Interference Cancellation Algorithms for Grant-Free Multiple Access With Massive MIMO

Lorenzo Valentini, *Graduate Student Member, IEEE*, Marco Chiani, *Fellow, IEEE*, and Enrico Paolini, *Senior Member, IEEE*

***Abstract*— In next generation Internet-of-Things, the overhead introduced by grant-based multiple access protocols may engulf the access network as a consequence of the unprecedented number of connected devices. Grant-free access protocols are therefore gaining an increasing interest to support massive access from machine-type devices with intermittent activity. In this paper, coded random access (CRA) with massive multiple input multiple output (MIMO) is investigated as a solution to design highly-scalable massive multiple access protocols, taking into account stringent requirements on latency and reliability. With a focus on signal processing aspects at the physical layer and their impact on the overall system performance, critical issues of successive interference cancellation (SIC) over fading channels are first analyzed. Then, SIC algorithms and a scheduler are proposed that can overcome some of the limitations of the current access protocols. The effectiveness of the proposed processing algorithms is validated by Monte Carlo simulation, for different CRA protocols and by comparisons with developed benchmarks.**

***Index Terms*— Coded random access, grant-free access, massive MIMO, massive multiple access, signal processing, successive interference cancellation, 6G.**

I. Introduction

THE rise of the Internet-of-Things (IoT) has progressively furthered attention on machine-type communication (MTC), meant as the autonomous communication between physical objects not directly operated by humans [1]. Due to the fast growing IoT pervasiveness in several application domains, the density of connected objects (in terms of devices per unit area) has recently become so large that the expression *massive* machine-type communication (mMTC) has been introduced [2], [3] to indicate wireless networking among a very large number of devices that are physically located in the same area. Each of these devices, usually battery-driven, generates data discontinuously and intermittently; the device's activity periods are typically used for transmission of one message, in the form of a short packet, to another device or to a remote server through the network. A usual situation is the one where a massive number of devices are connected wirelessly to the network through the same base station (BS). As such, in the uplink a typical massive multiple access (MMA) problem occurs, in which a myriad of transmitters contend to transmit short data packets to the same receiver over the radio access network [4], [5]. Since transmitters wake up intermittently, unpredictably, and independently of each other, the receiver has no a priori knowledge of the number and the subset of simultaneously active ones within a given time interval.

It should be remarked that the MMA setting deviates significantly from the traditional multiple access one. As opposed to the conventional context, where orthogonal channel access is feasible owing to the relatively small number of transmitters and where grant-based protocols with pre-allocation of radio resources are justified by each transmitter typically holding the assigned resources for a while, in MMA applications scheduled access schemes are very inconvenient and inefficient, with control signaling that may even outnumber data. Moreover, from a theoretical point of view, the traditional information-theoretic tools to analyze multi-access communication are insufficient to address fundamental limits of massive access. This problem, that has been known for a long time [6], [7], [8], has recently received a renewed interest [9], [10], [11], [12], [13].

The main challenge for next-generation MMA schemes is represented by the need of featuring a very high scalability, in terms of capability to support the ever-increasing connection densities, in presence of reliability and latency constraints that, although smoother than those characterizing ultra-reliable and low-latency communication (URLLC) services, may be much more tightening than the typical mMTC 5G ones (packet loss probability not exceeding 1% and latency not exceeding 10 s) [14], [15], [16], [17], [18], [19]. To cope with these requirements, next-generation MMA schemes should be designed to maximize the number of *simultaneously active* machine-type devices (or "users"), each contending for transmission of one data packet, for which a target reliability and a target latency can be guaranteed with no a priori knowledge of the users' activation pattern. In this respect, grant-free multiple access schemes, with no pre-existing resource allocation or handshake procedure between the user and the BS, have recently gained an increasing interest owing to their capability to substantially reduce control signalling for connection establishment, which is beneficial in terms of scalability and latency, as well as of protocol lightness and energy efficiency on the device side. Grant-free access schemes, on the other hand, tend to increase complexity at the physical (PHY) layer

Manuscript received 28 November 2022; revised 30 March 2023; accepted 14 May 2023. Date of publication 22 May 2023; date of current version 16 August 2023. This work was supported in part by the CNIT National Laboratory WiLab and the WiLab-Huawei Joint Innovation Center. An earlier version of this paper was presented in part at the IEEE International Conference on Communications, Seoul, South Korea, May 2022 [DOI: 10.1109/ICC45855.2022.9839121]. The associate editor coordinating the review of this article and approving it for publication was J. Wang. *(Corresponding author: Marco Chiani.)*

The authors are with the Department of Electrical, Electronic, and Information Engineering "Guglielmo Marconi" and the CNIT/WiLab, University of Bologna, 40136 Bologna, Italy (e-mail: lorenzo.valentini13@unibo.it; marco.chiani@unibo.it; e.paolini@unibo.it).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCOMM.2023.3277891.

Digital Object Identifier 10.1109/TCOMM.2023.3277891

on the receiver side, due to the need to perform packet detection and also channel estimation directly from the detected packets. Examples of grant-free schemes are the ones recently proposed in [20], [21], [22], [23], [24], [25], [26] and [27]. Typical MMA schemes are also uncoordinated, meaning that simultaneously active devices take actions independently of each other, without any coordination or cooperation.

Uncoordinated protocols based on the coded random access (CRA) paradigm [28], [29], [30], [31], [32], [33], [34], [35], a particular class of grant-free access schemes, ensure high reliability and are currently regarded as candidates for 6G [36] due to their capability of bridging random access with iterative decoding of codes on sparse graphs. Some of these protocol, e.g., contention resolution diversity slotted ALOHA [28] or irregular repetition slotted ALOHA [29], are based on packet repetition; some others, like coded slotted ALOHA (CSA) [30], on packet fragmentation and packet-level coding. Packet replicas or coded fragments are transmitted on different resources, and resource diversity is combined with interference cancellation performed by simple successive interference subtraction at the receiver [30]. As a matter of fact, the performance of CRA schemes does not depend only on the medium access control (MAC) protocol, e.g., the uncoordinated resource selection strategy on the device side; it also heavily relies on the effectiveness of the PHY layer processing algorithms at the receiver. Although part of the literature on CRA tends to model the PHY layer signal processing (including packet detection, channel estimation, and interference cancellation) as ideal, signal processing in a realistic setting may introduce considerable losses with respect to the performance under idealized conditions, especially in terrestrial scenarios characterized by fading. For instance, the often employed collision or multi-packet reception (MPR) channel models [37], [38] may sometimes turn inaccurate, jeopardizing effective system design and optimization [39].

In this paper, we address successive interference cancellation (SIC) algorithms at PHY layer for CRA schemes in massive access applications. We start by critically reviewing a low-complexity SIC algorithm proposed in [21] and tailored to massive multiple input multiple output (MIMO) processing at the receiver; an in-depth analysis for this algorithm is developed and possible vulnerabilities are highlighted. Motivated by this analysis, we then propose an innovative massive MIMO SIC algorithm that is able to considerably improve the number of supported simultaneously active devices for given reliability and latency constraints. The algorithm relies on two main observations. The first one is that, in CRA protocols, it is possible to effectively exploit resource diversity to accurately estimate the channel coefficients in the resources in which interference must be subtracted. The second one is that not all interference cancellation operations are equally effective: Introducing a prioritization to schedule the most effective ones first is expected to improve the overall performance. The key contributions of the paper can be summarized as follows:

- we exploit resource diversity and operation scheduling to improve SIC in presence of a massive number of BS antennas;
- we theoretically analyze the interference effects within a slot, providing system design guidelines;
- we investigate scalability of several PHY and MAC protocol configurations.

This paper is organized as follows. Section II introduces preliminary concepts, the system model, and some background material. Section III describes the proposed SIC technique along with an analysis that justifies the gain introduced by the new scheme. Numerical results are shown in Section IV. Finally, conclusions are drawn in Section V. A subset of the results presented in this work appeared in the conference paper [40]. With respect to [40]: (i) the proposed schemes are addressed in a more thorough way, providing all details and extending the analysis to include noise (besides interference) and to generic M-quadrature amplitude modulation (QAM) constellations; (ii) the concept of cancellation scheduling is introduced to further improve performance; (iii) performance benchmarks are obtained and used as a reference in the numerical results; (iv) richer numerical results are presented.

Notation: Throughout the paper, capital and lowercase bold letters denote matrices and vectors, respectively. The conjugate transposition of a matrix or vector is denoted by $(\cdot)^H$, while $\|\cdot\|$ indicates the Euclidean norm. The operator $\mathbb{E}\{\cdot\}$ denotes expectation, while $\mathbb{V}\{\cdot\}$ is used for variance.

