = 0.1$ in Figure 4.9, which demonstrates that the proposed

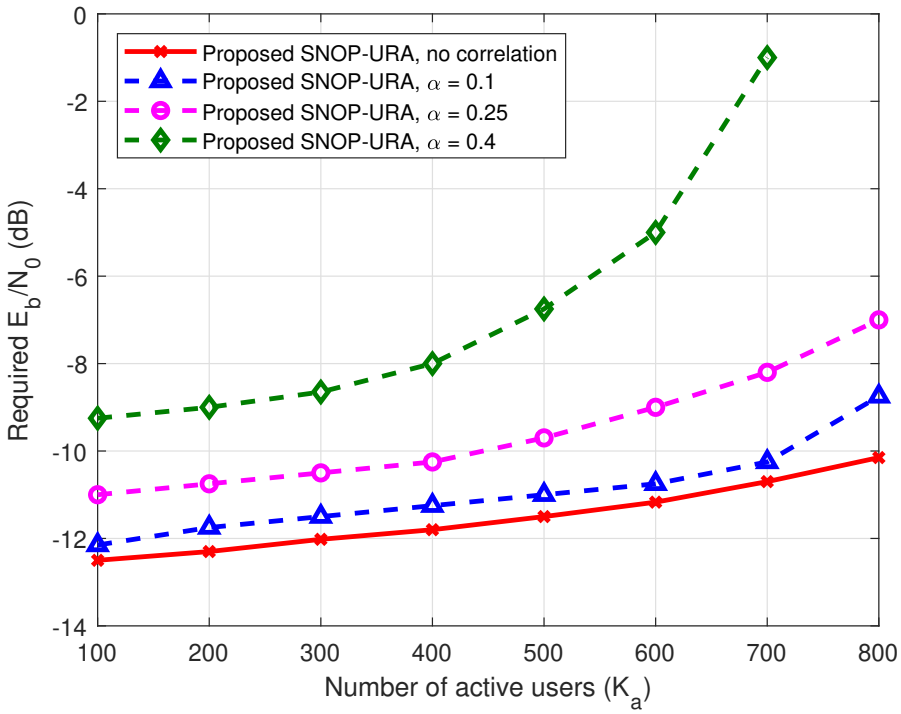

Figure 4.8: Required E_b/N_0 versus number of active users for different correlation levels among the antennas for $\epsilon = 0.05$.

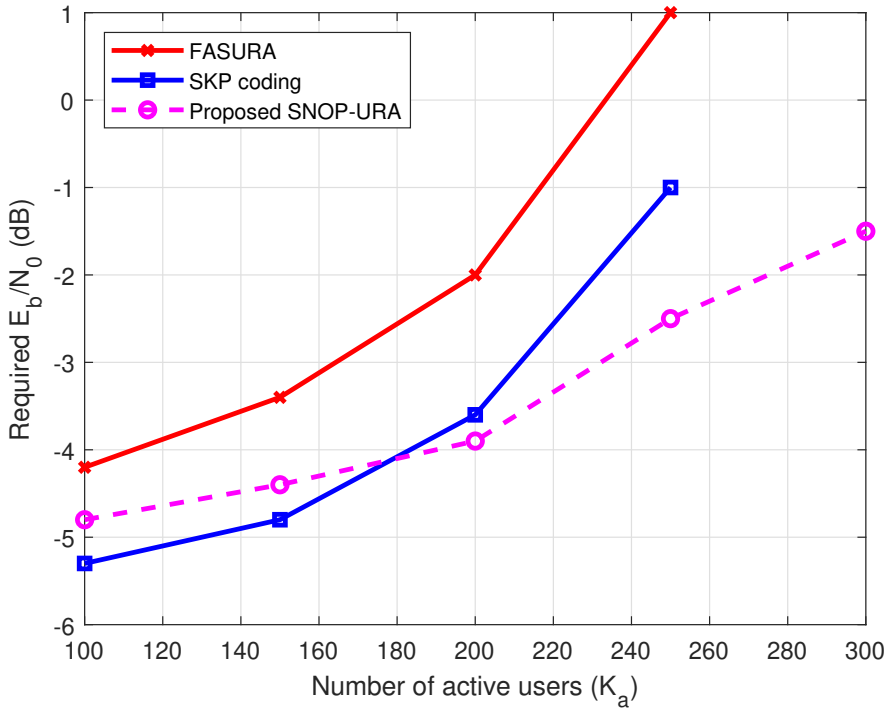

Figure 4.9: Required E_b/N_0 versus the number of active users of the proposed scheme and the state-of-the-art schemes in the literature for $M = 8$ and $\epsilon = 0.1$.

SNOP-URA outperforms FASURA for all active user loads by up to 3.5 dB and SKP coding for $K_a \geq 200$ by up to 1.5 dB.

4.4.3 Verification of Analytical Approximations

In this section, we verify the normal approximation based analytical results obtained in Sec. 4.3. For this purpose, we first obtain the required input power for a probability of error $p_e = \epsilon - p_{e,c}$ where $p_{e,c}$ is the approximate error probability resulting from the collisions. Then, we calculate the required $\frac{E_b}{N_0}$ using the obtained results in Sec. 4.3. We also calculate the $\frac{E_b}{N_0}$ using the SINR approximation in (4.11) which is referred as *analytical, est. CSI* to compare it with the actual system performance taking the channel estimation errors into account. Note that this SINR approximation is not rigorously related to FBL analysis and we only use it as an approximation. For the simulation results, we do not take into account the SIC step since we are interested in computing the SINR in the first iteration. Figure 4.10 illustrates that the analysis accurately characterizes the performance of the proposed scheme as the performance gap is less than 0.85 dB for the entire range of active user loads when $M = 100$, less than 1.5 dB for $K_a \leq 600$ for the estimated CSI, and less than 1.75 dB for $K_a \leq 700$ when $M = 50$. Also, one can observe that the analytical results become tighter with increasing the number of receive antennas.

We also investigate the effect of the unequal distribution of the total power to the pilot and data parts by sweeping over the ratio between P_d and P_p in (4.11) and calculating the SINR for different numbers of slots and pilot lengths for $n_d = 512$ and $M = 50$. This setting is selected since it is a good representative for the actual system performance and performing the actual system simulations for each set of parameter is difficult. Figure 4.11 demonstrates that for the considered set of parameters of our system, the data and pilot power can be the same but an unequal power assignment can be beneficial for 5 slots and $n_p = 128$ as the SINR is maximized when $\frac{P_d}{P_p} = -4$ dB. We verify this result in Figure 4.12 by calculating the required $\frac{E_b}{N_0}$ for the same set of parameters. The results show

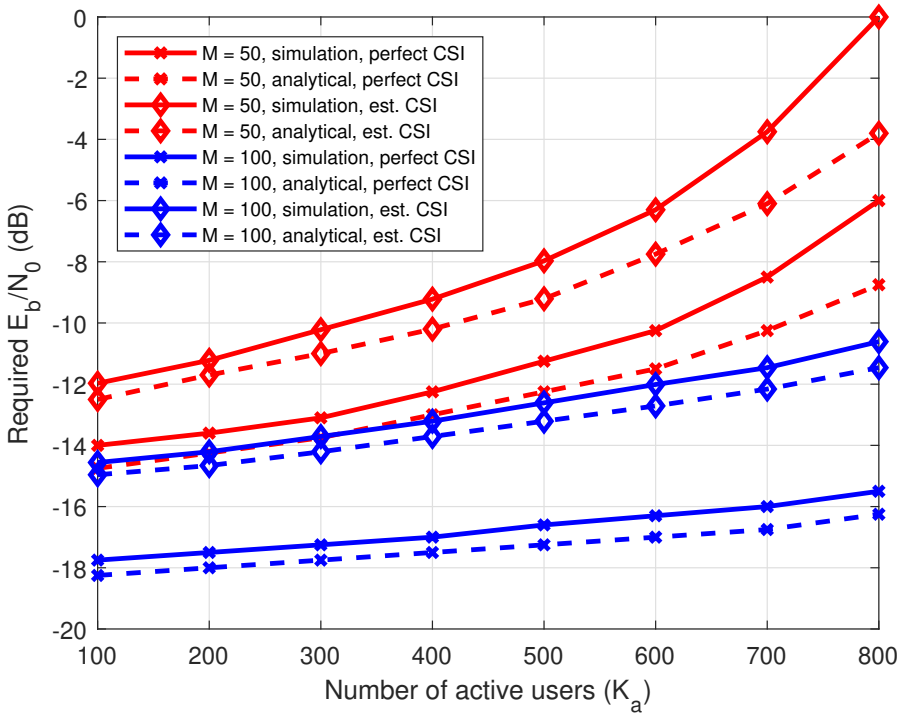

Figure 4.10: Required $\frac{E_b}{N_0}$ versus the number of active users for different number of BS antennas.

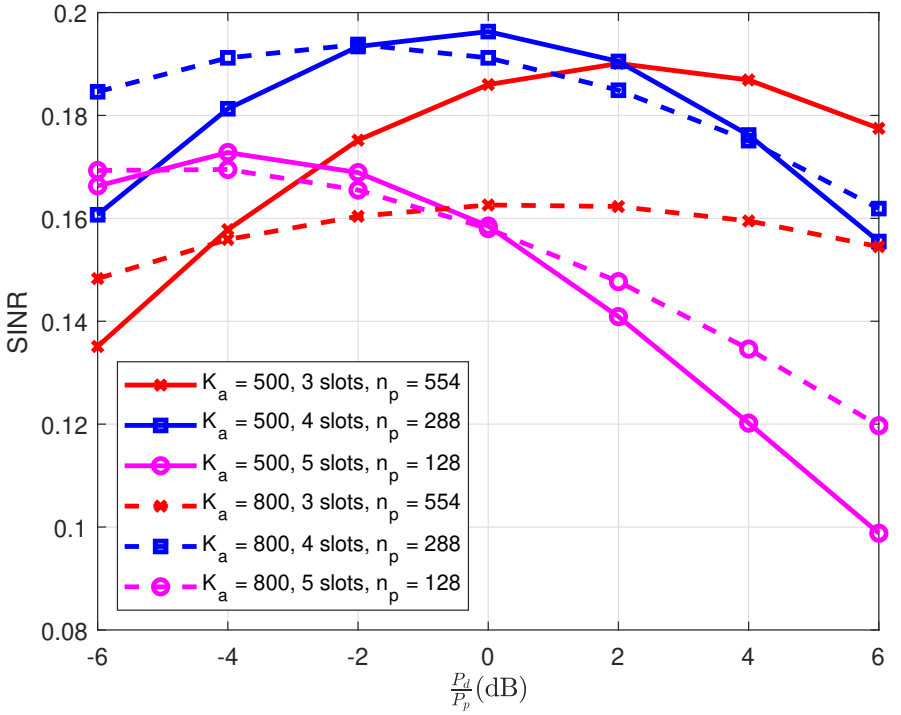

Figure 4.11: SINR values for different sets of parameters with respect to power distribution between pilot and data parts.

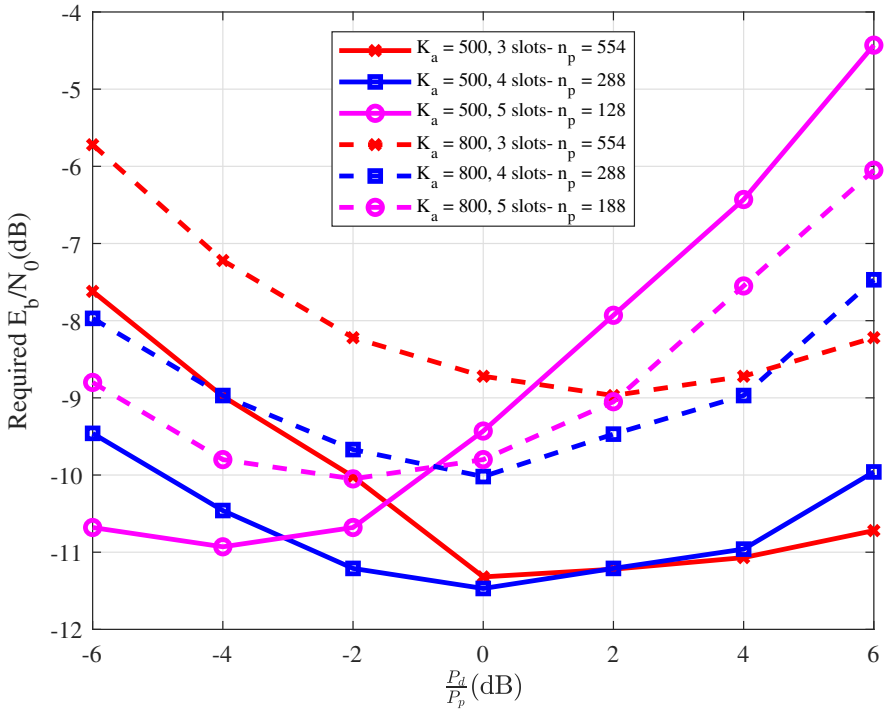

Figure 4.12: Required $\frac{E_b}{N_0}$ with respect to power distribution of data and pilot parts for different pilot lengths.

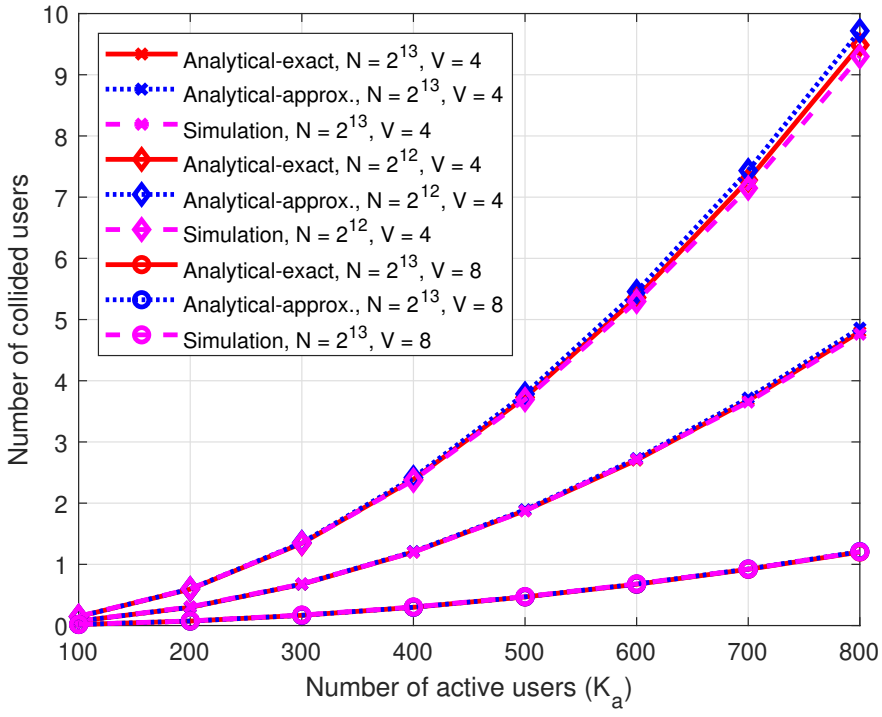

Figure 4.13: Average number of users in a collision with respect to number of active users for different number of pilot sequences and number of slots.

that the minimum required $\frac{E_b}{N_0}$ is obtained when $\frac{P_d}{P_p} = -4$ dB for $n_p = 128$, while it is around $\frac{P_d}{P_p} = 0$ dB for the other two scenarios, i.e., assigning equal pilot and data powers is optimal. Note that we consider a slightly higher transmission frame for $K_a = 800$ and 5 slots by taking $n_p = 188$ since LMMSE estimation works when $n_p \geq K_s$. Moreover, in order to check the accuracy of the collision analysis expressions obtained in (4.31) and (4.33), the average number of users involved in a collision is depicted in Figure 4.13, which clearly illustrates the tightness of the derived approximate expression.

4.5 Chapter Summary

We consider unsourced random access over Rayleigh block fading channels where the BS is equipped with a massive number of antennas. We propose a slotted scheme called SNOP-URA to decrease the multiuser interference, and divide each slot into pilot and data parts. After randomly selecting a slot, each user transmits a pilot sequence from a non-orthogonal set in the pilot part and a polar encoded and modulated signal in the data part. We employ a gOMP based algorithm for activity detection, and an iterative algorithm with LMMSE channel estimation and MRC-based polar decoding of data symbols at the receiver. At the end of each iteration, we re-estimate the channel vectors and use these estimates for SIC. We also provide a performance and complexity analysis, and examine the effect of pilot collisions on the system performance. Our results demonstrate that the proposed SNOP-URA solution offers superior performance to the state-of-the-art, or has a competitive performance with a lower complexity. Note that since the proposed scheme works well in the short blocklength regime as well, it may also be suitable for fast fading scenarios.

Chapter 5

Unsourced Random Access with Hardware Impairments

In practical massive MIMO systems, the antenna elements may consist of inexpensive hardware which are particularly sensitive to impairments like I/Q imbalance, phase noise, and quantization errors, leading to new models with transceiver hardware impairments. These impairments can be considered as an aggregate, called residual hardware impairments (HWIs) [78, 79, 80]. Furthermore, since the potential number of users in a URA system is in the order of millions, the devices are likely to be very inexpensive, and hence HWIs are inevitable on the user side as well. With this motivation, in this chapter, we study URA systems with residual additive and multiplicative HWIs at both the user equipment (UE) and the BS with a massive MIMO receiver. Specifically, we propose a URA solution based on our work in the previous chapter by adopting it to the scenario with HWIs. Namely, we employ a slotted transmission scheme, and divide each slot into pilot and data parts. In the pilot part, each user selects a pilot sequence that is randomly picked from a common non-orthogonal codebook based on part of its message bits, and encodes the remaining bits by a polar code. At the receiver side, we adapt the receiver blocks to make them HWI-aware using the HWI statistics. Our extensive results reveal that the URA schemes designed for the case with no HWIs can still operate in the presence of HWIs albeit with some

performance penalty. On the other hand, the newly proposed HWI-aware decoding scheme increases the energy efficiency while supporting more active users for both cases of additive and multiplicative HWIs.

The rest of the chapter is organized as follows. We describe the system model in Section 5.1, and the proposed coding scheme in Section 5.2. We then present a set of numerical examples in Section 5.3, and conclude the chapter in Section 5.4.

5.1 System Model

We consider a massive random access set-up where K_a out of K_{tot} users ($K_a \ll K_{\text{tot}}$) transmit B information bits to a BS equipped with M antennas through a transmission frame of length-n. We divide the transmission frame into V slots of length L and assume a quasi-static fading scenario, i.e., the channel vectors remain constant over a slot. We consider additive and multiplicative residual HWIs that are modeled as distortion noises at both the BS and the UE.

For the case of additive HWI, the received signal for the s-th slot at the BS can be written as [78]

$$\mathbf{Y}_{s,a} = \sum_{k \in \mathcal{K}_s} (\mathbf{x}(\mathbf{m}_k) + \boldsymbol{\eta}_k^{UE})\mathbf{h}_k + \boldsymbol{\eta}_s^{BS} + \mathbf{Z}_s, \tag{5.1}$$

where $\mathbf{Y}_{s,a} \in \mathbb{C}^{L \times M}$, $\mathcal{K}_s$ is the set of active users transmitting in the s-th slot, $\mathbf{m}_k$ is the message of the k-th user, $\mathbf{x}(\mathbf{m}_k) \in \mathbb{C}^{L \times 1}$ is the transmitted signal, $\mathbf{h}_k \in \mathbb{C}^{1 \times M}$ is the channel vector, $\mathbf{Z}_s$ is the circularly symmetric complex additive white Gaussian noise (AWGN) with independent and identically distributed (i.i.d.) zero-mean and variance N_0 elements, i.e., $\mathbf{Z}_s \sim \mathcal{CN}(\mathbf{0}, N_0\mathbf{I}_L)$. We assume that the elements of $\mathbf{x}(\mathbf{m}_k)$ have an average power of P, and the elements of $\mathbf{h}_k$ are i.i.d. and modelled as $h_{k,i} \sim \mathcal{CN}(0,1)$. $\boldsymbol{\eta}_k^{UE} \in \mathbb{C}^{L \times 1}$ is the additive HWI vector of the k-th user whose elements are $\eta_{k,i}^{UE} \sim \mathcal{CN}(0, (\kappa^{UE})^2 P)$, where κ^{UE}

is the proportionality coefficient of the additive HWI at the user side. Similarly, $\boldsymbol{\eta}_s^{BS} \in \mathbb{C}^{L\times M}$ models the hardware impairment at the BS in the s-th slot whose rows follow $\mathcal{CN}(0, \Upsilon^{BS})$, where $\Upsilon^{BS} = (\kappa^{BS})^2 P \sum_{k\in\mathcal{K}_s} \text{diag}(|h_{k,1}|^2, \ldots, |h_{k,M}|^2)$ with $\mathbf{X}_{(i,:)}$ being the i-th row of $\mathbf{X}$ and κ^{BS} is the proportionality coefficient of the additive HWI at the base station.

The received signal in the s-th slot at the BS with multiplicative HWIs can be written as follows [79]

$$\mathbf{Y}_{s,m} = \sum_{k\in\mathcal{K}_s} ((\mathbf{x}(\mathbf{m}_k) \odot \mathbf{c}_k^{UE})\mathbf{h}_k) \odot \mathbf{C}_s^{BS} + \mathbf{Z}_s, \tag{5.2}$$

where $\mathbf{Y}_{s,m} \in \mathbb{C}^{L\times M}$, $\odot$ denotes elementwise multiplication, $\mathbf{c}_k^{UE} \in \mathbb{C}^{L\times 1}$ and $\mathbf{C}_s^{BS} \in \mathbb{C}^{L\times M}$ are the multiplicative HWI vector of the k-th user and the multiplicative HWI at the BS in the s-th slot, respectively. The elements of $\mathbf{c}_k^{UE}$ and $\mathbf{C}_s^{BS}$ can be written as $\alpha e^{j\phi}$ where α and ϕ are random variables with a given distribution. For instance, for the case that α has a log-normal and ϕ has a uniform distribution, the natural logarithm of α is a normal random variable with mean μ_c and variance σ_c^2, i.e., $\alpha \sim \text{Lognormal}(\mu_c^{UE}, (\sigma_c^{UE})^2)$ and $\phi \sim \mathcal{U}(-\phi_{\max}^{UE}, \phi_{\max}^{UE})$ for HWI at the UE and $\alpha \sim \text{Lognormal}(\mu_c^{BS}, (\sigma_c^{BS})^2)$ and $\phi \sim \mathcal{U}(-\phi_{\max}^{BS}, \phi_{\max}^{BS})$ for HWI at the BS.

Following the terminology in [59, 60, 4, 62], the PUPE of the system P_e is defined as the sum of the probability of misdetection P_{md} and the probability of false alarm P_{fa}, i.e., $P_e = P_{\text{md}} + P_{\text{fa}}$. P_{md} and P_{fa} are given as

$$P_{\text{md}} = \frac{\mathbb{E}\left[\sum_{k\in\mathcal{K}_a} \mathbb{1}_{\{\mathbf{m}_k \notin \mathcal{L}\}}\right]}{K_a}, \tag{5.3}$$

$$P_{\text{fa}} = \mathbb{E}\left[\frac{|\mathcal{L} \setminus \{\mathbf{m}_k : k \in \mathcal{K}_a\}|}{|\mathcal{L}|}\right], \tag{5.4}$$

where $\mathcal{L}$ is the list of the decoded messages produced by the receiver, $\mathcal{K}_a$ is the

set of active users, $\mathbb{1}_{\{\cdot\}}$ is the indicator function, and $|\cdot|$ denotes the cardinality of a set.

The energy efficiency of the system is measured in terms of the required $\frac{E_b}{N_0}$, which can be written as

$$\frac{E_b}{N_0} = \frac{LP}{BN_0}, \tag{5.5}$$

where E_b is the energy-per-bit. The aim is to minimize the required $\frac{E_b}{N_0}$ while achieving a target PUPE of ϵ.

5.2 Proposed Scheme

5.2.1 Encoding

We divide the transmission slot into pilot and data parts with lengths of n_p and n_d, respectively to encode the user messages in each slot. We assume that each user picks a pilot sequence from a non-orthogonal pilot codebook $\mathbf{A} \in \mathbb{C}^{n_p \times N}$ based on the first J message bits $\mathbf{m}_p$ for the transmission in the pilot part, where $N = 2^J$. Each column of $\mathbf{A}$ is normalized to have a squared Euclidean norm $n_p P$. The pilot part of the received signal in the s-th slot for additive and multiplicative HWIs can be written as

$$\mathbf{Y}_{s,p,a} = (\mathbf{A}_s + \boldsymbol{\eta}_{s,p}^{UE})\mathbf{H} + \boldsymbol{\eta}_{s,p}^{BS} + \mathbf{Z}_{s,p}, \tag{5.6}$$

$$\mathbf{Y}_{s,p,m} = ((\mathbf{A}_s \odot \mathbf{C}_{s,p}^{UE})\mathbf{H}) \odot \mathbf{C}_{s,p}^{BS} + \mathbf{Z}_{s,p}, \tag{5.7}$$

respectively, where $\mathbf{Y}_{s,p,a}$ and $\mathbf{Y}_{s,p,m} \in \mathbb{C}^{n_p \times M}$ are the pilot parts of the received signal in the s-th slot with additive and multiplicative HWI, respectively, $\mathbf{A}_s$ is the matrix with the selected pilots as its columns, $\mathbf{H} = [\mathbf{h}_1^T, \mathbf{h}_2^T, \ldots, \mathbf{h}_{K_s}^T]^T \in \mathbb{C}^{K_s \times M}$ is

the channel matrix consisting of user channels, K_s is the number of transmitting users in the s-th slot, and $\mathbf{Z}_{s,p}$ is the first n_p rows of $\mathbf{Z}_s$. $\boldsymbol{\eta}_{s,p}^{UE} \in \mathbb{C}^{n_p \times K_s}$ is the first n_p rows of the set of additive HWI vectors of the transmitting users in the s-th slot, and $\boldsymbol{\eta}_{s,p}^{BS}$ is the first n_p rows of $\boldsymbol{\eta}_s^{BS}$. $\mathbf{C}_{s,p}^{UE}$ is the first n_p rows of the set of the multiplicative HWI vectors of the transmitting users in the s-th slot and $\mathbf{C}_{s,p}^{BS}$ is the multiplicative HWI at the BS affecting the pilot part of the received signal in the s-th slot.

To encode the last $B - J$ message bits $\mathbf{m}_d$, we add r cyclic redundancy check (CRC) bits to $\mathbf{m}_d$ and employ a polar code of length $2n_d$ followed by quadrature phase shift keying (QPSK) modulation. The data part of the received signal in the s-th slot for the additive and multiplicative HWIs can be written as

$$\mathbf{Y}_{s,d,a} = \left(\mathbf{S}_s + \boldsymbol{\eta}_{s,d}^{UE}\right)\mathbf{H} + \boldsymbol{\eta}_{s,d}^{BS} + \mathbf{Z}_{s,d}, \tag{5.8}$$

$$\mathbf{Y}_{s,d,m} = \left(\left(\mathbf{S}_s \odot \mathbf{C}_{s,d}^{UE}\right)\mathbf{H}\right) \odot \mathbf{C}_{s,d}^{BS} + \mathbf{Z}_{s,d}, \tag{5.9}$$

respectively, where $\mathbf{Y}_{s,d,a}, \mathbf{Y}_{s,d,m} \in \mathbb{C}^{n_d \times M}$ corresponds to the last n_d rows of $\mathbf{Y}_{s,a}$ and $\mathbf{Y}_{s,m}$, respectively, $\mathbf{S}_s \in \mathbb{C}^{n_d \times K_s}$ with elements in $\{\sqrt{P/2}(\pm 1 \pm j)\}$ denotes the matrix of the user symbols transmitting in the s-th slot, $\boldsymbol{\eta}_{s,d}^{UE} \in \mathbb{C}^{n_d \times K_s}$ and $\boldsymbol{\eta}_{s,d}^{BS} \in \mathbb{C}^{n_d \times M}$ are the last n_d rows of the set of additive HWI vectors of the transmitting users in the s-th slot and $\boldsymbol{\eta}_s^{BS}$, respectively, $\mathbf{C}_{s,d}^{UE} \in \mathbb{C}^{n_d \times K_s}$ and $\mathbf{C}_{s,d}^{BS} \in \mathbb{C}^{n_d \times M}$ are the last n_d rows of the set of the multiplicative HWI vectors of the transmitting users in the s-th slot and $\mathbf{C}_s^{BS}$, respectively, and $\mathbf{Z}_{s,d}$ is the last n_d rows of $\mathbf{Z}_s$. The encoding process is illustrated in Figure 5.1, where $b(\mathbf{m}_p)$ is the pilot index corresponding to $\mathbf{m}_p$, and the user index is dropped for the ease of presentation.

5.2.2 Iterative Hardware Impairment Aware Decoding

We follow an iterative decoding approach at the receiver where we first recover the set of selected pilot signatures $\mathcal{I}$ and estimate the corresponding channel

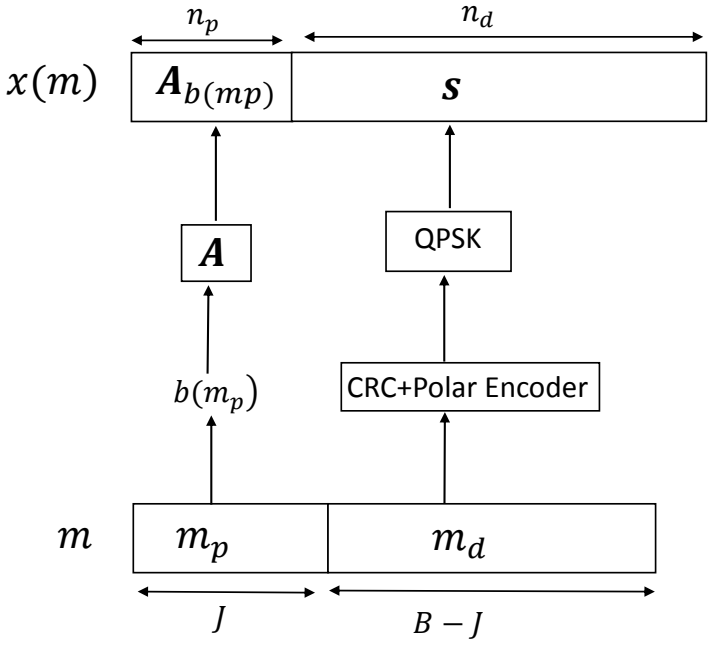

Figure 5.1: Encoding process of the user messages for the proposed scheme.

vectors followed by the decoding of the data part, accompanied by successive interference cancellation to improve the performance. We have presented the original version of this scheme with some slight differences for the case of no hardware impairments in [4], which can also be used in the presence of hardware impairments (as an HWI ignorant receiver). However, we can modify the receiver blocks using the statistics of hardware impairments to improve the performance. We explain these processes in detail in the following, where we drop the slot index s to simplify the notation.

5.2.2.1 Activity Detection

The detection of the active pilot sequences is essentially a compressed sensing problem as $K_a \ll N$, where $\mathbf{A}$ is the sensing matrix. In order to solve this problem, we employ the gOMP algorithm [73], which is a generalization of orthogonal matching pursuit (OMP) [74]. OMP is a greedy iterative algorithm that is used to solve compressed sensing problems by finding the column of the sensing matrix having the maximum correlation with the current residual signal, adding its index

to the output list, and subtracting the effect of the detected column at each iteration. On the other hand, gOMP allows for adding multiple indices to the output list at each iteration rather than only one index, which reduces complexity.