II. Preliminaries and Background

In this section we define the reference scenario, including the channel access protocols and the channel model, also reviewing some PHY layer signal processing techniques performed at the receiver, that will be useful in the sequel.

A. Scenario Definition

We consider an scenario with K single-antenna users (K very large) which aim at transmitting simple uplink messages to one receiving BS equipped with multiple antennas. The BS time is organized into frames, with a periodic beacon signal broadcast by the BS at the beginning of each frame. The frames are divided into N_{s} slots, and users are frame- and slot-synchronous by relying on the beacon signal. The K users wake up unpredictably to transmit data in a frame and therefore they are not all simultaneously active. The number of simultaneously active users, contending for transmission of their packet in the same frame, is denoted by K_{a} and we assume that the receiver has no prior knowledge about this number. The K_{a} active users contend for the channel in a grant-free and uncoordinated fashion to send their uplink data to the BS.

In this paper, we consider the channel access protocols belonging the class of CSA [30]. We focus on the specific case of CSA with repetition codes of a given rate $1/r$ for all users. This means that each active user generates r replicas of its data payload and transmits them in r different slots of the frame. Different strategies for replica placement in the frame have been proposed in [35]. The availability of a BS with a massive number of antenna elements is a key feature to enable MPR at the receiver. In addition to the massive number of BS antennas, MPR is enabled by the use of orthogonal preamble

(or pilot) sequences. In MMA K is typically much larger than the number of available pilots N_{P}, so that pre-assignment of a specific orthogonal pilot to each user is not possible. As a strategy to cope with this issue, each active user picks one pilot randomly from the set of N_{P} available preambles, without any coordination with the other active users. In this setting, if a user is the only one picking a particular pilot in a slot, decoding of the user message in that slot may succeed. Otherwise, under power control decoding fails and the SIC procedure will be in charge of resolving the "collision". The use of CSA-based access and random pilot selection was proposed in [21].

Regarding the channel model, we consider a block Rayleigh fading channel with additive white Gaussian noise (AWGN). The channel coherence time is assumed equal to the slot duration T_{s}, which implies statistical independence of the channel coefficients of the same user across different slots. When coherence times are large, it is possible to subdivide slots into sub-slots as done recently in [27] (where compressed sensing and SIC across sub-slots are used), with the advantage that the user channel remains the same in all sub-slots. In this paper we consider relatively small coherence times and for this reason we stick with the framed and slotted structure. We do not consider shadowing effects owing to the assumption of perfect power control. Coherently with the above-mentioned access protocol and use of orthogonal pilots, each user active in a slot transmits a packet replica composed of one of the N_{P} orthogonal pilot sequences, of length N_{P} symbols, concatenated with a data payload of length N_{D} symbols. Denoting the number of BS antennas by M, the signal received in a slot may be expressed as $[\boldsymbol{P}, \boldsymbol{Y}] \in \mathbb{C}^{M\times(N_{\mathrm{P}}+N_{\mathrm{D}})}$ where

$$\begin{aligned}\boldsymbol{P} &= \sum_{k\in\mathcal{A}} \boldsymbol{h}_k \boldsymbol{s}(k) + \boldsymbol{Z}_p \\ \boldsymbol{Y} &= \sum_{k\in\mathcal{A}} \boldsymbol{h}_k \boldsymbol{x}(k) + \boldsymbol{Z}. \end{aligned} \tag{1}$$

In (1), $\mathcal{A}$ is the set of users transmitting a replica in the considered slot, while $\boldsymbol{h}_k = (h_{k,1}, \ldots, h_{k,M})^T \in \mathbb{C}^{M\times 1}$ is the vector of channel coefficients of the k-th user. The elements of $\boldsymbol{h}_k$ are modeled as zero-mean, circularly symmetric, complex Gaussian independent and identically distributed (i.i.d.) random variables, i.e., $h_{k,i} \sim \mathcal{CN}(0, \sigma_{\mathrm{h}}^2)$ for all $k \in \mathcal{A}$ and $i \in \{1, \ldots, M\}$. Moreover, $\boldsymbol{s}(k) \in \mathbb{C}^{1\times N_{\mathrm{P}}}$ and $\boldsymbol{x}(k) \in \mathbb{C}^{1\times N_{\mathrm{D}}}$ are the orthogonal pilot sequence picked by user k in the current slot and the user's payload, respectively, both with a unitary average energy per symbol. Finally, $\boldsymbol{Z}_p \in \mathbb{C}^{M\times N_{\mathrm{P}}}$ and $\boldsymbol{Z} \in \mathbb{C}^{M\times N_{\mathrm{D}}}$ are matrices whose elements are Gaussian noise samples. The elements of both $\boldsymbol{Z}_p$ and $\boldsymbol{Z}$ are i.i.d. random variables with distribution $\mathcal{CN}(0, \sigma_{\mathrm{n}}^2)$. Due to power control, through the paper we adopt the normalization $\sigma_{\mathrm{h}}^2 = 1$ for all users' channel coefficients.

B. Channel and Payload Estimation

As mentioned above, the BS receives a signal in the form $[\boldsymbol{P}, \boldsymbol{Y}]$ in each slot of the frame. The processing can be split into two phases [21], [35]. In the first one, the BS attempts channel estimation for all possible pilots by computing $\boldsymbol{\phi}_j \in \mathbb{C}^{M\times 1}$, for all $j \in \{1, \ldots, N_{\mathrm{P}}\}$, as

$$\boldsymbol{\phi}_j = \frac{\boldsymbol{P}\,\boldsymbol{s}_j^H}{\|\boldsymbol{s}_j\|^2} = \sum_{k\in\mathcal{A}^j} \boldsymbol{h}_k + \boldsymbol{z}_j \tag{2}$$

where $\mathcal{A}^j$ is the set of active devices employing pilot j in the current slot, $\boldsymbol{s}_j \in \mathbb{C}^{1\times N_{\mathrm{P}}}$ is the j-th pilot sequence, and $\boldsymbol{z}_j \in \mathbb{C}^{M\times 1}$ is a noise vector with i.i.d. $\mathcal{CN}(0, \sigma_{\mathrm{n}}^2/N_{\mathrm{P}})$ entries. Note that in absence of noise, when pilot j is picked by a single user in the current slot, $\boldsymbol{\phi}_j$ equals the vector of channel coefficients for that user.

In the second phase, the BS computes the quantities $\boldsymbol{f}_j \in \mathbb{C}^{1\times N_{\mathrm{D}}}$ and $g_j \in \mathbb{R}$ as

$$\begin{aligned}\boldsymbol{f}_j &= \boldsymbol{\phi}_j^H \boldsymbol{Y} \\ &= \sum_{k\in\mathcal{A}^j} \|\boldsymbol{h}_k\|^2 \boldsymbol{x}(k) + \sum_{k\in\mathcal{A}^j} \sum_{m\in\mathcal{A}\setminus\{k\}} \boldsymbol{h}_k^H \boldsymbol{h}_m \boldsymbol{x}(m) + \tilde{\boldsymbol{z}}_j \end{aligned} \tag{3}$$

and

$$\begin{aligned} g_j &= \|\boldsymbol{\phi}_j\|^2 \\ &= \sum_{k\in\mathcal{A}^j} \left(\|\boldsymbol{h}_k\|^2 + \sum_{m\in\mathcal{A}^j\setminus\{k\}} \boldsymbol{h}_m^H \boldsymbol{h}_k \right) + \tilde{n}_j \end{aligned} \tag{4}$$

where $\tilde{\boldsymbol{z}}_j \in \mathbb{C}^{1\times N_{\mathrm{D}}}$ and $\tilde{n}_j$ are noise terms. Then, the BS attempts estimation of the payload using conventional maximal ratio combining (MRC) as

$$\hat{\boldsymbol{x}} = \frac{\boldsymbol{f}_j}{g_j} = \frac{\boldsymbol{\phi}_j^H \boldsymbol{Y}}{\|\boldsymbol{\phi}_j\|^2}\,. \tag{5}$$

In the case where a generic user ℓ is the only one transmitting with pilot j in a given slot, hereafter referred to as singleton user ($\mathcal{A}^j = \{\ell\}$), we have $\hat{\boldsymbol{x}} \approx \boldsymbol{x}_\ell$. Demapping and decoding operations are performed on $\hat{\boldsymbol{x}}$ and, upon successful channel decoding, the packet symbols are stored in a buffer waiting for the successive interference cancellation phase. The aim of this latter iterative processing, that will be explained in detail in the next section, is to subtract the interference of a packet in a slot using the information retrieved in another slot from one of its replicas. In fact, whenever a packet is successfully decoded, the BS acquires information about the positions of its replicas along with the employed preambles. This can be implemented in several ways, e.g., letting this information be a function of the information bits. This information can be used to cancel interference from a slot and attempt the decoding procedure again. Here, we separately computed $\boldsymbol{f}_j$ and g_j for reasons that will be clear in Section III-A.

III. Analysis of Successive Interference Cancellation Techniques

In this section we present our main contributions. We first review in detail a state-of-the-art SIC technique for CSA with massive MIMO [21], discussing some critical points. Then, we present a theoretical analysis of this technique to assess and investigate the role of interference. Motivated by the carried out analysis, we propose a SIC algorithm to improve the overall CSA performance.

A. Channel Hardening-Based Interference Cancellation

Consider the low-complexity SIC algorithm, here indicated as channel-hardening-based (CHB), proposed in [21] and also recently exploited in [35]. This algorithm relies heavily on channel hardening and favorable propagation effects, which hold when the number of BS antennas, M, is large [41]. Accordingly, in a massive MIMO setting, (3) and (4) can be approximated as

$$\boldsymbol{f}_j \approx \sum_{k\in\mathcal{A}^j} \|\boldsymbol{h}_k\|^2 \boldsymbol{x}(k) + \tilde{\boldsymbol{z}} \tag{6}$$

$$g_j \approx \sum_{k\in\mathcal{A}^j} \|\boldsymbol{h}_k\|^2 + \tilde{n} \tag{7}$$

respectively. In other words, the algorithm relies on assuming that the cross-terms in (3) and (4) (i.e., terms featuring a product $\boldsymbol{h}_k^H \boldsymbol{h}_m$ with $k \neq m$) can be neglected with respect to the main terms. Assume that we initially compute $\boldsymbol{f}_j$ and g_j, $j = 1, \ldots, N_{\mathrm{P}}$, in all slots and that the payload of user ℓ is successfully decoded in a slot. Then, the above approximations lead naturally to the SIC procedure where we update $\boldsymbol{f}_j$ and g_j as $\boldsymbol{f}_j \leftarrow \boldsymbol{f}_j - \|\boldsymbol{h}_\ell\|^2 \boldsymbol{x}(\ell)$ and $g_j \leftarrow g_j - \|\boldsymbol{h}_\ell\|^2$, respectively, in all slots where replicas of the ℓ-th user's payload are present. As such, this SIC algorithm subtracts only the main interfering term from (3) and (4). The update requires knowledge of $\|\boldsymbol{h}_\ell\|^2$ in the replica slots where, due to the block fading assumption, the channel coefficients are different. For this issue, in [21] the authors invoke temporal stability of $\|\boldsymbol{h}_\ell\|^2$ through the whole frame. Here, we simply use the expectation $\mathbb{E}\{\|\boldsymbol{h}_\ell\|^2\} = M$ to perform SIC which is more accurate under block Rayleigh fading assumptions with $\sigma_{\mathrm{h}}^2 = 1$. Hence, the SIC procedure can be described by the updates

$$\boldsymbol{f}_j \leftarrow \boldsymbol{f}_j - M\,\boldsymbol{x}(\ell) \quad \text{and} \quad g_j \leftarrow g_j - M. \tag{8}$$

Before the next section, we want to foreshadow that the approximations (6) and (7) are not very accurate when the cardinality of $\mathcal{A}$ is large. In fact, since for $m \neq k$ we have

$$\begin{aligned} \mathbb{E}\left\{\boldsymbol{h}_k^H \boldsymbol{h}_m\right\} &= 0 \\ \mathbb{V}\left\{\boldsymbol{h}_k^H \boldsymbol{h}_m\right\} &= M \end{aligned} \tag{9}$$

the corresponding interfering terms in (3) and (4) may prevent from decoding a user packet even if it is the only one with a specific pilot. In the following we analyze this phenomenon by evaluating the probability that a user, being the only one with a specific pilot in a slot, is nevertheless not decoded.

B. Theoretical Analysis of the Interference Effects

We use the terminology "logical" to refer to an idealized setting in which: (i) whenever a user is the only one using a pilot in a given slot it is successfully decoded with probability one; (ii) channel estimation is perfect so that interference subtraction is ideal. Hereafter we provide a theoretical analysis of the effects of interference by removing hypotheses (i) and (ii), to understand their impact in a realistic setting.

Let us consider a situation where $|\mathcal{A}|$ users transmit simultaneously in a slot, $|\mathcal{A}^j|$ of them using pilot j. Assume $|\mathcal{A}^j| - 1$ users from the set $\mathcal{A}^j$ have been successfully decoded in other slots of the frame. Then, in the current slot, we can apply CHB interference subtraction which, as mentioned above, mitigates but does not eliminate completely the interference. At this point, there is only one undecoded user adopting the j-th pilot (singleton). To analyze the probability that this user is successfully decoded, we focus on the interfering and noisy terms in (3). Then, from (3) we can write

$$\boldsymbol{f}_j = \sum_{k\in\mathcal{A}^j} \|\boldsymbol{h}_k\|^2 \boldsymbol{x}(k) + \boldsymbol{I}_j \tag{10}$$

where

$$\begin{aligned} \boldsymbol{I}_j = \sum_{k\in\mathcal{A}^j} \sum_{m\in\mathcal{A}\setminus\{k\}} \boldsymbol{h}_k^H \boldsymbol{h}_m \boldsymbol{x}(m) + \sum_{m\in\mathcal{A}} \boldsymbol{z}_j^H \boldsymbol{h}_m \boldsymbol{x}(m) \\ + \sum_{k\in\mathcal{A}^j} \boldsymbol{h}_k^H \boldsymbol{Z} + \sum_{m\in\mathcal{A}} \boldsymbol{z}_j^H \boldsymbol{Z}. \end{aligned} \tag{11}$$

Let us define $\boldsymbol{\xi}_1(k,m) = \boldsymbol{h}_k^H \boldsymbol{h}_m \boldsymbol{x}(m)$. Since $\boldsymbol{h}_k$ and $\boldsymbol{h}_m$ are length-M vectors whose entries are modeled as i.i.d. $\mathcal{CN}(0,1)$ random variables and $\boldsymbol{x}$ is a length-N_{D} payload vector with i.i.d. entries, it follows that each entry $\xi_1(k,m)$ of $\boldsymbol{\xi}_1(k,m)$, $k \neq m$, fulfills

$$\begin{aligned} \mathbb{E}\{\xi_1(k,m)\} &= 0 \\ \mathbb{V}\{\xi_1(k,m)\} &= M. \end{aligned} \tag{12}$$

The second group of terms in (11) can be represented by $\boldsymbol{\xi}_2(m) = \boldsymbol{z}_j^H \boldsymbol{h}_m \boldsymbol{x}(m)$ where $\boldsymbol{z}_j$ is a noise vector with i.i.d. $\mathcal{CN}(0, \sigma_{\mathrm{n}}^2/N_{\mathrm{P}})$ entries. Therefore each entry $\xi_2(m)$ of $\boldsymbol{\xi}_2(m)$ fulfills

$$\begin{aligned} \mathbb{E}\{\xi_2(m)\} &= 0 \\ \mathbb{V}\{\xi_2(m)\} &= \frac{M}{N_{\mathrm{P}}}\,\sigma_{\mathrm{n}}^2. \end{aligned} \tag{13}$$

Similarly, the third group of terms in (11) can be represented by $\boldsymbol{\xi}_3(k) = \boldsymbol{h}_k^H \boldsymbol{Z}$ where $\boldsymbol{Z}$ is a matrix whose elements are i.i.d. $\mathcal{CN}(0, \sigma_{\mathrm{n}}^2)$. Then, each entry $\xi_3(k)$ of $\boldsymbol{\xi}_3(k)$ fulfills

$$\begin{aligned} \mathbb{E}\{\xi_3(k)\} &= 0 \\ \mathbb{V}\{\xi_3(k)\} &= M\,\sigma_{\mathrm{n}}^2. \end{aligned} \tag{14}$$

Finally the last term $\boldsymbol{\xi}_4 = \boldsymbol{z}_j^H \boldsymbol{Z}$ has entries characterized by

$$\begin{aligned} \mathbb{E}\{\xi_4\} &= 0 \\ \mathbb{V}\{\xi_4\} &= \frac{M}{N_{\mathrm{P}}}\,\sigma_{\mathrm{n}}^4. \end{aligned} \tag{15}$$