To implement the gOMP algorithm, we first calculate the following correlation in each iteration

$$\mathbf{R}^{(t)} = \mathbf{A}^H \mathbf{Y}_{p,i}^{(t)}, \tag{5.10}$$

where $\mathbf{R}^{(t)} \in \mathbb{C}^{N \times M}$, $\mathbf{Y}_{p,i}^{(t)}$ is the residual received pilot signal with HWI at the t-th gOMP iteration, $i \in \{a, m\}$ where a is used for the additive HWI and m for the multiplicative one, and $(\cdot)^H$ denotes the Hermitian operator. We set $\mathbf{Y}_p^{(0)} = \mathbf{Y}_p$. We then calculate the Euclidean norm of each row of $\mathbf{R}^{(t)}$ to obtain $\mathbf{R}_c^{(t)}$ by exploiting the receive (antenna) diversity, and take $i_{\text{OMP}} = \left\lceil (\frac{K_a}{V} + \Delta)/n_{\text{OMP}} \right\rceil$ largest elements of $\mathbf{R}_c^{(t)}$, where $\frac{K_a}{V}$ is the average active user load in a slot. Here, Δ is an integer parameter used to minimize the undercount probability, and n_{OMP} is the number of gOMP iterations. We add the selected indices to the list of detected signatures $\hat{\mathcal{I}}$, and subtract their effects using their projection on the received pilot signal. We continue the iterations for n_{OMP} steps, then we perform thresholding to decrease the number of overcounts. The details of this process are similar to the algorithm used in [4].

5.2.2.2 Channel Estimation

We estimate the user channel vectors employing the detected pilot signatures via a linear minimum mean square error (LMMSE) solution. For this purpose, for the case of additive HWIs, given the received pilot signal $\mathbf{Y}_{p,a}$, taking the expectation over the channel vectors, hardware impairments and noise, and using the fact that the additive distortions due to the hardware impairments are independent of each other and the AWGN term, the auto-covariance matrix $\mathbf{C}_{Y,a}$ and the cross-covariance matrix $\mathbf{C}_{\tilde{H}Y,a}$ can be calculated as

$$
\begin{aligned}
\mathbf{C}_{\tilde{H}Y,a} &= \mathbb{E}[\tilde{\mathbf{H}}\mathbf{Y}_{p,a}^{H}] \\
&= \mathbb{E}[\tilde{\mathbf{H}}(\tilde{\mathbf{H}}^{H}(\hat{\mathbf{A}} + \boldsymbol{\eta}_p^{UE})^{H} + (\boldsymbol{\eta}_p^{BS})^{H} + \mathbf{Z}_p^{H})] \\
&= \mathbb{E}[\tilde{\mathbf{H}}\tilde{\mathbf{H}}^{H}(\hat{\mathbf{A}}^{H} + (\boldsymbol{\eta}_p^{UE})^{H})] \\
&= M\hat{\mathbf{A}}^{H},
\end{aligned}
\tag{5.11}
$$

$$
\begin{aligned}
\mathbf{C}_{Y,a} &= \mathbb{E}[\mathbf{Y}_{p,a}\mathbf{Y}_{p,a}^{H}] \\
&= \mathbb{E}\left[(\hat{\mathbf{A}} + \boldsymbol{\eta}_p^{UE})\tilde{\mathbf{H}}\tilde{\mathbf{H}}^{\mathbf{H}}\left(\hat{\mathbf{A}}^{H} + (\boldsymbol{\eta}_p^{UE})^{H}\right)\right] \\
&\quad + \mathbb{E}\left[\boldsymbol{\eta}_p^{BS}(\boldsymbol{\eta}_p^{BS})^{H}\right] + \mathbb{E}\left[\mathbf{Z}_p\mathbf{Z}_p^{H}\right] \\
&= M\left(\hat{\mathbf{A}}\hat{\mathbf{A}}^{H} + \mathbb{E}[\boldsymbol{\eta}_p^{UE}(\boldsymbol{\eta}_p^{UE})^{H})]\right) \\
&\quad + \mathbb{E}\left[(\kappa^{BS})^2 P\left\|\tilde{\mathbf{H}}\right\|_F^2\right]\mathbf{I}_{n_p} + MN_0\mathbf{I}_{n_p} \\
&= M\left(\hat{\mathbf{A}}\hat{\mathbf{A}}^{H} + \left(K_s P((\kappa^{UE})^2 + (\kappa^{BS})^2) + N_0\right)\mathbf{I}_{n_p}\right),
\end{aligned}
\tag{5.12}
$$

respectively, where $\hat{\mathbf{A}}$ is the set of detected pilots, $\tilde{\mathbf{H}}$ is the set of channel vectors corresponding to the detected signatures, and $\|\cdot\|_F$ denotes the Frobenius norm.

Similarly, in order to estimate the channel vectors of the users for the case of multiplicative HWIs, the cross-covariance and auto-covariance matrices of $\mathbf{Y}_{p,m}$ can be calculated as

$$
\begin{aligned}
\mathbf{C}_{\tilde{H}Y,m} &= \mathbb{E}[\tilde{\mathbf{H}}\mathbf{Y}_{p,m}^{H}] \\
&= \mathbb{E}[\tilde{\mathbf{H}}(\mathbf{Z}_p^{H} + (\mathbf{C}_p^{BS})^{H} \odot (\tilde{\mathbf{H}}^{H}((\mathbf{C}_p^{UE})^{H} \odot \hat{\mathbf{A}}^{H})))] \\
&= e^{\left(\mu_c^{BS} + \frac{(\sigma_c^{BS})^2}{2}\right)}\mathbb{E}[\tilde{\mathbf{H}}(\tilde{\mathbf{H}}^{H}e^{\left(\mu_c^{UE} + \frac{(\sigma_c^{UE})^2}{2}\right)}\tilde{\mathbf{A}}^{H})] \\
&= e^{\left(\mu_c^{BS} + \mu_c^{UE} + \frac{(\sigma_c^{UE})^2 + (\sigma_c^{BS})^2}{2}\right)}\mathbb{E}[\tilde{\mathbf{H}}\tilde{\mathbf{H}}^{H}\hat{\mathbf{A}}^{H}] \\
&= Me^{\left(\mu_c^{BS} + \mu_c^{UE} + \frac{(\sigma_c^{UE})^2 + (\sigma_c^{BS})^2}{2}\right)}\tilde{\mathbf{A}}^{H},
\end{aligned}
\tag{5.13}
$$

$$
\begin{aligned}
\mathbf{C}_{Y,m} &= \mathbb{E}[\mathbf{Y}_{p,m}\mathbf{Y}_{p,m}^{H}] \\
&= \mathbb{E}[(((\hat{\mathbf{A}} \odot \mathbf{C}_p^{UE})\tilde{\mathbf{H}}) \odot \mathbf{C}_p^{BS} + \mathbf{Z}_p)(\mathbf{Z}_p^{H} + (\mathbf{C}_p^{BS})^{H} \\
&\quad \odot (\tilde{\mathbf{H}}^{H}((\mathbf{C}_p^{UE})^{H} \odot \hat{\mathbf{A}}^{H})))] \\
&= \mathbf{G} + MN_0\mathbf{I}_{n_p},
\end{aligned}
\tag{5.14}
$$

respectively, where $\exp\left(\mu_c^{BS} + \frac{(\sigma_c^{BS})^2}{2}\right)$ and $\exp\left(\mu_c^{UE} + \frac{(\sigma_c^{UE})^2}{2}\right)$ are the mean of the HWI at the BS and the UE sides, respectively, and $\mathbf{G}$ is given as

$$\mathbf{G} = \mathbb{E}[\tilde{\mathbf{A}}\tilde{\mathbf{H}} \odot \mathbf{C}_p^{BS}((\mathbf{C}_p^{BS})^H \odot (\tilde{\mathbf{H}}^H \tilde{\mathbf{A}}^H))], \tag{5.15}$$

where $\tilde{\mathbf{A}} = \hat{\mathbf{A}} \odot \mathbf{C}_p^{UE}$. Then, the ij-th element of $\mathbf{G}$ can be determined as

$$\begin{aligned}
\mathbf{G}_{i,j} &= \mathbb{E}\left[(\tilde{\mathbf{A}}_{(i,:)}\tilde{\mathbf{H}}) \odot \mathbf{C}_{p(i,:)}^{BS}((\mathbf{C}_{p(j,:)}^{BS})^H \odot (\tilde{\mathbf{H}}^H \tilde{\mathbf{A}}_{(:,j)}^H))\right] \\
&= \mathbb{E}\left[(\tilde{\mathbf{B}}_{(i,:)} \odot \mathbf{C}_{p(i,:)}^{BS})\left((\mathbf{C}_{p(j,:)}^{BS})^H \odot (\tilde{\mathbf{B}}_{(j,:)})^H\right)\right] \\
&= \mathbb{E}\left[\sum_r \tilde{b}_{i,r} c_{i,r}^{BS} \tilde{b}_{j,r}^* (c_{j,r}^*)^{BS}\right] \\
&= \begin{cases} \sum_r \mathbb{E}\left[\left|\tilde{b}_{i,r}\right|^2\right]((\mu_C^{BS})^2 + (\sigma_C^{BS})^2) & \text{if } i = j \\ \sum_r \mathbb{E}\left[\tilde{b}_{i,r}\tilde{b}_{j,r}^*\right](\mu_C^{BS})^2 & \text{if } i \neq j, \end{cases}
\end{aligned} \tag{5.16}$$

where $\mu_C^{BS} = e^{\mu_c^{BS} + (\sigma_c^{BS})^2/2}$, $(\sigma_C^{BS})^2 = (e^{(\sigma_c^{BS})^2} - 1)e^{(2\mu_c^{BS} + (\sigma_c^{BS})^2)}$, $\mathbf{X}_{(:,j)}$ is the j-th column of $\mathbf{X}$ and $\tilde{\mathbf{B}} = \tilde{\mathbf{A}}\tilde{\mathbf{H}}$. The expectations in the last step of (16) can be calculated as

$$\begin{aligned}
\mathbb{E}\left[\left|\tilde{b}_{i,r}\right|^2\right] &= \mathbb{E}\left[\sum_s \left|\tilde{a}_{i,s}\tilde{h}_{s,r}\right|^2\right] \\
&= \sum_s \mathbb{E}\left[|\tilde{a}_{i,s}|^2\right]\mathbb{E}\left[\left|\tilde{h}_{s,r}\right|^2\right] \\
&= \sum_s \mathbb{E}\left[|\tilde{a}_{i,s}|^2\right] \\
&= \sum_s |a_{i,s}|^2 \mathbb{E}\left[|c_{i,s}|^2\right] \\
&= \left((\mu_C^{UE})^2 + (\sigma_C^{UE})^2\right)\sum_s |a_{i,s}|^2,
\end{aligned} \tag{5.17}$$

$$\begin{aligned}\mathbb{E}\left[\tilde{b}_{i,r}\tilde{b}^{*}_{j,r}\right] &= \mathbb{E}[\tilde{\mathbf{A}}_{(i,:)}\tilde{\mathbf{H}}_{(:,r)}\tilde{\mathbf{H}}^{\mathbf{H}}_{(r,:)}\tilde{\mathbf{A}}^{H}_{(:,j)}] \\ &= \sum_{s}\mathbb{E}[\tilde{a}_{i,s}\tilde{a}^{*}_{j,s}] \\ &= \sum_{s} a_{i,s}a^{*}_{j,s}\mathbb{E}[c_{i,s}c^{*}_{j,s}] \\ &= (\mu_{C}^{UE})^{2}\sum_{s} a_{i,s}a^{*}_{j,s},\end{aligned} \tag{5.18}$$

where $\mu_{C}^{UE} = e^{\mu_{c}^{UE}+(\sigma_{c}^{UE})^{2}/2}$, $(\sigma_{C}^{UE})^{2} = (e^{(\sigma_{c}^{UE})^{2}}-1)e^{(2\mu_{c}^{UE}+(\sigma_{c}^{UE})^{2})}$. The estimated channel coefficients can be calculated as

$$\hat{\mathbf{H}}_{i} = \mathbf{W}_{i}\mathbf{Y}_{p,i}, \tag{5.19}$$

where $\hat{\mathbf{H}}_{i}$ is the set of the estimated channel vectors, and $\mathbf{W}_{i} = \mathbf{C}_{\tilde{H}Y,i}\mathbf{C}^{-1}_{Y,i}$ is the LMMSE matrix for the proposed scheme with additive or multiplicative HWIs indicated by $i \in \{a, m\}$. Note that the process is given for the first iteration, however, it can also be applied in the same way after changing $\mathbf{Y}_{p,i}$ with its residual after SIC, and updating $\hat{\mathbf{A}}$ by removing the pilot signatures corresponding to the decoded users in the subsequent iterations.

5.2.2.3 User Separation and Polar Decoding

We decode the data part by first employing MRC/MMSE for user separation and then utilizing a single-user polar decoder. Using MRC, the symbol estimates of the k-th user can be produced as follows

$$\hat{\mathbf{s}}_{k,i} = \mathbf{Y}^{(j)}_{d,i}\hat{\mathbf{h}}^{H}_{k,i}, \tag{5.20}$$

where $\mathbf{Y}^{(j)}_{d,i}$ is the data part of the received signal at the j-th decoding iteration, and $\hat{\mathbf{h}}_{k,i}$ is the channel vector estimate of the k-th user, which is the k-th row of $\hat{\mathbf{H}}_{i}$.