We can now make the approximation which considers entry independence between $\boldsymbol{\xi}_1(k,m)$, $\boldsymbol{\xi}_2(m)$, $\boldsymbol{\xi}_3(k)$, and $\boldsymbol{\xi}_4$. Under this approximation it follows that the entries I_j of $\boldsymbol{I}_j$ have

$$\begin{aligned} &\mathbb{E}\{I_j\} = 0 \\ &\mathbb{V}\{I_j\} = M\left(|\mathcal{A}^j|\left(|\mathcal{A}| - 1 + \sigma_{\mathrm{n}}^2\right) + \frac{\sigma_{\mathrm{n}}^2}{N_{\mathrm{P}}}\left(|\mathcal{A}| + \sigma_{\mathrm{n}}^2\right)\right). \end{aligned} \tag{16}$$

As mentioned above, let $|\mathcal{A}^j| - 1$ users employing pilot j be decoded in other slots. Performing CHB SIC (8), new residual interfering terms arise. Equation (10) can be rewritten as

$$\boldsymbol{f}_j = \|\boldsymbol{h}_\ell\|^2 \boldsymbol{x}(\ell) + \sum_{k\in\mathcal{A}^j\setminus\{\ell\}} \left(\|\boldsymbol{h}_k\|^2 - M\right) \boldsymbol{x}(k) + \boldsymbol{I}_j$$

$$= \|\boldsymbol{h}_\ell\|^2 \boldsymbol{x}(\ell) + \tilde{\boldsymbol{I}}_j \tag{17}$$

where the subscript ℓ denotes the only remaining user employing pilot j in the slot. Since $\mathbb{E}\{\|\boldsymbol{h}_k\|^2\} = M$ and $\mathbb{V}\{\|\boldsymbol{h}_k\|^2\} = M$, we can incorporate these terms in our approximation, leading to

$$\mathbb{E}\left\{\tilde{I}_j\right\} = 0$$
$$\mathbb{V}\left\{\tilde{I}_j\right\} = M\left(|\mathcal{A}^j|\left(|\mathcal{A}| + \sigma_{\mathrm{n}}^2\right) - 1 + \frac{\sigma_{\mathrm{n}}^2}{N_{\mathrm{P}}}\left(|\mathcal{A}| + \sigma_{\mathrm{n}}^2\right)\right). \tag{18}$$

Due to summation of a large amount of terms, we can approximate $\tilde{I}_j$ as a circularly symmetric complex Gaussian distribution with the mean and variance given in (18). Then, dividing by M we can estimate the payload of user ℓ as

$$\hat{\boldsymbol{x}}(\ell) = \frac{\|\boldsymbol{h}_\ell\|^2}{M}\boldsymbol{x}(\ell) + \frac{\tilde{\boldsymbol{I}}_j}{M}. \tag{19}$$

For a realistic analysis we also consider modulation and channel coding. Employing an M-QAM modulation and hard-decision decoding, the symbol error probability for given $w = \frac{2}{\sigma_h^2}\|\boldsymbol{h}_\ell\|^2$ can be written as [42]

$$P_{\mathrm{e}|w} = A_{\mathsf{M}}\,\mathrm{erfc}\left(\sqrt{\frac{C_{\mathsf{M}}\,w^2}{\mathbb{V}\left\{\tilde{I}_j\right\}}}\right) - \frac{A_{\mathsf{M}}^2}{4}\,\mathrm{erfc}^2\left(\sqrt{\frac{C_{\mathsf{M}}\,w^2}{\mathbb{V}\left\{\tilde{I}_j\right\}}}\right) \tag{20}$$

where $A_{\mathsf{M}} = 2 - 2/\sqrt{\mathsf{M}}$ and $C_{\mathsf{M}} = 3/(8\mathsf{M} - 8)$. Finally, we assume an error correcting code with bounded-distance hard-decision decoding, able to correct up to t errors, and constellation Gray mapping. We can express the probability that decoding of a user packet is unsuccessful given w as

$$P_{\mathrm{fail}|w} \approx 1 - \sum_{d=0}^{t}\binom{N_{\mathrm{D}}}{d}P_{\mathrm{e}|w}^d\left(1 - P_{\mathrm{e}|w}\right)^{N_{\mathrm{D}}-d} \tag{21}$$

where N_D is the number of payload symbols. Equality in (21) would hold if, whenever a symbol is failed, only one of its bits was received in error. In general this is not true, but exploiting Gray mapping this is a well-fitting approximation.

In conclusion, under CHB SIC, the probability that decoding of a user packet is unsuccessful in a slot where its $|\mathcal{A}^j| - 1$ pilot-interferers have been subtracted and a total of $|\mathcal{A}|$ users were initially allocated in the slot can be expressed as

$$P_{\mathrm{fail}} = \int_0^\infty P_{\mathrm{fail}|w}\,\frac{1}{2^M\,\Gamma(M)}\,w^{M-1}\,e^{-w/2}\,dw \tag{22}$$

where $\Gamma(\cdot)$ is the gamma function. This follows from w being chi-squared distributed with $2M$ degrees of freedom ($\sigma_h^2 = 1$). The expression assumes M-QAM constellation with Gray mapping and hard-decision decoding. The expression of $P_{\mathrm{fail}|w}$ in (22) is given by (21), where $P_{\mathrm{e}|w}$ is provided in (20) with the approximation (18). We observe that, to increase the resilience of singleton users to interference in terms of packet error probability, we can increase either the number of BS antennas M or the code error correction capability t for fixed N_{D} (which however decreases the error correcting code rate). In Appendix A, we show how to extend this analysis to general modulation and coding schemes.

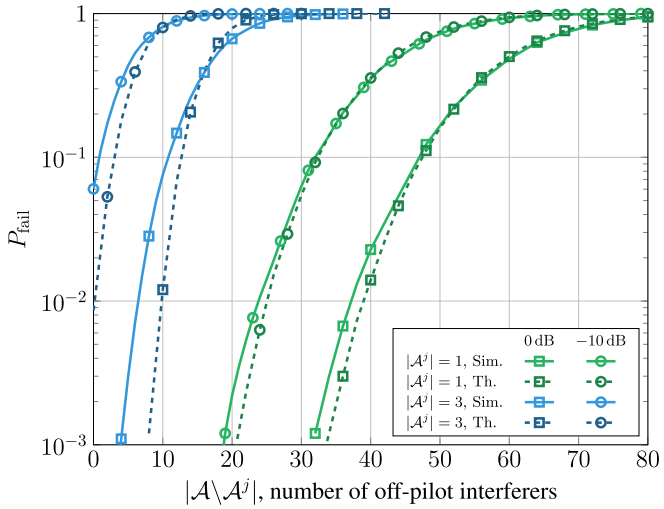

Fig. 1. Probability of decoding failure of a singleton user after $|\mathcal{A}^j| - 1$ CHB interference subtraction operations. Comparison between the analytical approximation (22) and the simulation for $N_{\mathrm{D}} = 256$, $t = 10$, $M = 256$, QPSK constellation, and $\sigma_{\mathrm{n}}^2 \in \{1, 10\}$.

Remark 1: A simple approximation can be made, observing that for large M the probability density function (PDF) narrowed around the mean value and therefore $\mathbb{E}\{f(w)\} \simeq f(\mathbb{E}\{w\})$. This was made in [40] neglecting the noise (i.e., $\sigma_{\mathrm{n}}^2 = 0$).

Example 1: We report in Fig. 1 the analytical approximations derived in (22) in comparison with Monte Carlo simulations for $N_{\mathrm{D}} = 256$, $t = 10$, $M = 256$, quadrature phase-shift keying (QPSK) constellation, and two noise levels $\sigma_{\mathrm{n}}^2 \in \{1, 10\}$. Despite approximations, the analytical results provide a good estimate of the simulated curves also in the presence of noise. In particular, when $|\mathcal{A}^j| = 1$, no interference subtractions are performed and the user experiences the most favorable interference conditions. The $|\mathcal{A}^j| = 1$ curve in Fig. 1 reveals the actual performance of MRC payload estimation in (5) when interferers, using different orthogonal preambles, are captured in the model. Indeed, this is a major non-ideality, degrading the general performance of MAC protocols when a realistic channel model is considered. On the other hand, when $|\mathcal{A}^j| > 1$, the estimation deteriorates even more, revealing the non-ideality of the SIC procedure. Moreover, we point out that, whenever a device using pilot j in the current slot is successfully decoded and CHB is performed, the interference on pilots different from j is not mitigated. This is the critical point of this SIC procedure and in Section III-C we will propose a technique able to overcome this problem.

Remark 2: The analysis conducted in this section, not only provides system design guidelines, but can be relevant to jointly optimize PHY and MAC layer. For example, in [39] an optimization is proposed based on density evolution recursion under realistic channel and PHY layer processing relying on (22).

C. Payload Aided Subtractions

Motivated by the analysis carried out in the previous subsection, we aim at changing the SIC algorithm to make it more

effective and improve the overall performance. In repetition-based CSA, users send multiple copies of the same payload over the frame. Hereafter, we refer to the slots in which a packet is successfully decoded as "generator" slots.

Assume one of the replicas sent by a user, say user ℓ, is successfully decoded in a slot, in correspondence of some pilot $\boldsymbol{s}_j$. The BS available information consists of the user's payload $\boldsymbol{x}(\ell)$, which is common to all replicas, the indexes of the slots where the other replicas have been transmitted, the indexes of the pilots used in each such replica, and the estimate $\boldsymbol{\phi}_j$ of the channel coefficients in the generator slot computed as per (2). The interference subtraction operation in the generator slot is performed as

$$\begin{aligned} \boldsymbol{P}^{(i+1)} &= \boldsymbol{P}^{(i)} - \boldsymbol{\phi}_j \boldsymbol{s}_j \\ \boldsymbol{Y}^{(i+1)} &= \boldsymbol{Y}^{(i)} - \boldsymbol{\phi}_j \boldsymbol{x}(\ell) \end{aligned} \tag{23}$$

where we let $\boldsymbol{P}^{(0)} = \boldsymbol{P}$ and $\boldsymbol{Y}^{(0)} = \boldsymbol{Y}$. As from (23), in the generator slot we do not recompute the channel estimate since the estimation provided by $\boldsymbol{\phi}_j$ is impaired only by noise. Regarding the replica slots, we exploit knowledge of the payload (that is the same in all replicas) to estimate the channel coefficients as

$$\hat{\boldsymbol{h}}_\ell^{(i)} = \frac{\boldsymbol{Y}^{(i)} \boldsymbol{x}(\ell)^H}{\|\boldsymbol{x}(\ell)\|^2} = \boldsymbol{h}_\ell + \tilde{\boldsymbol{h}}_\ell \,. \tag{24}$$

Then, using the "payload-based" channel estimate, in the replica slots we can perform subtraction of interference, similar to (23), as

$$\begin{aligned} \boldsymbol{P}^{(i+1)} &= \boldsymbol{P}^{(i)} - \hat{\boldsymbol{h}}_\ell^{(i)} \boldsymbol{s}(\ell) \\ \boldsymbol{Y}^{(i+1)} &= \boldsymbol{Y}^{(i)} - \hat{\boldsymbol{h}}_\ell^{(i)} \boldsymbol{x}(\ell) \,. \end{aligned} \tag{25}$$

In this SIC algorithm, hereafter referred to as payload-aided-based (PAB), each time an update of the matrices $\boldsymbol{P}$ and $\boldsymbol{Y}$ has been carried out we re-compute (2) and (5) for each pilot in the current slot, to check if any other user can be successfully decoded after interference subtraction. We point out that exploiting the preamble (instead of the payload) to perform channel estimation in slots where we wish to subtract interference may heavily deteriorate the estimation quality due to preamble collisions. For the sake of clarity, we report in Algorithm 1 the base station SIC processing.

Remark 3: In the particular case in which we perform the first subtraction operation in a slot using (25), we have

$$\hat{\boldsymbol{h}}_\ell^{(0)} = \boldsymbol{h}_\ell + \sum_{k \in \mathcal{A} \setminus \{\ell\}} \boldsymbol{h}_k \frac{\boldsymbol{x}(k)\, \boldsymbol{x}(\ell)^H}{\|\boldsymbol{x}(\ell)\|^2} + \boldsymbol{z}_\mathrm{h} \tag{26}$$

where $\boldsymbol{z}_\mathrm{h}$ is the residual noise term. In this specific case, we can derive the statistical properties of the estimation error $\tilde{\boldsymbol{h}}_\ell$, given that the payload symbols are independent among users, as

$$\begin{aligned} \mathbb{E}\left\{\tilde{h}_{\ell,n}\right\} &= 0 \\ \mathbb{V}\left\{\tilde{h}_{\ell,n}\right\} &= \frac{|\mathcal{A}| - 1 + \sigma_n^2}{N_\mathrm{D}} \end{aligned} \tag{27}$$

where $n = 1, \ldots, M$. We observe that, as expected, the accuracy of the channel coefficients estimate improves as the number of payload symbols increases. On the other hand, the channel estimate deteriorates as the number of users transmitting in the slot increases. Among all possible $\tilde{\boldsymbol{h}}_\ell$ obtained running the SIC algorithm, this represents the worst case in terms of estimation accuracy.

Algorithm 1 Base Station SIC Processing Summary

```
while "decoded user buffer is not empty" do
    pop user ℓ information from the buffer;
    for "all replicas of user ℓ" do
        res = slot-pilot pair of the current replica of ℓ;
        if SIC == CHB then
            Use (8) and re-attempt decoding in res;
        else // SIC == PAB
            if "res is the resource where the user was found" then
                Use (23) in the slot of res;
            else
                Re-estimate channel with (24) and use (25);
            Re-attempt decoding for all pilots in the slot of res;
        Update the buffer with new found users if necessary;
```

In general, at step $i = n_\mathrm{up} + n_\mathrm{pa}$ of the PAB SIC algorithm, n_up subtractions using (23) have been performed (due to uncollided pilots), and n_pa ones based on payload-aided channel coefficients estimation as from (25). In this regard, Fig. 2 illustrates the results of a variation of the experiment described in Section III-B for the two SIC techniques, CHB and PAB, and for $|\mathcal{A}^j| = 2$. More specifically, we assume that a fraction $0 \leq p \leq 1$ of users in the set $\mathcal{A} \backslash \mathcal{A}^j$ have been successfully decoded and subtracted. For them, we consider the worst case scenario ($n_\mathrm{up} = 0$) where the SIC is performed using (25). As expected, the PAB performance improves as p increases. On the other hand, CHB is not influenced by p. Then, averaging on p, PAB outperforms CHB also in the worst case scenario. Moreover, as in a real scenario we have $n_\mathrm{up} > 0$, the PAB technique is expected to outperform the CHB one to a larger extent; this is confirmed by the numerical results presented in Section IV.

Remark 4: A variation of the PAB technique can be applied also in case of protocols featuring payload segmentation and packet-level coding applied to the segments [30]. When a sufficient number of segments have been decoded in their slots, packet-level decoding allows reconstructing the missing segments. To subtract the interference generated by the missing segments in their slots, "segment-based" channel estimation can be performed, similar to payload-based one when payloads are replicated.

D. Scheduling of Interference Subtraction Operations

In this section we propose a processing technique that is able to enhance the overall system performance and that can be applied in different MAC and PHY layer configurations. The approach consists of introducing a priority scheduler for

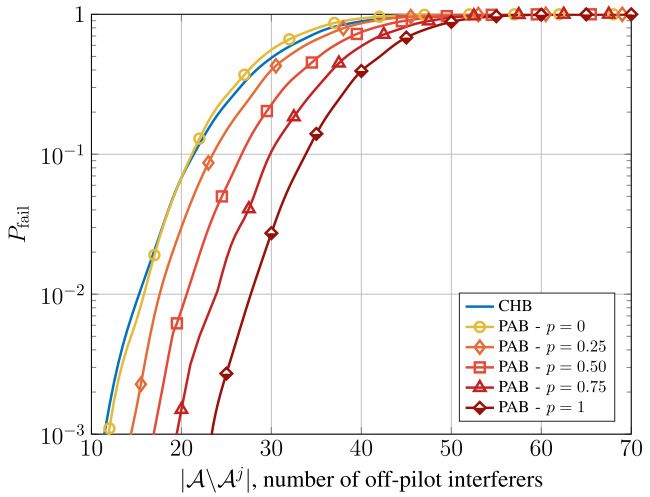

Fig. 2. Probability of decoding failure of a singleton user after one interference subtraction operation ($|\mathcal{A}^j| = 2$). Comparison between CHB and PAB for $N_{\text{D}} = 256$, $t = 10$, $M = 256$, and $\sigma_{\text{n}}^2 = 0$.

interference subtraction operations based on the accuracy of the corresponding channel estimation.