Alternatively, one can employ an MMSE estimator for the estimation of the user symbols. In the case that the receiver does not aware of the HWIs, the symbol estimates of the k-th user can be obtained as

$$\hat{\mathbf{s}}_{k,i} = \mathbf{Y}_{d,i}^{(j)} \left(\hat{\mathbf{H}}_{\mathbf{i}}^H \hat{\mathbf{H}}_{\mathbf{i}} + \frac{N_0}{P} \mathbf{I}_M \right)^{-1} \hat{\mathbf{h}}_{k,i}^H. \tag{5.21}$$

In addition, the MMSE estimation can be modified to make it more robust against HWIs using the HWI statistics. For this purpose, for the case of additive HWIs, the objective function can be defined as follows:

$$\begin{aligned}
\varepsilon_k &= \mathbb{E}\left[\|\mathbf{s}_k - \hat{\mathbf{s}}_k\|^2 \right] \\
&= \mathbb{E}_{s,\eta,Z}\left[\mathbf{s}_k^H \mathbf{s}_k - \mathbf{s}_k^H \mathbf{Y}_{d,a} \mathbf{W}_k - \mathbf{W}_k^H \mathbf{Y}_{d,a}^H \mathbf{s}_k + \mathbf{W}_k^H \mathbf{Y}_{d,a}^H \mathbf{Y}_{d,a} \mathbf{W}_k \right] \\
&= \mathrm{Tr}\{\mathbf{R}_{s,k}\} - 2\,\mathrm{Re}\{ \mathbb{E}_{s,\eta,Z}\left[\mathbf{s}_k^H \mathbf{Y}_{d,a} \mathbf{W}_k \right] \} + \mathbb{E}_{s,\eta,Z}\left[\mathbf{W}_k^H \mathbf{Y}_{d,a}^H \mathbf{Y}_{d,a} \mathbf{W}_k \right] \\
&= n_d P - 2\,\mathrm{Re}\left\{ \mathbb{E}\left[\mathbf{s}_k^H \left((\mathbf{s}_k + \boldsymbol{\eta}_{k,d}^{UE}) \mathbf{h}_k + \sum_{i,i\neq k} (\mathbf{s}_i + \boldsymbol{\eta}_{i,d}^{UE}) \mathbf{h}_i + \boldsymbol{\eta}_{s,d}^{BS} + \mathbf{Z}_{s,d} \right) \mathbf{W}_k \right] \right\} + \\
&\quad \mathbb{E}_{s,\eta,Z}\left[\mathbf{W}_k^H \left(\sum_k \mathbf{h}_k^H \left(\mathbf{s}_k^H + \left(\boldsymbol{\eta}_{k,d}^{UE} \right)^H \right) + \left(\boldsymbol{\eta}_{s,d}^{BS} \right)^H + \mathbf{Z}_{s,d}^H \right) \left(\sum_k \left(\mathbf{s}_k + \boldsymbol{\eta}_{k,d}^{UE} \right) \mathbf{h}_k + \boldsymbol{\eta}_{s,d}^{BS} + \mathbf{Z}_{s,d} \right) \mathbf{W}_k \right]
\end{aligned} \tag{5.22}$$

Then, using the fact that the HWIs are independent of each other and the AWGN term, we obtain

$$\begin{aligned}
\varepsilon_k &= n_d P - \mathbb{E}\left[\left(\mathbf{s}_k^H \mathbf{s}_k + \mathbf{s}_k^H \eta_{k,d}^{UE} \right) \mathbf{h}_k \mathbf{W}_k \right] - \mathbb{E}\left[\mathbf{W}_k^H \mathbf{h}_k^H (\mathbf{s}_k^H \mathbf{s}_k + (\eta_{k,d}^{UE})^H \mathbf{s}_k) \right] + \\
&\quad \mathbb{E}_{s,\eta,Z}\left[\mathbf{W}_k^H \left(\sum_k \mathbf{h}_k^H (\mathbf{s}_k^H + (\boldsymbol{\eta}_{k,d}^{UE})^H)(\mathbf{s}_k + \eta_{k,d}^{UE}) \mathbf{h}_k + (\boldsymbol{\eta}_{s,d}^{BS})^H (\boldsymbol{\eta}_{s,d}^{BS}) + n_d N_0 \mathbf{I}_M \right) \mathbf{W}_k \right] \\
&= n_d P - n_d P\, \mathbf{h}_k \mathbf{W}_k - n_d P\, \mathbf{W}_k^H \mathbf{h}_k^H \\
&\quad + \mathbb{E}\left[\mathbf{W}_k^H \left((n_d P + n_d (\kappa^{UE})^2 P) \mathbf{H}^H \mathbf{H}) + n_d \Upsilon^{BS} + n_d N_0 \mathbf{I}_M \right) \mathbf{W}_k \right] \\
&= n_d P - n_d P\, \mathbf{h}_k \mathbf{W}_k - n_d P\, \mathbf{W}_k^H \mathbf{h}_k^H + \mathrm{Tr}\{ \mathbf{X}_1 \mathbf{W}_k \mathbf{W}_k^H \}
\end{aligned} \tag{5.23}$$

where we take expectation over the transmitted symbols, noise, and hardware impairments, $\hat{\mathbf{s}}_k = \mathbf{Y}_{d,a} W_k$ is the estimated symbol, $\mathbf{W}_k \in \mathbb{C}^{M\times 1}$ is the separation matrix for the k-th user, $\mathbf{X}_1 = \mathbf{W}_k^H \left((n_d P + n_d (\kappa^{UE})^2 P) \mathbf{H}^H \mathbf{H}) + n_d \Upsilon^{BS} + n_d N_0 \mathbf{I}_M \right)$,

and we use the rotation of trace for the last term in the equation. Then, a closed from for the separation matrix can be obtained by taking the derivative of ε_k with respect to $\mathbf{W}_k$ to minimize ε_k as

$$\frac{\partial \gamma}{\partial \mathbf{W}} \varepsilon_k = -n_d P \mathbf{h}_k^T + \mathbf{X}_1^T \mathbf{W}_k^* = 0 \rightarrow \mathbf{X}_1^T \mathbf{W}_k^* = n_d P \mathbf{h}_k^T \tag{5.24}$$

Therefore, we obtain $\mathbf{W}_k^* = n_d P (\mathbf{X}_1^T)^{-1} \mathbf{h}_k^T$, hence $\mathbf{W}_k = n_d P (\mathbf{X}_1^H)^{-1} \mathbf{h}_k^H$.

Similarly, for the case of multiplicative HWIs, the objective function of the k-th user can be defined as

$$\begin{aligned}
\varepsilon_k &= \mathbb{E}[\|\mathbf{s}_k - \hat{\mathbf{s}}_k\|^2] \\
&= \mathbb{E}_{s,\mathbf{c},Z}[(\mathbf{s}_k - \mathbf{Y}_{d,m}\mathbf{W}_k)^H (\mathbf{s}_k - \mathbf{Y}_{d,m}\mathbf{W}_k)] \\
&= \mathrm{Tr}\{\mathbf{R}_{s,k}\} - 2\,\mathrm{Re}\big\{\mathbb{E}_{s,\mathbf{c},Z}[\mathbf{s}_k^H \mathbf{Y}_{d,m}\mathbf{W}_k]\big\} + \mathbb{E}_{s,\mathbf{c},Z}[\mathbf{W}_k^H \mathbf{Y}_{d,m}^H \mathbf{Y}_{d,m}\mathbf{W}_k] \\
&= n_d P - 2\,\mathrm{Re}\Bigg\{\mathbb{E}_{s,\mathbf{c},Z}\left[\mathbf{s}_k^H \left(\left((\mathbf{s}_k \odot \mathbf{c}_{k,d}^{UE})\mathbf{h}_k + \sum_{i,i\neq k}(\mathbf{s}_i \odot \mathbf{c}_{i,d}^{UE})\mathbf{h}_i\right) \odot \mathbf{C}_{s,d}^{BS} + \mathbf{Z}_{s,d}\right)\mathbf{W}_k\right]\Bigg\} + \\
&\quad \mathbb{E}_{s,\mathbf{c},Z}\left[\mathbf{W}_k^H \left(\left(\sum_k \mathbf{h}_k^H (\mathbf{s}_k^H \odot (\mathbf{c}_{k,d}^{UE})^H)\right) \odot (\mathbf{C}_{s,d}^{BS})^H + \mathbf{Z}_{s,d}^H\right)\left(\left(\sum_k (\mathbf{s}_k \odot \mathbf{c}_{k,d}^{UE})\mathbf{h}_k\right) \odot \mathbf{C}_{s,d}^{BS} + \mathbf{Z}_{s,d}\right)\mathbf{W}_k\right]
\end{aligned} \tag{5.25}$$

Then, using the fact that the HWIs are independent of each other and the AWGN term, we obtain

$$\begin{aligned}
\varepsilon_k &= n_d P - 2\,\mathrm{Re}\big\{\mathbb{E}_{s,\mathbf{c},Z}\left[\mathbf{s}_k^H \left(((\mathbf{s}_k \odot \boldsymbol{\mu_1})\mathbf{h}_k) \odot \boldsymbol{\mu_2}\right)\mathbf{W}_k\right]\big\} + \\
&\quad \mathbb{E}_{s,\mathbf{c},Z}\left[\mathbf{W}_k^H \left(\sum_k \left(\mathbf{h}_k^H (\mathbf{s}_k^H \odot (\mathbf{c}_{k,d}^{UE})^H)\right) \odot (\mathbf{C}_{s,d}^{BS})^H \left((\mathbf{s}_k \odot \mathbf{c}_{k,d}^{UE})\mathbf{h}_k\right) \odot \mathbf{C}_{s,d}^{BS} + n_d N_0 \mathbf{I}_M\right)\mathbf{W}_k\right] \\
&= n_d P - n_d P\, e^{\alpha} \mathbf{h}_k \mathbf{W}_k - n_d P\, e^{\alpha} \mathbf{W}_k^H \mathbf{h}_k^H \\
&\quad + n_d N_0 \mathbf{W}_k^H \mathbf{I}_M \mathbf{W}_k + \mathbb{E}\left[\mathbf{W}_k^H \left(\sum_k \left(\mathbf{h}_k^H \tilde{\mathbf{s}}_{\mathbf{k}}^H\right) \odot (\mathbf{C}_{s,d}^{BS})^H (\tilde{\mathbf{s}}_{\mathbf{k}} \mathbf{h}_k) \odot \mathbf{C}_{s,d}^{BS}\right)\mathbf{W}_k\right]
\end{aligned} \tag{5.26}$$

where we take expectation over the transmitted symbols, noise, and hardware impairments, $\boldsymbol{\mu}_1 = \exp\left(\mu_c^{BS} + \frac{(\sigma_c^{BS})^2}{2}\right)\mathbf{1}^{n_p \times M}$, $\boldsymbol{\mu}_2 = \exp\left(\mu_c^{UE} + \frac{(\sigma_c^{UE})^2}{2}\right)\mathbf{1}^{n_p \times K_s}$, $\tilde{\mathbf{s}}_{\mathbf{k}} = \mathbf{s}_k \odot \mathbf{c}_{k,d}^{UE}$, and $\alpha = \left(\mu_c^{BS} + \mu_c^{UE} + \frac{(\sigma_c^{UE})^2 + (\sigma_c^{BS})^2}{2}\right)$.

Then, $\mathbf{F} = \mathbb{E}\left[\sum_k \left(\mathbf{h}_k^H \tilde{\mathbf{s}}_\mathbf{k}^H\right) \odot (\mathbf{C}_{s,d}^{BS})^H (\tilde{\mathbf{s}}_\mathbf{k}\mathbf{h}_k) \odot \mathbf{C}_{s,d}^{BS}\right]$ can be calculated as

$$\begin{aligned}\mathbf{F} &= \mathbb{E}\left[\sum_k \left(\mathbf{h}_k^H \tilde{\mathbf{s}}_\mathbf{k}^H\right) \odot (\mathbf{C}_{s,d}^{BS})^H (\tilde{\mathbf{s}}_\mathbf{k}\mathbf{h}_k) \odot \mathbf{C}_{s,d}^{BS}\right] \\ &= \mathbb{E}\left[\sum_k \left(\tilde{\mathbf{R}}_k^H \odot (\mathbf{C}_{s,d}^{BS})^H\right)\left(\tilde{R}_k \odot \mathbf{C}_{s,d}^{BS}\right)\right]\end{aligned} \tag{5.27}$$

where $\tilde{\mathbf{R}}_k = \tilde{\mathbf{s}}_\mathbf{k}\mathbf{h}_k$. Then, the ij-th element of $\mathbf{F}$ can be calculated as

$$\begin{aligned}\mathbf{F}_{i,j} &= \mathbb{E}\left[(\tilde{\mathbf{R}}_{\mathbf{k}(i,:)}^{\ H}) \odot (\mathbf{C}_{s(i,:)}^H)^{BS} (\tilde{\mathbf{R}}_{\mathbf{k}(:,j)} \odot (\mathbf{C}_{s(:,j)}^{BS}))\right] \\ &= \mathbb{E}\left[\sum_k \sum_r \tilde{r}_{k(r,i)}^* \left(c_{r,i}^{BS}\right)^* \tilde{r}_{k(r,j)} c_{r,j}^{BS}\right] \\ &= \begin{cases} \sum_k \sum_r \mathbb{E}\left[\left|\tilde{r}_{k(r,i)}\right|^2\right] ((\mu_C^{BS})^2 + (\sigma_C^{BS})^2) & \text{if } i = j \\ \sum_k \sum_r \mathbb{E}\left[\tilde{r}_{k(r,i)} \tilde{r}_{k(r,j)}^*\right] (\mu_C^{BS})^2 & \text{if } i \neq j, \end{cases}\end{aligned} \tag{5.28}$$

where $\mu_C^{BS} = e^{\mu_c^{BS} + (\sigma_c^{BS})^2/2}$, $(\sigma_C^{BS})^2 = (e^{(\sigma_c^{BS})^2} - 1) e^{(2\mu_c^{BS} + (\sigma_c^{BS})^2)}$, $\mathbf{X}_{(:,j)}$ is the j-th column of $\mathbf{X}$. The expectations in the last step can be calculated as

$$\begin{aligned}\mathbb{E}\left[\left|\tilde{r}_{k(r,i)}\right|^2\right] &= \mathbb{E}\left[\sum_s \left|\tilde{s}_{r,s} h_{s,i}\right|^2\right] \\ &= \sum_s \mathbb{E}\left[\left|\tilde{s}_{r,s}\right|^2\right] \left|h_{s,i}\right|^2 \\ &= \sum_s \mathbb{E}\left[\left|s_{r,s}\right|^2\right] \mathbb{E}\left[\left|(c_k^{UE})_{r,s}\right|^2\right] \left|h_{s,i}\right|^2 \\ &= \left((\mu_C^{UE})^2 + (\sigma_C^{UE})^2\right) n_d P \sum_s \left|h_{s,i}\right|^2\end{aligned} \tag{5.29}$$

$$
\begin{aligned}
\mathbb{E}\left[\tilde{r}_{k(r,i)}\tilde{r}^{*}_{k(r,j)}\right] &= \mathbb{E}\left[\sum_{s}\tilde{s}_{r,s}h_{s,i}\tilde{s}^{*}_{r,s}h^{*}_{s,j}\right] \\
&= \sum_{s}\mathbb{E}\left[\left|\tilde{s}_{r,s}\right|^{2}\right]h_{s,i}h^{*}_{s,j} \\
&= \sum_{s}\mathbb{E}\left[\left|s_{r,s}\right|^{2}\right]\mathbb{E}\left[\left|(c_{k}^{UE})_{r,s}\right|^{2}\right]h_{s,i}h^{*}_{s,j} \\
&= \left((\mu_{C}^{UE})^{2}+(\sigma_{C}^{UE})^{2}\right)n_{d}P\sum_{s}h_{s,i}h^{*}_{s,j}
\end{aligned}
\tag{5.30}
$$

where $\mu_{C}^{UE} = e^{\mu_{c}^{UE}+(\sigma_{c}^{UE})^{2}/2}$, $(\sigma_{C}^{UE})^{2}=(e^{(\sigma_{c}^{UE})^{2}}-1)e^{(2\mu_{c}^{UE}+(\sigma_{c}^{UE})^{2})}$. Therefore,

$$
\mathbf{F}_{i,j} = \begin{cases} n_{d}\,P((\mu_{C}^{UE})^{2}+(\sigma_{C}^{UE})^{2})((\mu_{C}^{BS})^{2}+(\sigma_{C}^{BS})^{2})\sum_{s}|h_{s,i}|^{2} & \text{if } i=j \\ n_{d}\,P((\mu_{C}^{UE})^{2}+(\sigma_{C}^{UE})^{2})((\mu_{C}^{BS})^{2})\sum_{s}h_{s,i}h^{*}_{s,j}, & \text{if } i\neq j \end{cases}
\tag{5.31}
$$

Taking the derivative of ε_k with respect to $\mathbf{W}_k$ to minimize ε_k:

$$
\begin{aligned}
\frac{\partial\gamma}{\partial\mathbf{W}}\varepsilon_{k} &= -n_{d}\,Pe^{\alpha}\mathbf{h}_{k}^{T}+n_{d}N_{0}\mathbf{W}_{k}^{*}+\mathbf{F}^{T}\mathbf{W}_{k}^{*} \\
&\rightarrow \mathbf{W}_{k}^{*}(n_{d}N_{0}\mathbf{I}_{M}+\mathbf{F}^{T}) = n_{d}Pe^{\alpha}\mathbf{h}_{k}^{T} \\
&\rightarrow \mathbf{W}_{k}^{*} = n_{d}Pe^{\alpha}(n_{d}N_{0}\mathbf{I}_{M}+\mathbf{F}^{T})^{-1}\mathbf{h}_{k}^{T} \\
&\rightarrow \mathbf{W}_{k} = n_{d}Pe^{\alpha}(n_{d}N_{0}\mathbf{I}_{M}+\mathbf{F}^{H})^{-1}\mathbf{h}_{k}^{H}
\end{aligned}
\tag{5.32}
$$

where $\alpha = \left(\mu_{c}^{BS}+\mu_{c}^{UE}+\frac{(\sigma_{c}^{UE})^{2}+(\sigma_{c}^{BS})^{2}}{2}\right)$.