Let us initially focus our attention on PAB schemes. Recalling (2), we see that pilot-based channel estimation is impaired by noise only in case of a singleton user (namely, when $|\mathcal{A}^j| = 1$ for some j). On the other hand, the samples corresponding to replicas of a successfully decoded packet are subtracted from the received matrices $\boldsymbol{P}$ and $\boldsymbol{Y}$ using the payload-based estimation of the channel coefficient according to (25). Since the payloads are not orthogonal with each other, payload-based estimation is impaired by both noise and interference. We can therefore categorize interference subtraction operations based on the accuracy of the channel estimation on which they rely and schedule "high quality" subtractions first. Since in each SIC iteration channel estimations are performed on the current $\boldsymbol{P}$ and $\boldsymbol{Y}$ matrices, as per (2) and (24), it is expected that giving priority to those subtractions that deteriorate these matrices less (in terms of interference residue after the subtraction is performed) helps to increase the number of successful channel decoding operations, triggering further SIC iterations and avoiding a premature stop of the SIC process.

The proposed scheduling of interference subtraction operations, hereafter referred to as "instantaneous cancellation", works as follows. Consider phase 1 of the BS processing. After reception of each block of symbols corresponding to a slot, the BS attempts packet decoding for each pilot by computing and processing $\hat{\boldsymbol{x}}$ in (5). In the baseline scheduling described in literature (e.g., in [21]), all packets successfully decoded in the slot are buffered, awaiting for SIC phase. Then, after all slots have been processed and the SIC phase starts, the subtraction operations are scheduled following the order in which the decoded packets are extracted from the buffer (e.g., according to a first-in-first-out or last-in-first-out policy): for each extracted packet, the samples of all its replicas are subtracted in parallel from the corresponding slots. Instead of this, we propose to perform subtraction operations of singleton users, i.e., high-quality subtraction operations relying on pilot-based channel estimates, immediately after a packet has been decoded in slot and to immediately reprocess

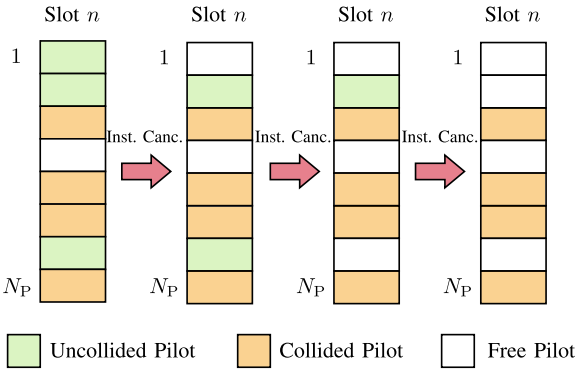

Fig. 3. Pictorial representation of the instantaneous cancellation technique. In the example have been used $N_{\text{P}} = 8$ orthogonal pilots per slot. In green are represented pilots chosen by one user (singleton), in orange the pilots used by two or more users, and in white the unused pilots.

the other pilots in the same slot, iterating the procedure and moving to the next slot only when no new packets can be successfully decoded. The successfully decoded packets are still buffered awaiting for the SIC phase, but the samples of these packets are "instantaneously" subtracted from the slots where they have been decoded.

This provides a second benefit which is exemplified in Fig. 3. In this example, the total number of pilots is $N_{\text{P}} = 8$ and we are processing the generic slot n. There are three singleton users in pilot $p \in \{1, 2, 7\}$, pilot 4 is unused, while the other pilots have been chosen by more than one user. Let pilots be considered in order from 1 to N_{P}, and assume the decoder successfully decodes a packet in correspondence of pilot 1. It performs instantaneous cancellation and re-attempts decoding from pilot 1.[1] Next, assume that when the receiver attempts decoding in pilot 2, a decoding failure occurs. Such an event is consistent with the curves in Fig. 1, for $|\mathcal{A}^j| = 1$, which illustrate that even a singleton user may not been correctly decoded due to interference and noise. Then, let the receiver successfully decode a packet using pilot 7 (decoding failures necessarily occur in pilots from 3 to 6) and immediately subtract the corresponding samples from the slot: Since the receiver restarts again from pilot 1 and now in the slot there is less interference compared to the previous decoding step, it is possible that the packet using pilot 2 is now decoded. Deferring all subtractions to the SIC phase and performing in parallel all subtractions associated with the same packet, the user in pilot 2 could not be found; even if the user was found in another slot, the subtraction in the slot n would be impaired by both noise and interference, deteriorating the overall performance.

This algorithm synergizes effectively also with MAC protocols that foresee a feedback channel used by the BS to broadcast acknowledgement (ACK) messages, e.g., at the end of each slot. This is because, when the scheduling algorithm is applied, a larger number of ACK messages are more likely to be triggered. In general, the instantaneous cancellation technique can be seen as a pre-SIC processing that is performed slot by slot and, as such, can be employed by both CHB

[1]Considering that only a singleton user can be successfully decoded, the procedure can be optimized avoiding to search for packets in pilots where a user has already been found.

and PAB processing schemes. In Section IV we will show the effectiveness of this technique for different choices of the MAC access protocol and PHY layer processing.

E. Complexity Analysis

In this subsection, we discuss the BS processing complexity. Firstly, we carry out a worst case complexity analysis, assuming that no particular strategy aimed at reducing the cost of processing is applied. Possible optimization techniques to lower complexity are pointed out at the end of the subsection. As from Section II-B, we can split the BS processing into two phases, the initialization one (i.e., slot-by-slot processing) and the SIC one. Hereafter, we denote by C_{INIT} the cost of initializing one slot and C_{SIC} the cost of subtracting the inference of one user, such that the total cost is given by $C_{\mathrm{TOT}} = N_{\mathrm{s}}\, C_{\mathrm{INIT}} + K_{\mathrm{a}}\, C_{\mathrm{SIC}}$ (assuming all K_{a} users active in the frame are processed, otherwise the expression is an upper bound).

The typical situation is the one where the cost of channel decoding, here referred to as C_{DEC}, dominates all the other costs involved in (2), (5), (8), (23), (24), and (25), including matrix multiplications, matrix subtractions, and scalar divisions. This is true not only for the here considered algebraic linear block codes with bounded-distance hard-decision decoding, but also for low-density parity-check (LDPC) codes under belief-propagation decoding or polar codes under successive cancellation list decoding. Then, since during the initialization phase we attempt decoding $N_{\mathrm{P}}\beta$ times in each slot, where $\beta = 1$ when instantaneous cancellation is not applied and $\beta = N_{\mathrm{P}}$ (in the worst case) otherwise, we have $C_{\mathrm{INIT}} \approx N_{\mathrm{P}}\, \beta\, C_{\mathrm{DEC}}$. Similarly, during the SIC phase we perform decoding $\gamma\, r \alpha$ times per each interfering user, where $\gamma = 1$ when instantaneous cancellation is not applied and $\gamma \leq 1$ otherwise, and where $\alpha = 1$ for CHB and $\alpha = N_{\mathrm{P}}$ (in the worst case) for PAB. Here, α represents the average number of decoding re-attempts per slot, while $\gamma\, r$ may be regarded as the "effective" number of replicas to be subtracted per user in the SIC phase. This leads us to $C_{\mathrm{SIC}} \approx \gamma\, r\, \alpha\, C_{\mathrm{DEC}}$. We conclude that the total cost may be expressed as $C_{\mathrm{TOT}} \approx (N_{\mathrm{s}} N_{\mathrm{P}}\, \beta + K_{\mathrm{a}}\, \gamma\, r\, \alpha)\, C_{\mathrm{DEC}}$. Comparing the total cost of the low-complexity CHB scheme with that of PAB with instantaneous cancellation (highest complexity) we see that, in this worst case analysis, the increase in complexity is linear by a factor of approximately N_{P}.

Let us finally discuss how, in practice, the cost of PAB with instantaneous cancellation can be significantly reduced. Activity detection techniques (e.g., a simple energy detector) are effective in decreasing the value of α as they allow avoiding to attempt decoding on empty or too crowded slot-pilot pairs. The value of β can be lowered in the same way by avoiding useless decoding attempts in initialization phase. Also very simple (and easy to implement) tricks are effective to substantially reduce complexity. For example, simply avoiding to reprocess pilots where a user has already been found during instantaneous cancellation, it is possible to drop the value of β from N_{P} to $(N_{\mathrm{P}} + 1)/2$. In general, by the means of such optimizations it is possible to reduce α and β to values much smaller than N_{P} (usually between 1 and 5, depending on the traffic, with $N_{\mathrm{P}} = 64$). Lastly, it is very important to point out that the actual processing time is very dependent of the architecture: for example, the operations increasing complexity (i.e., decoding attempts) well-fit parallel computational architectures since are independent of each other.

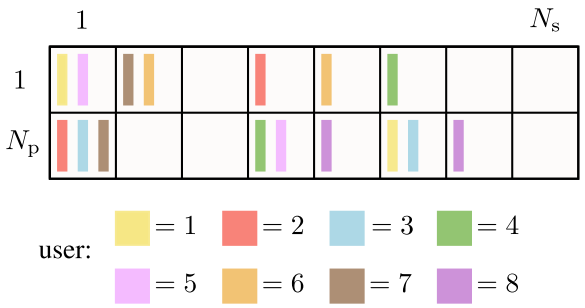

Fig. 4. An example of user replicas allocation in a frame with $N_{\mathrm{s}} = 8$ slots and $N_{\mathrm{P}} = 2$ orthogonal pilots. There are $K_{\mathrm{a}} = 8$ active users, each of them transmitting $r = 2$ packets in the frame.

F. Collision Channel Benchmarks

In this section we introduce some performance benchmarks that will be used in Section IV. These benchmarks are based on a collision channel model over "resources" (slot-pilot pairs), on a collision channel model without SIC, and on a more realistic setting we name perfect replica channel estimation (PRCE), respectively. In addition, we provide the analytical expression for the no-SIC performance, and show that the PRCE benchmark is approachable under specific conditions.

The system performance assuming a collision channel over resources provides an upper bound on the number of simultaneously active users at a target reliability. In this idealized setting: (i) a packet arriving alone in a slot-pilot pair is successfully decoded with probability one (meaning perfect channel estimation and very high signal-to-noise ratio); (ii) interference cancellation in the generator slot and across slots is perfect (meaning perfect channel estimation for the replicas); (iii) no decoding is possible of multiple packets arriving in the same resource (typical in presence of power control). This assumption can be seen as an extension of the classical collision channel over slots. When evaluating numerical results we refer to this benchmark as "logical performance with SIC".

Example 2: In Fig. 4 we provide an example assuming allocation of the users' replicas in a frame with $N_{\mathrm{s}} = 8$ slots and $N_{\mathrm{P}} = 2$ orthogonal pilots. There are $K_{\mathrm{a}} = 8$ active users, each of them transmitting $r = 2$ packets. We use the notation (s, p) to indicate the resource corresponding to slot s and pilot p. Considering collision channel over resources, the messages of users 2, 6, 8, and 4 are successfully decoded in resources $(4, 1)$, $(5, 1)$, $(5, 2)$, and $(6, 1)$ respectively. Note that the message of user 8 is decoded also in resource $(7, 2)$. Then, SIC is performed for all decoded users, leaving user 7 and 5 in $(2, 1)$ and $(4, 2)$ uncollided. Iterating this procedure until no more packets are found, it is easy to verify that all users are retrieved in the order $2, 6, 8, 4, 7, 5, 1, 3$.

As a "worst case" benchmark, we consider also the situation where collision channel over resources model is adopted,

but no SIC procedure is run at the receiver. When replicas are randomly placed in the frame, the performance curve in terms of packet loss probability, given that there are K_a simultaneously active users, can be analytically derived (see Appendix B) as

$$P_{\mathrm{L,\,noSIC}} = \left(1 - \left(1 - \frac{r}{N_\mathrm{s} N_\mathrm{P}}\right)^{K_\mathrm{a}-1}\right)^{r} \tag{28}$$

for N_s slots per frame, N_P orthogonal pilots, and r replicas per user. This analysis allows assessing the improvement on the massive access schemes attributable to the SIC processing. When evaluating numerical results, we refer to this benchmark as "logical performance without SIC".

Example 3: As reported in Example 2 initialization phase, only user 2, 6, 8, and 4 are retrieved referring to Fig. 4. Then, since no SIC algorithm is considered, all the other user messages are lost.

As a third benchmark, we consider a more realistic setting (compared to collision channel assumptions) in which payload estimation is performed as in (5); upon successful message decoding in a slot, PAB processing is applied under the assumption that the subtractions are perfect (ideal SIC). In this setting, referred to as PRCE, the performance is therefore limited by payload estimation (5) only. This establishes a second upper bound on the number of simultaneously active users; this upper bound is generally tighter than the logical performance with SIC one.

Example 4: With reference again to Fig. 4, some of the replicas from users 2, 6, 8, and 4 are singleton ones in the corresponding resources. Under a collision channel over resources model, these replicas would be decoded with probability one. However, since in the PRCE setting payload estimation is realistic and might fail (as it was revealed in the analysis yielding Fig. 1), the process SIC may stop prematurely.

Remark 5: The PRCE performance can be approached, under real channel estimation conditions, when the coherence time of the channel is larger than r times the slot duration and we adopt the access protocol proposed in [35] and called intra-frame spatial coupling (SC). This access strategy consists of letting each active device transmit its replicas in nearby slots, again with a random pilot selection for each replica. In such a setting, in the high signal-to-noise ratio (SNR) regime, the channel estimations of singleton users are almost perfect and, due to block fading channel assumption, the coefficients remain constant for all replicas. This assumption is realistic in all situations in which the slot time is short compared to the coherence time and in which the SC strategy is applied. Making this assumption when no SC protocol is enabled could instead be too optimistic.

IV. Performance Evaluation

In this section, we present numerical results about several PHY layer processing strategies. To make fair comparisons in the context of mMTC with reliability and latency constraints, we impose a common maximum latency and we plot the reliability in terms of packet loss rate P_L against the scalability represented by the number of simultaneously active user per frame K_a. Moreover, we compare the techniques discussed in previous sections with some representative benchmarks, using also different MAC protocols. In particular, we call "baseline MAC" the standard repetition-based CSA protocol with a constant number r of replicas per packet transmitted in r slots chosen uniformly at random in the frame. As a variation of this baseline protocol, we also adopt the recently proposed repetition-based CSA with intra-frame SC and ACK messages [35].

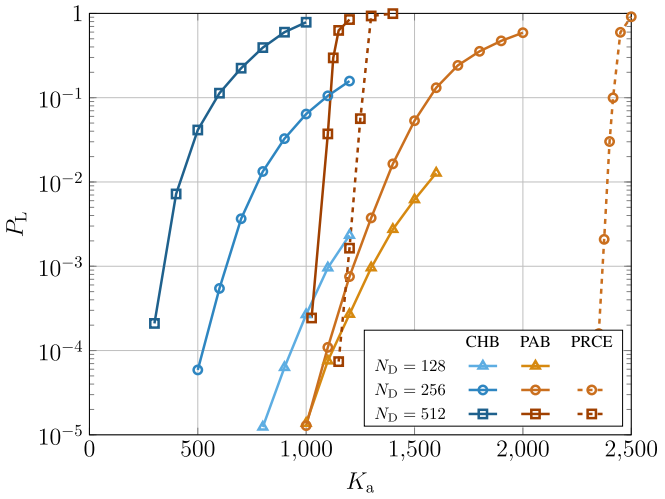

Fig. 5. Packet loss rate values of schemes characterized by different SIC techniques and payload sizes $N_\mathrm{D} = \{128, 256, 512\}$. Baseline MAC with $N_\mathrm{P} = 64$, $N_\mathrm{s} = \{130, 78, 43\}$, and $M = 256$ antennas. Comparison between the CHB, the proposed PAB and the ideal SIC case (PRCE). For the sake of completeness, the PRCE curve at $N_\mathrm{D} = 128$ intersect $P_\mathrm{L}^* = 10^{-3}$ around $K_\mathrm{a} = 4500$.

A. Simulation Setup

We consider a system where users transmit payloads encoded with an (n, k, t) narrow-sense binary Bose–Chaudhuri–Hocquenghem (BCH) code. A cyclic redundancy check (CRC) code is also used to validate decoded packets, avoiding that the SIC procedure adds interference instead of subtracting it. Zero padding the BCH codeword with a final bit, we can map the encoded bits onto a QPSK constellation with Gray mapping, obtaining N_D symbols per codeword. The QPSK symbol energy is normalized to one. Simulations have been carried out with symbol rate $B_\mathrm{s} = 1$ Msps, $M = 256$ BS antennas, $N_\mathrm{P} = 64$ orthogonal pilot sequences, CSA repetition degree $r = 3$, and $\sigma_\mathrm{n}^2 = 0.1$. We impose a maximum latency constraint $\Omega = 50$ ms. For a given maximum latency Ω, the number of slots per frame N_s is equal to [35]

$$N_\mathrm{s} = \left\lfloor \frac{\Omega\, B_\mathrm{s}}{2\,(N_\mathrm{P} + N_\mathrm{D})} \right\rfloor . \tag{29}$$

Note that the length of each orthogonal pilot equals the total number of available pilot sequences N_P. These sequences are constructed using Hadamard matrices. Unless otherwise stated, we will consider that the coherence time is equal to the slot time.