We then extract the bit-wise log-likelihood ratio (LLR) values by treating the complex symbol estimate $\hat{\mathbf{s}}_{k,i}$ as a single-user channel output, and feed them to a single-user polar decoder utilizing successive cancellation list decoding (SCLD). The decoded sequence is deemed successful and added to the output list $\hat{\mathcal{D}}$ if it satisfies the CRC check. After this step, the decoded messages and the corresponding pilot signature estimates become available. Then, the successfully decoded message is re-encoded and modulated to obtain the actual QPSK symbols, and the estimate of the transmitted signal of the k-th user in the output list is obtained as follows

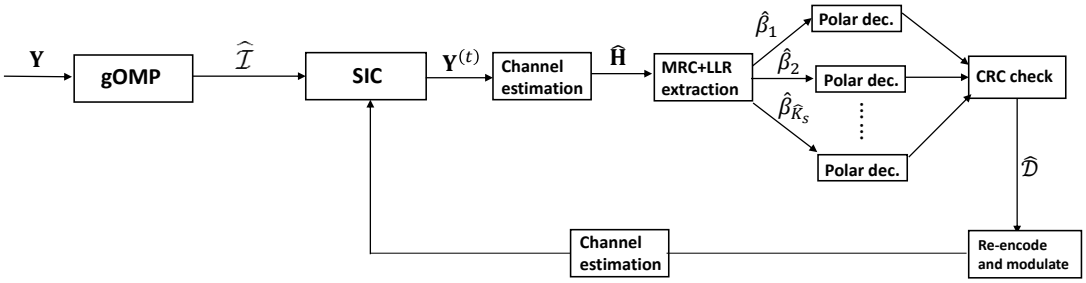

Figure 5.2: Decoding procedure of the proposed scheme in each slot.

$$\hat{\mathbf{x}}_k = \begin{bmatrix} \hat{\mathbf{A}}_k \\ \mathbf{s}'_k \end{bmatrix}, \tag{5.33}$$

where $\hat{\mathbf{A}}_k$ is the estimated pilot signature, and $\mathbf{s}'_k$ is the re-constructed transmitted symbol for the k-th user.

Ignoring the interference due to the non-decoded users, for the case that the receiver is not aware of the hardware impairments, the channel vectors corresponding to the successfully decoded messages using an LMMSE filter can be re-estimated as

$$\hat{\mathbf{H}}_{\text{SIC}} = (\hat{\mathbf{X}}_{\hat{\mathcal{D}}}^H \hat{\mathbf{X}}_{\hat{\mathcal{D}}} + N_0 \mathbf{I}_{|\hat{\mathcal{D}}|})^{-1} \hat{\mathbf{X}}_{\hat{\mathcal{D}}}^H \mathbf{Y}_i^{(j)}, \tag{5.34}$$

where $\hat{\mathbf{X}}_{\hat{\mathcal{D}}}$ is the matrix of the re-constructed transmitted signals for the successfully decoded user messages, and $\mathbf{Y}_i^{(j)}$ is the residual received signal at the j-th iteration for additive or multiplicative HWIs depending on i. Considering the HWIs, we modify the LMMSE procedure using the same way as in Sec. 5.2.2.2 by replacing the set of detected pilot signatures $\hat{\mathbf{A}}$ with the set of re-constructed transmitted signals $\hat{\mathbf{X}}_{\hat{\mathcal{D}}}$. Using these re-estimated channel vectors we perform SIC as

$$\mathbf{Y}_i^{(j+1)} = \mathbf{Y}_i^{(j)} - \hat{\mathbf{X}}_{\hat{\mathcal{D}}} \hat{\mathbf{H}}_{\text{SIC,i}}\,, \tag{5.35}$$

where $\hat{\mathbf{H}}_{\mathrm{SIC},i}$ is the set of re-estimated channel vectors obtained by HWI-aware LMMSE filtering. The iterations continue until there is no successfully decoded message in the current iteration, or a predetermined number of iterations is reached. The decoding process in a slot is illustrated in Figure 5.2.

5.3 Numerical Results

In this section, we present a performance evaluation of the proposed scheme with HWIs. We take $n = 3200$ and $B = 100$ to be consistent with the existing URA literature with a massive MIMO receiver, and $M = 50$. We choose the number of slots as $V = 4$, which leads to a slot length of 800, and set $n_p = 288$ and $n_d = 512$, employ 5G polar codes with a length of 1024, and set the CRC length to 16 and the list size of SCLD to 128 as in [4]. For the pilot codebook, we employ a sub-sampled DFT matrix with $N = 2^J$ with $J = 13$. We set $n_{\mathrm{OMP}} = 4$ and $\Delta = 0.2\frac{K_a}{V}$, which are selected based on our extensive simulations.

The energy efficiency of the proposed scheme is evaluated by the minimum required E_b/N_0 for $\epsilon = 0.05$ for different additive HWI scenarios in Figure 5.3. We consider two scenarios; (i) only the UE has HWI; (ii) there are HWIs on both the BS and the UE sides. We set both κ^{UE} and κ^{BS} to 0.4, which is a relatively high HWI level. The results in Figure 5.3 demonstrate that, although the presence of the HWIs can be tolerated, i.e., the scheme in [4] with re-estimation of the channel vectors for SIC (HWI-ignorant receiver) works with acceptable performance, the proposed hardware impairment aware solution is more energy efficient. For instance, the new algorithm decreases the required E_b/N_0 by 0.8 dB for $K_a = 800$ in scenario (i), and by 1 dB in scenario (ii) for $K_a = 700$. The proposed scheme can also increase the range of the supported active user load. For instance, the scheme can support up to 900 active users in scenario (i) and 800 active users in scenario (ii), while the scheme in [4] can support up to 800 and 700 active users, respectively. Moreover, the proposed scheme is more robust against HWIs compared to the HWI-ignorant receiver. For instance, increasing κ^{UE} and κ^{BS} from 0.3 to 0.4 for $K_a = 700$ in scenario (ii) deteriorates the performance of the

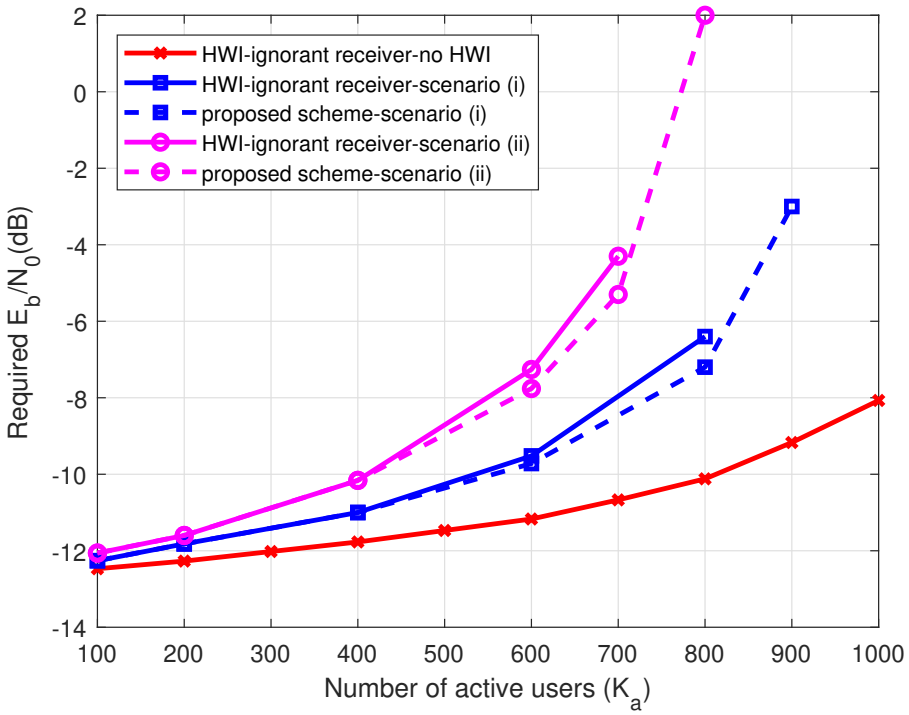

Figure 5.3: Required E_b/N_0 versus number of active users for different additive HWI scenarios and $P_e \leq 0.05$.

ignorant receiver by 4.2 dB while the performance loss for the proposed scheme is 3.4 dB. Note that we have opted for a comparison with the state-of-the-art scheme in [4] since this paper is the first work on URA with hardware impairments and there is no other URA scheme for direct comparison. Furthermore, the performance loss compared to the no HWI case is less than 3 dB in scenario (i) for $K_a \leq 800$, and 5.4 dB in scenario (ii) for $K_a \leq 700$.

We evaluate the energy efficiencies of the proposed scheme and the one in [4] with multiplicative HWIs in Figure 5.4. We assume that α follows a log-normal distribution with $\mu_c = 0$, $\sigma_c^2 = 0.1$, and ϕ has uniformly distributed with $\phi_{\max} = \frac{\pi}{9}$ at both the UE and the BS sides. The results in Figure 5.4 illustrate that with the HWI-aware receiver, the proposed scheme requires less energy to achieve the same PUPE in this case as well. For instance, it is superior to the scheme in [4] by 2.5 dB in scenario (i) for $K_a = 900$, and by 1 dB in scenario (ii) for $K_a = 700$. In addition, the proposed scheme can support 1000 active users in scenario (i), and 800 active users in scenario (ii), while the ignorant receiver in [4] can serve

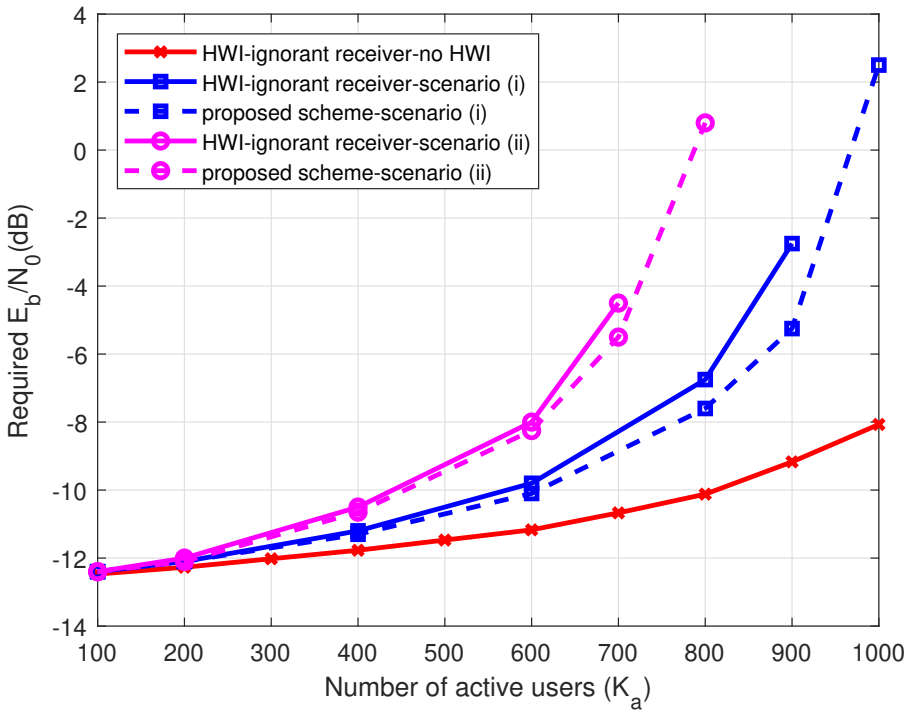

Figure 5.4: Required E_b/N_0 versus the number of active users with different multiplicative HWI scenarios and $P_e \leq 0.05$.

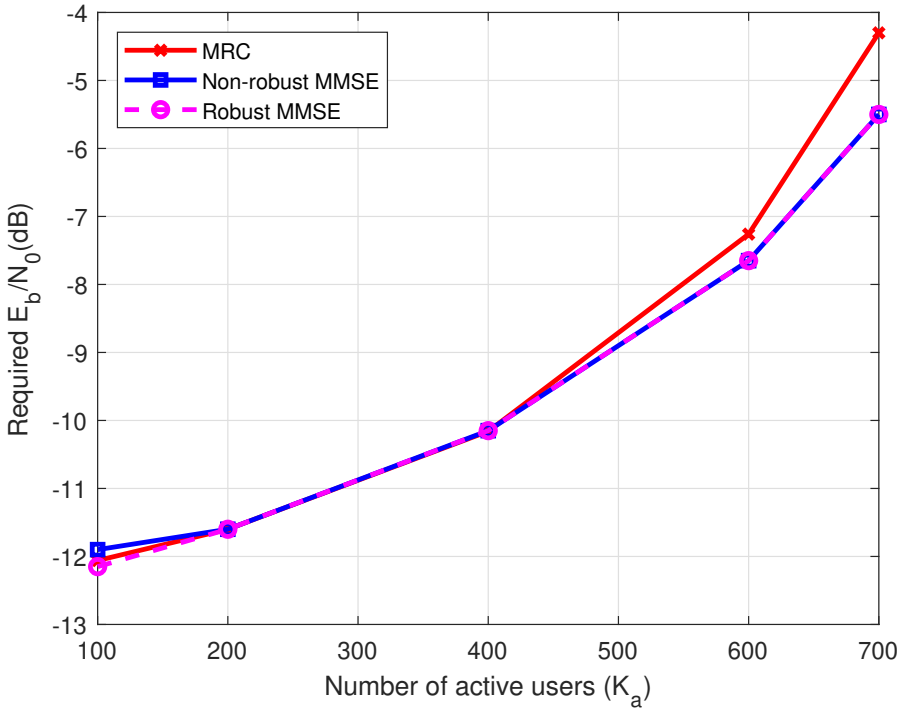

Figure 5.5: Required E_b/N_0 versus the number of active users with different user symbol estimators for additive HWIs.

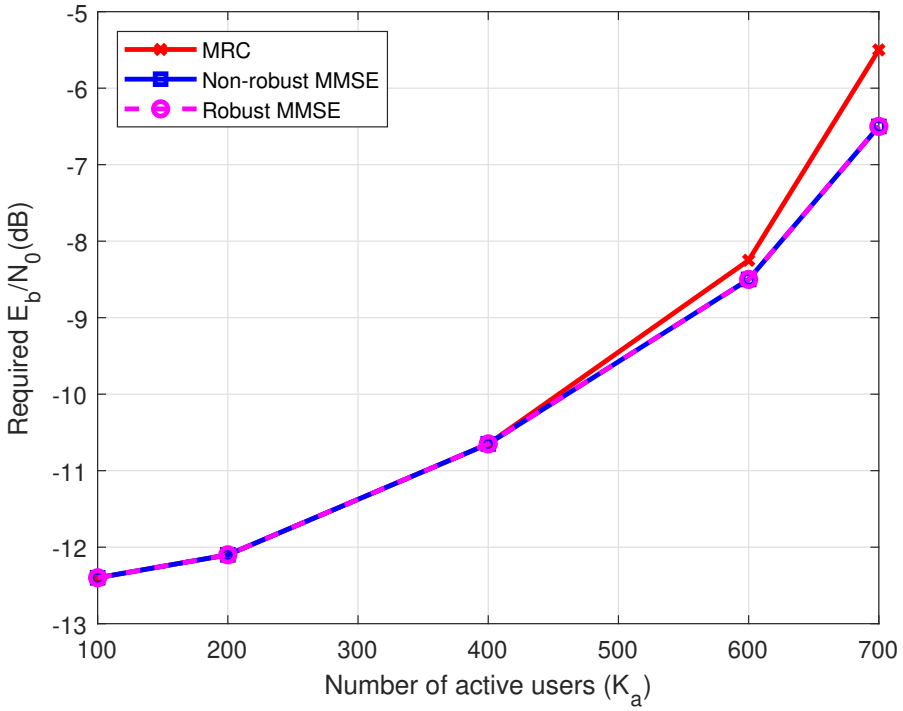

Figure 5.6: Required E_b/N_0 versus the number of active users with different user symbol estimators for multiplicative HWIs.

up to 900 and 700 active users in scenarios (i) and (ii), respectively.

We also compare the performance of different choices to estimate the transmitted user symbols and the derived robust MMSE with its non-robust counterpart for additive and multiplicative HWI scenarios in Figure 5.5 and 5.6, respectively. The results in both figures show that MMSE estimator can improve the performance of MRC up to 1 dB in high active user loads with the cost of increasing complexity, however, the robust MMSE derivation using the HWI statistics does not provide any performance advantage.

5.4 Chapter Summary

We investigated the URA problem with a massive MIMO receiver in the presence of HWIs. We proposed a coding scheme combining the ideas of slotting the transmission frame, gOMP for AD, LMMSE channel estimation, and polar coding

with the re-estimation of the channel vectors for SIC. We designed an HWI-aware receiver by exploiting the HWI statistics. Numerical examples demonstrate that while the existing schemes can still be used with some success, the proposed solution decreases the required energy by up to 2.5 dB while supporting more active users.

Chapter 6

Unsourced Random Access using ODMA and Polar Codes

In this chapter, we investigate ODMA [43], which is a newly proposed technique for non-orthogonal multiple access systems, where each user exploits a small fraction of time slots based on its transmission pattern in the context of URA. Specifically, in our proposed scheme, each active user divides its message into two parts, where the first part determines the transmission pattern (activity indices), and the second part is encoded with a polar code to be transmitted on the active indices. We propose an $l1$-norm based blind pattern detection method to recover the activity patterns of the active users followed by single-user polar decoding and successive interference cancellation (SIC). Numerical examples demonstrate that the proposed scheme is effective in both Gaussian and fading MAC. Specifically, it offers a superior performance in fading MAC, and a competitive performance with a lower complexity in Gaussian MAC with respect to the state-of-the-art.

The rest of the chapter is organized as follows. We introduce the system model in Section 6.1, and provide the details of the proposed solution in Section 6.2. We present a set of numerical results in Section 6.3, and conclude the chapter Section 6.4.

6.1 System Model

We consider a URA set-up, where K_a active users out of an unbounded number of total users transmit B bits of information to a common BS equipped with a single antenna through a transmission frame of n channel uses. The channel can be either a Gaussian MAC or a quasi-static fading MAC. If the transmission is over a quasi-static fading MAC, the received signal at the BS can be written as

$$\mathbf{y} = \sum_{k=1}^{K_a} h_k \mathbf{x}_k(\mathbf{m}_k) + \mathbf{z}, \tag{6.1}$$

where $\mathbf{m}_k$ is the message of the k-th user, $\mathbf{x}_k(\mathbf{m}_k) \in \mathbb{R}^{n\times 1}$ is the transmitted signal of the k-th user corresponding to the message $\mathbf{m}_k$, $\mathbf{y} \in \mathbb{C}^{n\times 1}$ is the received signal, $\mathbf{z} \in \mathbb{C}^{n\times 1}$ is the circularly symmetric complex additive white Gaussian noise (AWGN) consisting of independent and identically distributed (i.i.d.) elements with zero mean and variance N_0, i.e., $\mathbf{z} \sim \mathcal{CN}(\mathbf{0}, N_0\mathbf{I}_n)$, and $h_k \sim \mathcal{CN}(0,1)$ is the channel coefficient of the k-th user. Note that if the channel gains in (6.1) are deterministic, then the channel becomes a Gaussian MAC (GMAC).

We assume that the users utilize n_p symbols of the transmission frame for the pilot transmission (for the fading case) and n_c symbols out of the remaining $n-n_p$ ones for data transmission while the remaining symbol periods are unoccupied (by this specific user). The required energy per-bit of the system can be written as

$$\frac{E_b}{N_0} = \frac{n_p P_p + n_c P}{B N_0}, \tag{6.2}$$

where P_p is the symbol power of the pilot part and P is the symbol power of the data part. Note that there are no pilots in the GMAC case since there is no need for channel estimation.

The receiver aims to recover the list of messages $\mathcal{L}(y) = \{\hat{\mathbf{m}}_1, \ldots, \hat{\mathbf{m}}_{K_a}\}$ up to a permutation. The per-user probability of error (PUPE) is used as the performance

metric in URA, defined as

$$P_e = \frac{1}{K_a} \sum_{i=1}^{K_a} Pr(\mathbf{m}_i \notin \mathcal{L}(y)). \tag{6.3}$$

Our objective is to design a URA scheme with minimum required $\frac{E_b}{N_0}$ while satisfying $P_e \leq \epsilon$, where ϵ is the target PUPE.

6.2 Proposed Solution

6.2.1 Encoding

We assume that the message of each user is divided into two parts of $\mathbf{m}_p$ and $\mathbf{m}_c$ with lengths of B_p and $B_c = B - B_p$, respectively. If the channel is a fading MAC, we allocate the first n_p time instances for pilot transmission and the rest of the frame is utilized for the data transmission. In the pilot part, each user transmits a pilot sequence that is selected from a non-orthogonal pilot codebook $\mathbf{A} \in \mathbb{C}^{n_p \times M_s}$ based on its first B_p bits m_p where $M_s = 2^{B_p}$. On the other hand, the whole transmission frame is exploited for the data transmission in the GMAC scenario.

The encoding of the data part in both cases is performed as follows. The second part $\mathbf{m}_c$ is encoded using a $(n_c, B_c + r)$ polar code where r is the number of the cyclic redundancy check (CRC) bits, modulated by binary phase shift keying (BPSK), and transmitted according to a transmission pattern dictated by the first B_p message bits. Namely, each user picks a column from the pattern matrix $\mathbf{P}$ based on its first B_p bits. Each column of $\mathbf{P}$ has n_c non-zero elements denoting the active indices and the remaining elements corresponding to the idle indices. Then, the data part of the transmitted signal of the k-th user $\mathbf{x}_k$ can be formed by placing the elements of the BPSK modulated polar codeword of the k-th user $\mathbf{v}_k$ on to the active indices dictated by the p_k-th column of $\mathbf{P}$. Encoded signals

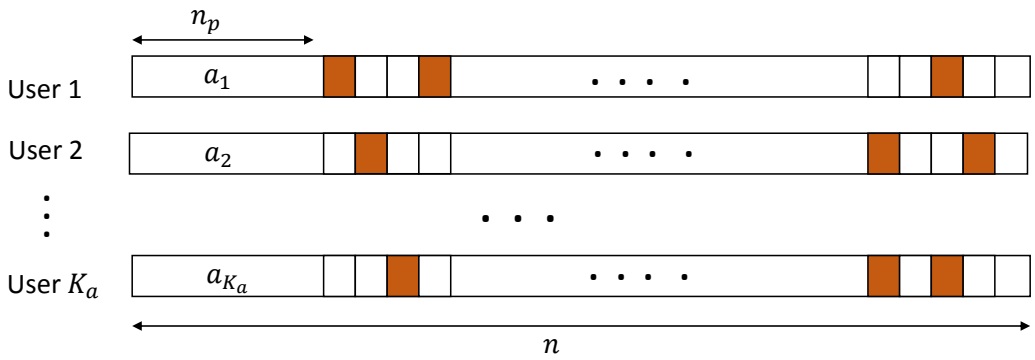

Figure 6.1: An illustration of the transmit signal structure for the fading MAC scenario. Colored boxes show the utilized symbol periods in the data part. For the GMAC scenario, there are no pilot symbols since there is no need for channel estimation.

of active users are illustrated in Figure 6.1 showing the ODMA idea, where $\mathbf{a}_k$ is the pilot sequence of the k-th user.

6.2.2 Receiver Operation

6.2.2.1 Pilot/Pattern Detection

For fading MAC, at the receiver side, we first need to detect the selected pilot sequences. We utilize the generalized orthogonal matching pursuit (gOMP) algorithm [73], which is a lower complexity generalization of OMP. To perform gOMP, we first calculate the following correlation between the candidate pilot sequences and the received pilot signal:

$$\mathbf{R} = \left|\mathbf{A}^H \mathbf{y}_p^{(j)}\right|, \tag{6.4}$$

where $\mathbf{y}_p^{(j)}$ is the received pilot at the j-th iteration that is initialized as $\mathbf{y}_p^{(1)} = \mathbf{y}_p$, where $\mathbf{y}_p = \mathbf{A}_s\mathbf{h} + \mathbf{z}_p$ is the received signal in the pilot part, $\mathbf{A}_s \in \mathbb{C}^{n_p \times K_a}$ is the matrix of the selected pilot sequences and $\mathbf{h} \in \mathbb{C}^{K_a \times 1}$ is the set of user channel coefficients. We add the i_{OMP} indices (depending on K_a and the number of gOMP iterations n_{OMP}) corresponding to the largest elements of $\mathbf{R}$ to the output list $\hat{\mathcal{I}}$,

subtract the effect of the detected sequences and continue the gOMP iterations for n_{OMP} times. Please refer to [5] for the further details on the application of gOMP in URA for pilot detection. Note that the transmission patterns are detected here as well as they are also picked based on the first B_p bits for fading MAC.

For the GMAC scenario, to detect the active transmission patterns, we employ an $l1$-norm energy test on the received signal, exploiting the sparseness of the scheme due to the short blocklengths of the utilized polar codes. Namely, for each transmission pattern, we calculate the received signal power at the corresponding locations of the received signal as follows

$$e_i = \left\| \mathbf{y}^{(j)} \odot \mathbf{p}_i \right\|_1 \quad i = 1, 2 \dots M_s, \tag{6.5}$$

where $\|.\|_1$ denotes the $l1$ norm, $\odot$ indicates the elementwise multiplication, and $\mathbf{y}^{(j)}$ is the residual signal at the j-th iteration which is initialized as $\mathbf{y}^{(1)} = \mathbf{y}$. We then order these signal powers and take the indices corresponding to the $K_b = K_a + \delta - n_{\text{dec}}$ largest elements, where δ is a small integer and n_{dec} is the number of successfully decoded users up to the j-th iteration assuming that K_a is known at the receiver. Note that K_a can be estimated with great success by an energy test similar to [30]. Also, the first B_p message bits are inherently decoded in this step.

6.2.2.2 Channel Estimation

For the case of fading MAC, we also need to estimate the channel coefficients. To do that, given the detected pilot sequences, we employ a linear minimum mean-square error (MMSE) solution to estimate the set of channel coefficients using

$$\hat{\mathbf{h}} = (\mathbf{A}_{\hat{\mathcal{I}}}^H \mathbf{A}_{\hat{\mathcal{I}}} + N_0 \mathbf{I}_{K_s})^{-1} \mathbf{A}_{\hat{\mathcal{I}}}^H \mathbf{y}_p^{(j)}, \tag{6.6}$$

where K_s is the cardinality of $\hat{\mathcal{I}}$ and $\mathbf{A}_{\hat{\mathcal{I}}}$ is the set of the columns of $\mathbf{A}$ specified by $\hat{\mathcal{I}}$.

6.2.2.3 Channel Decoding and SIC

Given the set of detected active transmission patterns $\hat{\mathbf{P}}$ (and, the channel coefficient estimates in the fading MAC scenario), we extract the LLR values of the codeword bits by treating interference as noise (TIN), and feed them to a single-user polar decoder. Indeed, given the transmission pattern of the intended user, the received signal at the active indices dictated by the transmission pattern becomes the output of a single-user AWGN/quasi-static fading channel assuming that the multiuser interference is Gaussian. We assume that the output sequence is successfully decoded and added to the output list $\hat{\mathcal{D}}$ if it satisfies the CRC check. We then re-encode and modulate the decoded messages to re-construct the transmitted symbols followed by SIC. In the GMAC scenario, we perform SIC as

$$\mathbf{y}^{(j+1)} = \mathbf{y}^{(j)} - \sum_{k \in \hat{\mathcal{D}}^{(j)}} \hat{\mathbf{x}}_k, \tag{6.7}$$

where $\hat{\mathbf{x}}_k$ is the re-constructed transmitted signal of the k-th user and $\hat{\mathcal{D}}^{(j)}$ is the set of the detected users at the j-th iteration. For SIC in the fading case, we re-estimate the channel coefficients using the decoded symbols together with the pilot sequences as follows

$$\hat{\mathbf{h}}_{\text{SIC}} = (\hat{\mathbf{X}}_{\hat{\mathcal{D}}}^H \hat{\mathbf{X}}_{\hat{\mathcal{D}}} + N_0 \mathbf{I}_{\left|\hat{\mathcal{D}}\right|})^{-1} \hat{\mathbf{X}}_{\hat{\mathcal{D}}}^H \mathbf{y}^{(j)}, \tag{6.8}$$

where $\hat{\mathbf{X}}_{\hat{\mathcal{D}}}$ is the set of re-constructed transmitted symbols and utilize these re-estimated channel coefficients for SIC as

$$\mathbf{y}^{(j+1)} = \mathbf{y}^{(j)} - \hat{\mathbf{X}}_{\hat{\mathcal{D}}} \hat{\mathbf{h}}_{\text{SIC}}. \tag{6.9}$$

We give the residual signal back to the pilot/pattern detection step to start a new iteration, and continue the iterations until no new message satisfies the CRC check in the current iteration, or a predetermined number of iterations is reached.

Note that since the users pick the transmission patterns independently, there can be pattern collisions when the first B_p bits of the two users are the same. The probability of this case can be adjusted by changing B_p. In our simulations, we choose B_p such that the collision of three or more users is negligible. Note that in the case of two collided users, both of them can be successfully decoded with a high probability if the rate of the polar code is less than 1/2 [38].

6.2.3 Complexity Analysis

We now provide a computational complexity analysis of the proposed scheme in terms of the number of multiplications and additions. For the fading MAC scenario, the pilot detection and channel estimation steps are dominated by matrix multiplication operations with a complexity of $\mathcal{O}(K_a^2 n_p)$. Since we use the $l1$ norm to calculate the decision metrics, our proposed pattern detection method in GMAC requires no multiplications and the number of additions is $\mathcal{O}(M_s n_c)$. The number of multiplications for LLR extraction is $\mathcal{O}(K_a n_c)$, dominating the number of multiplications for the GMAC case. The single-user polar decoder can be implemented using only additions and subtractions [81], with a complexity of $\mathcal{O}(LK_a n_c \log n_c)$ which dominates the number of additions, where L is the list size of the polar decoder. The matrix multiplication in the re-estimation step of the channel coefficients in fading MAC has a complexity of $\mathcal{O}(K_a^2(n_p + n_c))$, dominating the number of multiplications in the case of fading MAC.

6.3 Numerical Results

We assess the system performance by calculating the required E_b/N_0 for a target PUPE of ϵ. For a more straightforward comparison with the existing literature, we take $n = 30000$, $B = 100$, $\epsilon = 0.05$ for Gaussian MAC, and $\epsilon = 0.1$ for the fading case. For the parameters that are specific to our scheme, we set $B_p = 12$ for $K_a = 50$, $B_p = 13$ for $50 < K_a \leq 150$, and $B_p = 14$ otherwise. We employ 5G polar codes, set the CRC length to 16, and utilize a successive cancellation list decoder (SCLD) with a list size of 128. For the pattern matrix, we utilize a randomly generated binary matrix of size $(n - n_p) \times M_s$.

We compare the energy efficiency of the proposed scheme with the state-of-the-art in Figure 6.2 for the fading MAC scenario assuming i.i.d. quasi-static Rayleigh fading channels. We set $n_p = 4000$ for $K_a \leq 500$ and $n_p = 7000$ otherwise, $n_c = 512$, and $n_{\text{OMP}} = 8$. The results in Figure 6.2 illustrate that the newly proposed scheme outperforms the state-of-the-art for the entire regime. For instance, it is superior to the scheme in [51] by up to 5.5 dB for $K_a \leq 600$, and to the one in [52] by up to 0.9 dB for $K_a \leq 700$. Note that for this scenario, we subtract the effect of the successfully decoded sequences within the iteration, too, which improves the performance in the high active user loads.

In the GMAC scenario, we employ unequal power levels among the users similar to [39] to improve the decoding performance further. Namely, the codebook of the transmission patterns $\mathbf{P}$ is equally divided by the number of power levels, and the signal of each user is scaled with a power level based on its transmission pattern index. For instance, in the case of two groups, if the pattern index of a user is less than $M_s/2$, its transmitted signal is scaled by $\sqrt{P_1}$, and by $\sqrt{P_2}$ otherwise, while $P_1 + P_2 = 2P$ to satisfy the energy constraint. Note that the power levels are numerically determined, which are listed in Table 6.1.

In Figure 6.3, we compare the energy efficiency of the proposed scheme with those of the two state-of-the-art solutions in [40] and [44]. We set $n_c = 512$ for $K_a \leq 200$ and $n_c = 256$ otherwise, and employ one power level for $K_a < 150$,

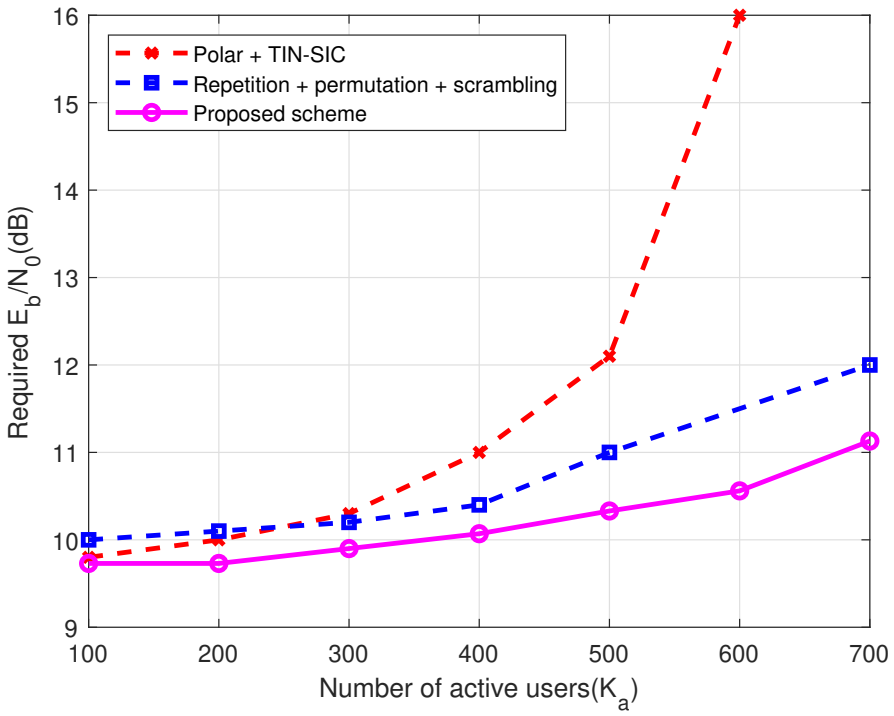

Figure 6.2: Required E_b/N_0 versus number of active users for fading MAC.

two for $150 \leq K_a \leq 200$, and three otherwise. These selections are based on our extensive simulations. The results in Figure 6.3 demonstrate that the proposed scheme performs better than the ODMA one in [44] up to 0.7 dB for low active user loads, and it has a competitive performance when the active user load is high. Also, the proposed scheme outperforms the technique based on random spreading and soft cancellation proposed in [40] in conjunction with a 5G-NR LDPC code [1]. Furthermore, the performance of the proposed scheme is comparable with the random coding achievability bound for $K_a \leq 100$, and the difference for higher active user loads is less than 2 dB.

[1]Note that this comparison is made with the scheme in [40] implemented using 5G-NR LDPC codes. We did not use the results in [40] directly since we did not have access to the specific LDPC code used. Nevertheless, performance of the proposed scheme is competitive even with those provided in [40] with a much lower complexity.

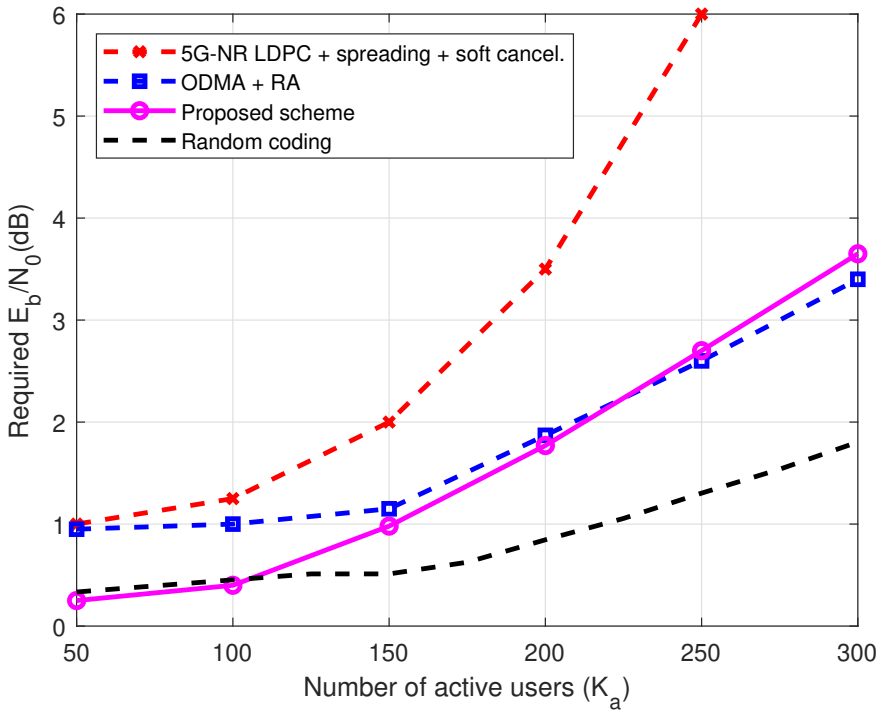

Figure 6.3: Required E_b/N_0 versus number of active users for GMAC.

Table 6.1: Assigned power levels for different active user loads

K_a	50-100	150	200	250	300
P_1	1	0.87	0.73	0.64	0.50
P_2		1.13	1.27	0.92	0.92
P_3				1.44	1.58

Table 6.2: Comparison of the complexity orders

URA scheme	Multiplications	Additions
Proposed	$\mathcal{O}(K_a n_c)$	$\mathcal{O}(LK_a n_c \log n_c)$
ODMA + RA [44]	$\mathcal{O}(M_s n_c)$	$\mathcal{O}(M_s n_c)$
LDPC + spread. [40]	$\mathcal{O}(K_a^4 n_c)$	$\mathcal{O}(K_a^3 n_c n_s)$

The complexity of the ODMA scheme in [44] is dominated by the pattern detection and that of the LDPC coding-based random spreading scheme in [40] is determined by the soft-in soft-out MMSE filtering. We compute the complexity orders of these schemes and our newly proposed scheme in terms of the number of multiplications and additions in Table 6.2, where n_s is the spreading sequence

length in [40]. Note that the additions are much easier to implement in hardware, hence there is a clear complexity advantage for typical system parameters, since $M_s \gg K_a$ and $K_a^4 \gg K_a$, the order of multiplications in the proposed scheme is significantly less than those in [44] and [40].

6.4 Chapter Summary

We examine ODMA in a URA set-up, and develop a simple and energy-efficient solution combining the ideas of exploiting a sparse transmission pattern, blind pattern detection based on the received signal power, and polar coding. Numerical examples illustrate that the newly proposed scheme improves the current state-of-the-art by outperforming the existing works for URA over fading MAC, while achieving a similar performance for the GMAC scenario with a significantly lower complexity.

Chapter 7

Summary and Conclusions

In this dissertation, we develop low complexity solutions for unsourced random access for different scenarios. URA is a recently introduced approach for massive random access, which is important for the next generation communication systems as the expected number of connected devices in a wireless network for some applications will be in the order of millions in B5G and 6G communications. The communication of these devices has a sporadic and uncoordinated nature, hence, the conventional approaches become insufficient to ensure a reliable and energy efficient communication. URA addresses this problem by dictating that all the users share the same codebook, which removes the user identity. Then, the system operation only depends on the number of active users rather than the total number of them, and the receiver aims to produce a list of the transmitted messages up to a permutation. The main objective in URA is to minimize the required energy-per-bit for a given per-user probability of error while supporting a large number of users.

We first propose and study URA over frequency-selective channels, a more practical model than the widely considered flat fading in the URA context. We divide the transmission frame into slots and employ OFDM to mitigate the channel effects, hence, the system operates in the frequency domain. Each active user picks a random slot to transmit its data consists of its pilot sequence and polar

codeword. At the receiver side, we employ an OMP-based joint activity detection and channel estimation algorithm to estimate the channel taps in the time domain followed by TIN-SIC detection to recover the message bits. For comparison purposes, we consider grant-based FDMA and derive its approximate performance limits based on normal approximations. Via numerical examples, we observe that the proposed scheme offers a competitive performance with grant-based FDMA and there is a moderate gap with its approximate performance limits. Furthermore, the performance loss with no coordination with estimated CSI is less than 2 dB when the number of active users is not large.

In another line of investigation, we consider URA with a massive MIMO receiver, i.e., the receiver is equipped with a massive number of antennas. We propose to divide the transmission frame into slots and each slot is divided into pilot and data parts as well. In the pilot part, the active users transmit a non-orthogonal pilot sequence that is selected based on a part of the message bits from a common codebook, and polar coding is adapted as the channel coding approach. At the receiver, we utilize a gOMP algorithm followed by LMMSE to estimate the pilot sequences and channel vectors of users, respectively. We then employ MRC to estimate user symbols, pass these estimates to a single-user polar decoder, and re-estimate the channel vectors for SIC at the end of each iteration. We also analyze the performance of the proposed scheme via normal approximation, provide a detailed complexity analysis, and study the effect of the pilot collisions to the system performance. Numerical examples demonstrate that the proposed scheme has a superior performance compared to the schemes in the literature or has a comparable performance with a lower complexity. Furthermore, it offers an excellent performance in the short blocklength regime and in the case that the receiver is equipped with multiple antennas.

In the existing URA schemes, it is assumed that ideal hardware units are available at both the BS and UE equipment. However, due to the massive number of users, the employed hardware needs to be inexpensive, hence, such impairments are inevitable. Similarly, in massive MIMO systems, cheap antenna elements are also sensitive to hardware impairments. With this motivation, we study URA with a massive MIMO receiver with residual HWIs at both the BS and the UE

sides. Specifically, we adapt our proposed scheme for URA with a massive MIMO receiver in Chapter 4 to this case by deriving a new LMMSE solution using the HWI statistics. We observe that the original scheme designed for the case of no HWIs can still operate with a performance penalty, however, the newly proposed scheme with an HWI-adaptive receiver can increase the energy efficiency and the number of supported active users.

In the last part of the dissertation, we examine a new multiple access technique called ODMA in the URA context. In ODMA, the active users exploit a small part of the transmission resources based on a transmission pattern, leading to a very sparse structure with a low multiuser interference. Assuming that the active users employ polar coding to encode their data, we employ a simple pattern detection approach at the receiver and then recover the message bits using single-user decoding and SIC. Numerical examples reveal that the proposed scheme outperforms the existing schemes in the case of fading MAC, and has a superior or competitive performance with a lower complexity in the Gaussian MAC scenario.

There are several promising extensions of the work items in this thesis that can be considered for future research. In the case of URA over frequency-selective channels, in this thesis, we consider a single-antenna receiver. However, the proposed scheme can also be applied to the massive MIMO scenario that can provide spatial multiplexing gains and some performance enhancement. In addition, different compressed sensing algorithms can be considered for the pilot detection or the scenario of the time-varying channel taps can be studied.

Our work on URA with a massive MIMO receiver also has some interesting extensions. First of all, we assume a quasi-static fading scenario where the channel coefficients remain constant throughout the transmission in this work. However, a block fading scenario that the channel coefficients of the users only remain constant within a slot and change from one slot to another can be considered in conjunction with the repeating user transmissions in multiple slots. Then, the diversity among the slots can be exploited. Namely, if a user signal is successively recovered in one slot, its replica can be subtracted from another slot where it may be transmitted with a weaker channel coefficient. Also, we assume that

the active users are fully-synchronized. However, as a future direction, one can consider an asynchronous MIMO-URA setup where the users start their transmission whenever they have a packet to transmit. This assumption makes the system more practical and leading an easier implementation as it is hard to synchronize a massive number of users. Furthermore, in our proposed scheme and in the pilot-based schemes in the URA context, the pilot and data transmissions are performed separately by dividing the transmission frame or slot into pilot and data parts. However, as an attractive line of research, overlapped pilot and data transmission can be considered. This allows for the usage of longer pilot and data sequences that can provide some performance improvement. On the other hand, the residual pilot interference in the data decoding due to the undetected pilots after pilot detection and channel estimation may have some deleterious effects, which need to be quantified.

In the case of residual HWIs at the BS and the UE sides, we assume that the HWI statistics of all users are the same. However, a more practical extension is to assume different HWI statistics for the users as they can have different impairment levels. Also, we have worked on adapting the system blocks individually to the presence of HWIs, however, a joint optimization of them is also possible.

In our investigation on ODMA for URA, it is assumed that the receiver is equipped with a single antenna and the user transmissions are synchronized. It will be interesting to consider extensions to the MIMO setup and to the case of a fully asynchronous URA scenario using ODMA, namely, allowing the users to start their transmissions at a random instance in time. Moreover, the random signatures proposed in [32] can be exploited in conjunction with ODMA.

Bibliography

[1] Y. Wu, X. Gao, S. Zhou, W. Yang, Y. Polyanskiy and G. Caire, "Massive access for future wireless communication systems," *IEEE Wireless Commun.*, vol. 27, no. 4, pp. 148-156, Aug. 2020.

[2] Y. Polyanskiy, "A perspective on massive random-access," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, June 2017, pp. 2523-2527.

[3] M. Ozates and T. M. Duman, "Unsourced random access over frequency-selective channels," *IEEE Commun. Lett.*, vol. 27, no. 4, pp. 1230-1234, Apr. 2023.

[4] M. Ozates, M. Kazemi and T. M. Duman, "A slotted unsourced random access scheme with a massive MIMO receiver," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, 2022, pp. 2456-2461.

[5] M. Ozates, M. Kazemi and T. M. Duman, "A slotted pilot-based unsourced random access scheme with a multiple-antenna receiver," *IEEE Trans. Wireless Commun.*, early access.

[6] M. Ozates, M. Kazemi and T. M. Duman, "Unsourced random access with hardware impairments," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, 2023, accepted.

[7] A. Benjebbour, Y. Saito, Y. Kishiyama, A. Li, A. Harada and T. Nakamura, "Concept and practical considerations of non-orthogonal multiple access (NOMA) for future radio access," in *Proc. Int. Symp. on Intelligent Signal Process. Commun. Systems*, 2013, pp. 770-774.

[8] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen and L. Hanzo, "A survey of non-orthogonal multiple access for 5G," *IEEE Commun. Surveys & Tutorials*, vol. 20, no. 3, pp. 2294-2323, thirdquarter 2018.

[9] R. Hoshyar, F. P. Wathan and R. Tafazolli, "Novel Low-Density Signature for Synchronous CDMA Systems Over AWGN Channel," *IEEE Trans. Signal Process.*, vol. 56, no. 4, pp. 1616-1626, Apr. 2008.

[10] R. Hoshyar, R. Razavi and M. Al-Imari, "LDS-OFDM an efficient multiple access technique," in *Proc. IEEE 71st Veh. Technol. Conf.*, 2010, pp. 1-5.

[11] H. Nikopour and H. Baligh, "Sparse code multiple access," in *Proc. IEEE 24th Annual Int. Symp. Personal, Indoor, and Mobile Radio Commun. (PIMRC)*, London, UK, 2013, pp. 332-336.

[12] Z. Yuan, G. Yu, W. Li, Y. Yuan, X. Wang and J. Xu, "Multi-user shared access for Internet of Things," in Proc. *2016 IEEE 83rd Vehicular Technol. Conf. (VTC Spring)* Nanjing, China, 2016, pp. 1-5.

[13] M. B. Shahab, R. Abbas, M. Shirvanimoghaddam and S. J. Johnson, "Grant-free non-orthogonal multiple access for IoT: A survey," IEEE Commun. Surveys & Tutorials, vol. 22, no. 3, pp. 1805-1838, thirdquarter 2020.

[14] K. Au et al., "Uplink contention based SCMA for 5G radio access," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Austin, TX, USA, 2014, pp. 900-905.

[15] F. Monsees, C. Bockelmann and A. Dekorsy, "Reliable activity detection for massive machine to machine communication via multiple measurement vector compressed sensing," in *Proc. IEEE Globecom Workshops (GC Workshops)* , Austin, TX, USA, 2014, pp. 1057-1062.

[16] A. T. Abebe and C. G. Kang, "Compressive Sensing-Based Random Access with Multiple-Sequence Spreading for MTC," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, San Diego, CA, USA, 2015, pp. 1-6.

[17] B. Wang, L. Dai, Y. Zhang, T. Mir and J. Li, "Dynamic compressive sensing-based multi-user detection for uplink grant-free NOMA," *IEEE Commun. Lett.*, vol. 20, no. 11, pp. 2320-2323, Nov. 2016.

[18] L. Liu and W. Yu, "Massive connectivity with massive MIMO - Part I: Device activity detection and channel estimation," *IEEE Trans. Signal. Process.*, vol. 66, no. 11, June 2018.

[19] L. G. Roberts, "ALOHA packet system with and without slots and capture," *SIGCOMM Comput. Commun. Rev.*, vol. 5, pp. 28–42, Apr. 1975.

[20] E. Casini, R. De Gaudenzi and O. Del Rio Herrero, "Contention resolution diversity slotted ALOHA (CRDSA): An enhanced random access scheme for satellite access packet networks," *IEEE Trans. Wireless Commun.*, vol. 6, no. 4, pp. 1408-1419, Apr. 2007.

[21] G. Liva, "Graph-based analysis and optimization of contention resolution diversity slotted ALOHA," *IEEE Trans. Commun.*, vol. 59, no. 2, pp. 477-487, Feb. 2011.

[22] E. Paolini, G. Liva and M. Chiani, "Coded slotted ALOHA: A graph-based method for uncoordinated multiple access," *IEEE Trans. Inf. Theory*, vol. 61, no. 12, pp. 6815-6832, Dec. 2015.

[23] U. Demirhan and T. M. Duman, "Irregular repetition slotted ALOHA with energy harvesting nodes," *IEEE Trans. Wireless Commun.*, vol. 18, pp. 4505-4517, September 2019.

[24] T. Akyildiz, U. Demirhan, and T. M. Duman, "Energy harvesting irregular repetition ALOHA with replica concatenation," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 955-968, February 2021.

[25] J. Haghighat and T. M. Duman, "Analysis of coded slotted ALOHA with energy harvesting nodes for perfect and imperfect packet recovery scenarios," *IEEE Trans. Wireless Commun.*, early access.

[26] J. Haghighat and T. M. Duman, "An energy-efficient feedback-aided irregular repetition slotted ALOHA scheme and its asymptotic performance analysis," *IEEE Trans. Wireless Commun.*, early access.

[27] Y. Polyanskiy, H. V. Poor and S. Verdu, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307-2359, May 2010.

[28] E. MolavianJazi and J. N. Laneman, "A second-order achievable rate region for Gaussian multi-access channels via a central limit theorem for Functions," *IEEE Trans. Inf. Theory*, vol. 61, no. 12, pp. 6719-6733, Dec. 2015.

[29] O. Ordentlich and Y. Polyanskiy, "Low complexity schemes for the random access Gaussian channel," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, June 2017, pp. 2528-2532.

[30] A. Vem, K. R. Narayanan, J. Chamberland and J. Cheng, "A user-independent successive interference cancellation based coding scheme for the unsourced random access Gaussian channel," *IEEE Trans. Commun.*, vol. 67, no. 12, pp. 8258-8272, Dec. 2019.

[31] E. Marshakov, G. Balitskiy, K. Andreev and A. Frolov, "A polar code based unsourced random access for the Gaussian MAC," in *Proc. IEEE Veh. Technol. Conf. (VTC)*, 2019, pp. 1-5.

[32] A. K. Tanc and T. M. Duman, "Massive random access with trellis-based codes and random signatures," *IEEE Commun. Lett.*, vol. 25, no. 5, pp. 1496-1499, May 2021.

[33] V. K. Amalladinne, J. -F. Chamberland and K. R. Narayanan, "An enhanced decoding algorithm for coded compressed sensing," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2020, pp. 5270-5274.

[34] V. K. Amalladinne, J. -F. Chamberland and K. R. Narayanan, "A coded compressed sensing scheme for unsourced multiple access," *IEEE Trans. Inf. Theory*, vol. 66, no. 10, pp. 6509-6533, Oct. 2020.

[35] A. Fengler, P. Jung and G. Caire, "SPARCs for unsourced random access," *IEEE Trans. Inf. Theory*, vol. 67, no. 10, pp. 6894-6915, Oct. 2021.

[36] V. K. Amalladinne, A. K. Pradhan, C. Rush, J. -F. Chamberland and K. R. Narayanan, "Unsourced random access with coded compressed sensing: Integrating AMP and belief propagation" *IEEE Trans. Inf. Theory*, vol. 68, no. 4, pp. 2384-2409, April 2022.

[37] H. Cao, J. Xing and S. Liang, "CRC-aided sparse regression codes for unsourced random access," *IEEE Commun. Lett.*, vol. 27, no. 8, pp. 1944-1948, Aug. 2023.

[38] A. K. Pradhan, V. K. Amalladinne, K. R. Narayanan and J. Chamberland, "Polar coding and random spreading for unsourced multiple access," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2020, pp. 1-6.

[39] M. J. Ahmadi and T. M. Duman, "Random spreading for unsourced MAC with power diversity," *IEEE Commun. Lett.*, vol. 25, no. 12, pp. 3995-3999, Dec. 2021.

[40] A. K. Pradhan, V. K. Amalladinne, K. R. Narayanan and J. -F. Chamberland, "LDPC codes with soft interference cancellation for uncoordinated unsourced multiple access," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2021, pp. 1-6.

[41] A. K. Pradhan, V. Amalladinne, A. Vem, K. R. Narayanan and J. -F. Chamberland, "Sparse IDMA: A joint graph-based coding scheme for unsourced random access," *IEEE Trans. Commun.*, vol. 70, no. 11, pp. 7124-7133, Nov. 2022.

[42] Z. Han, X. Yuan, C. Xu, S. Jiang and X. Wang, "Sparse kronecker-product coding for unsourced multiple access," *IEEE Wireless Commun. Lett.*, vol. 10, no. 10, pp. 2274-2278, Oct. 2021.

[43] G. Song, K. Cai, Y. Chi, J. Guo and J. Cheng, "Super-sparse on-off division multiple access: Replacing repetition with idling," *IEEE Trans. Commun.*, vol. 68, no. 4, pp. 2251-2263, Apr. 2020.

[44] J. Yan, G. Song, Y. Li and J. Wang, "ODMA transmission and joint pattern and data recovery for unsourced multiple access," *IEEE Wireless Commun. Lett.*, vol. 12, no. 7, pp. 1224-1228, July 2023.

[45] V. K. Amalladinne, A. Hao, S. Rini and J. -F. Chamberland, "Multi-class unsourced random access via coded demixing," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2021, pp. 3080-3085.

[46] S. Rini, V. K. Amalladinne and J. . -F. Chamberland, "An exploration of the heterogeneous unsourced MAC," in *Proc. 55th Asilomar Conf. Signals, Systems, and Computers*, 2021, pp. 954-958.

[47] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Quasi-static SIMO fading channels at finite blocklength," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2013, pp. 1531–1535.

[48] Y. Polyanskiy, Channel coding: non-asymptotic fundamental limits. Princeton University, 2010.

[49] S. S. Kowshik, K. Andreev, A. Frolov and Y. Polyanskiy, "Energy efficient random access for the quasi-static fading MAC," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)* , 2019, pp. 2768-2772.

[50] S. S. Kowshik, K. Andreev, A. Frolov and Y. Polyanskiy, "Energy efficient coded random access for the wireless uplink," *IEEE Trans. Commun.*, vol. 68, no. 8, pp. 4694-4708, Aug. 2020.

[51] K. Andreev, E. Marshakov and A. Frolov, "A polar code based TIN-SIC scheme for the unsourced random access in the quasi-static fading MAC," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2020, pp. 3019-3024.

[52] E. Nassaji, M. Bashir and D. Truhachev, "Unsourced random access over fading channels via data repetition, permutation, and scrambling," *IEEE Trans. Commun.*, vol. 70, no. 2, pp. 1029-1042, Feb. 2022.

[53] K. Andreev, P. Rybin and A. Frolov, "Coded compressed sensing with list recoverable codes for the unsourced random access," *IEEE Trans. Commun.*, vol. 70, no. 12, pp. 7886-7898, Dec. 2022.

[54] S. S. Kowshik, K. Andreev, A. Frolov and Y. Polyanskiy, "Short-packet low-power coded access for massive MAC," in *Proc. 53rd Asilomar Conf. Signals, Syst., Comput.*, 2019, pp. 827-832.

[55] A. Fengler, S. Haghighatshoar, P. Jung and G. Caire, "Non-Bayesian activity detection, large-scale fading coefficient estimation, and unsourced random access with a massive MIMO receiver," *IEEE Trans. Inf. Theory*, vol. 67, no. 5, pp. 2925-2951, May 2021.

[56] A. Decurninge, I. Land and M. Guillaud, "Tensor-based modulation for unsourced massive random access," *IEEE Wireless Commun. Lett.*, vol. 10, no. 3, pp. 552-556, Mar. 2021.

[57] A. Decurninge, P. Ferrand and M. Guillaud, "Massive random access with tensor-based modulation in the presence of timing offsets," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, 2022, pp. 1061-1066.

[58] Z. Luan, Y. Wu, S. Liang, W. Han, B. Bai and L. Zhang, "Modulation for massive unsourced random access based on tensor block term decomposition," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, 2022, pp. 637-643.

[59] A. Fengler, O. Musa, P. Jung and G. Caire, "Pilot-based unsourced random access with a massive MIMO receiver, interference cancellation, and power control," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 5, pp. 1522-1534, May 2022.

[60] M. J. Ahmadi, M. Kazemi and T. M. Duman, "Unsourced random access using multiple stages of orthogonal pilots: MIMO and single-antenna structures," *IEEE Trans. Wireless Commun.*, early access.

[61] E. Nassaji, R. Soltani, M. Bashir and D. Truhachev, "Spread unsourced random access with an iterative MIMO receiver," *IEEE Commun. Lett.*, vol. 26, no. 10, pp. 2495-2499, Oct. 2022.

[62] M. Gkagkos, K. R. Narayanan, J. -F. Chamberland and C. N. Georghiades, "FASURA: A scheme for quasi-static massive MIMO unsourced random access channels," in *Proc. IEEE 23nd Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, 2022, pp. 1-5.

[63] Z. Han, X. Yuan, C. Xu and X. Wang, "Receiver design for MIMO unsourced random access with SKP coding," *IEEE Wireless Commun. Lett.*, vol. 12, no. 1, pp. 45-49, Jan. 2023.

[64] F. Tian, X. Chen, L. Liu and D. W. Kwan Ng, "Design of massive unsourced random access over Rician channels," in *Proc. IEEE 22nd Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, 2021, pp. 341-345.

[65] X. Xie, Y. Wu, J. Gao and W. Zhang, "Massive unsourced random access for massive MIMO correlated channels," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, 2020, pp. 1-6.

[66] W. Wang, J. You, S. Liang, W. Han and B. Bai, "Slotted concatenated coding scheme for asynchronous uplink unsourced random access with a massive MIMO receiver," in *Proc. IEEE 33rd Annual Int. Symp. Personal, Indoor and Mobile Radio Commun. (PIMRC)*, 2022, pp. 246-252.

[67] J. Gao, Y. Wu, T. Li and W. Zhang, "Energy efficiency of MIMO massive unsourced random access with finite blocklength," *IEEE Wireless Commun. Lett.*, vol. 12, no. 4, pp. 743-747, Apr. 2023.

[68] I. Khan, M. Singh, and D. Singh, "Compressive sensing-based sparsity adaptive channel estimation for 5G massive MIMO systems," *Appl. Sci.*, vol. 8, no. 5, pp. 1–13, May 2018.

[69] S. Li, L. Xiao and T. Jiang, "An efficient matching pursuit based compressive sensing detector for uplink grant-free NOMA," *IEEE Trans. Veh. Technol.*, vol. 70, no. 2, pp. 2012-2017, Feb. 2021.

[70] Y. Polyanskiy and S. Verdú, "Scalar coherent fading channel: Dispersion analysis," in *Proc. IEEE Int. Symp. on Information Theory (ISIT)*, 2011, pp. 2959-2963.

[71] T. Erseghe, "Coding in the finite-blocklength regime: Bounds based on Laplace integrals and their asymptotic approximations," *IEEE Trans. Inf. Theory*, vol. 62, no. 12, pp. 6854-6883, Dec. 2016.

[72] S. Rauh, T. Lauterbach, H. Lieske, J. Robert and A. Heuberger, "Multipath characteristics of indoor-to-outdoor radio channels in the 868-MHz band," in *Proc. Smart SysTech 2016; Eur. Conf. Smart Obj., Syst. Technol.*, 2016, pp. 1-11.

[73] J. Wang, S. Kwon and B. Shim, "Generalized orthogonal matching pursuit," *IEEE Trans. Signal Process.*, vol. 60, no. 12, pp. 6202-6216, Dec. 2012.

[74] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655-4666, Dec. 2007.

[75] W. Yang, G. Durisi, T. Koch and Y. Polyanskiy, "Block-fading channels at finite blocklength," in *ISWCS 2013; The Tenth Int. Symp. Wireless Commun. Systems*, 2013, pp. 1-4.

[76] A. Fengler, P. Jung and G. Caire, "Pilot-based unsourced random access with a massive MIMO receiver in the quasi-static fading regime," in *Proc. IEEE 22nd Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, 2021, pp. 356-360.

[77] G. E. P. Box, J. S. Hunter and W. G. Hunter, " Statistics for experimenters: Design, innovation and discovery," 2nd ed., Wiley, 2005.

[78] E. Björnson, J. Hoydis, M. Kountouris and M. Debbah, "Massive MIMO systems with non-ideal hardware: Energy efficiency, estimation, and capacity limits," *IEEE Trans. Inf. Theory*, vol. 60, no. 11, pp. 7112-7139, Nov. 2014.

[79] U. Gustavsson et al., "On the impact of hardware impairments on massive MIMO," in *Proc. IEEE Globecom Workshops*, 2014, pp. 294-300.

[80] M. Kazemi, A. Mohammadi and T. M. Duman, "Analysis of DF relay selection in massive MIMO systems with hardware impairments," *IEEE Trans. Vehicular Technol.*, vol. 69, no. 6, pp. 6141-6152, June 2020,.

[81] A. Balatsoukas-Stimming, M. B. Parizi and A. Burg, "LLR-based successive cancellation list decoding of polar codes," *IEEE Trans. Signal Process.*, vol. 63, no. 19, pp. 5165-5179, Oct. 2015.