B. Numerical Results

In Fig. 5 we report the packet loss rate (PLR) varying the symbol payload size N_D while keeping the rate of the BCH

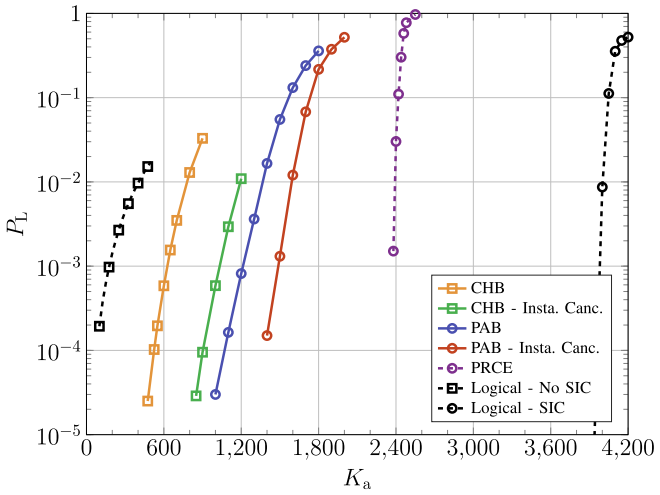

Fig. 6. Packet loss rate comparison between different PHY layer schemes, when a baseline MAC protocol based on CSA using repetition code with $r = 3$ is employed. Maximum latency $\Omega = 50$ ms, $M = 256$ antennas, $N_{\mathrm{P}} = 64$, $N_{\mathrm{s}} = 78$, and $N_{\mathrm{D}} = 256$.

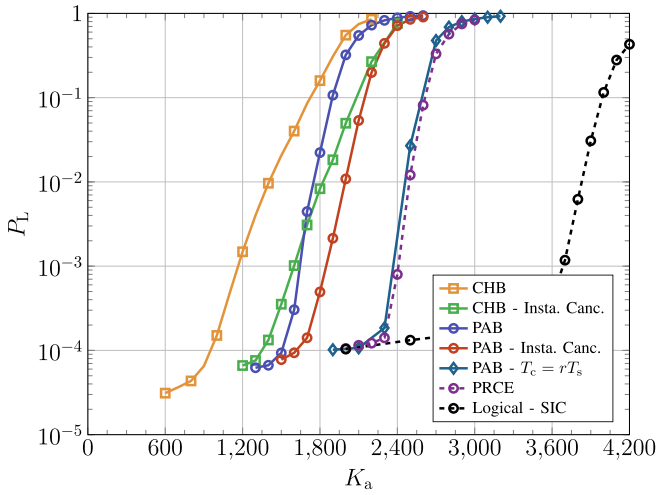

Fig. 7. Packet loss rate comparison between different PHY layer schemes, when intra-frame spatial coupling and ACKs are enabled. CSA using repetition code with $r = 3$ is employed, maximum latency $\Omega = 50$ ms, $M = 256$ antennas, $N_{\mathrm{P}} = 64$, $N_{\mathrm{s}} = 78$, and $N_{\mathrm{D}} = 256$.

code constant, for the CHB, PAB, and PRCE (ideal) interference cancellation. To be precise, for $N_{\mathrm{D}} \in \{128, 256, 512\}$ the corresponding BCH codes are $(255, 207, 6)$, $(511, 421, 10)$, and $(1023, 843, 18)$. In this particular example, we adopt the baseline MAC fixing $N_{\mathrm{P}} = 64$ leading to $N_{\mathrm{s}} \in \{130, 78, 43\}$ in accordance with (29). As expected, the CHB processing curves degrade when N_{D} increases due to the fact that the number of slots per frame N_{s} is decreasing. The same behavior can be observed for PRCE. In the case of PAB processing, instead, the trend is not so obvious. In fact, its performance tends to degrade when N_{s} decreases as for the other schemes, however, a gain in term of SIC quality is also expected from (27). In Fig. 5 we can see the gap between the PRCE and the PAB reduces, highlighting the effectiveness of the proposed technique in a complete scenario which accounts for both the PHY and MAC layers. In this particular example, these two effects counterbalance each other resulting in approximately 1000 active users per frame at $P_{\mathrm{L}} = 10^{-4}$, for all N_{D} under examination using PAB.

In Fig. 6 we plot a comparison between the CHB and PAB SIC techniques, using the baseline MAC protocol. We also apply instantaneous cancellation and plot the relative performance for both methods. The number of payload symbols is set to $N_{\mathrm{D}} = 256$, leading to a $(511, 421, 10)$ BCH code when an information payload of about 50 Bytes is considered. The PAB processing exhibits an improvement compared to the CHB. This is motivated by the fact that PAB subtractions have a beneficial effect on all users transmitting in a slot, while CHB ones influence only the users employing a particular pilot. Enabling instantaneous cancellation we obtain a remarkable performance boost in both SIC algorithms. Targeting for example a PLR $P_{\mathrm{L}} = 10^{-3}$, we see that the logical performance without SIC achieves up to 180 users per frame, the CHB processing increases this number to 650, and PAB with instantaneous cancellation achieves a K_{a} of approximately 1500. This $8\times$ increase in scalability motivates the interest on grant-free CRA schemes under a realistic PHY layer processing.

With reference to the same figure, we also point out the performance gap between a system performing realistic SIC and two idealized schemes, the PRCE and the logical one using SIC. The PAB and PRCE curves rely on the same payload estimation, and for this reason their performance gap depends on channel estimation imperfections. At the same time, there is a remarkable gap between the PRCE curve and the logical one using SIC as a result of payload estimation non-idealities addressed in Section III-B. Comparing the performance of actual schemes with these benchmarks reveals how neglecting the PHY layer processing in real scenarios may lead to wrong conclusions and suboptimum optimizations.

In Fig. 7 we report the performance of the same PHY layer processing techniques of Fig. 6, when the MAC access protocol recently presented in [35] is adopted. In particular, we consider intra-frame spatial coupling packet scheduling, where users are forced to transmit in adjacent slots. In addition, the BS can send ACK messages to notify successfully decoded users at the end of each slot to interrupt useless replica transmissions, resulting in interference attenuation and energy saving [35]. Despite the MAC protocol change, the proposed PHY layer processing techniques provide again a considerable performance improvement.

Let us now discuss how the PRCE performance (i.e., same processing as PAB but with ideal SIC) can be approached using the proposed techniques. As anticipated when discussing Fig. 5, one possibility to reduce the gap between PAB and PRCE is to increase N_{D}. However, since we are considering a scenario where maximum latency is constrained, the degrading effect caused by N_{s} reduction is dominant. Hence, reaching PRCE in this way could not give an overall boost in performance. Another case in which PRCE curve can be reached is depicted in Fig. 7. So far we have considered block fading channel where the coherence time T_{c} is equal to the slot time T_{s}. However, if the time slot is sufficiently small it is possible that, in some scenarios, the coherence time is several times T_{s}. Exploiting the characteristic of intra-frame spatial coupling, we can therefore have the same user channel coefficients among all the replicas ($T_{\mathrm{c}} \geqslant r\,T_{\mathrm{s}}$). Hence, when noise is

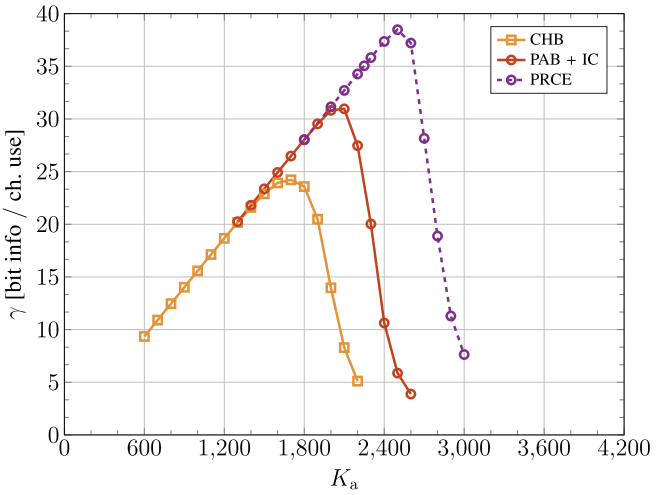

Fig. 8. Sum rates in information bits per channel use of different PHY layer schemes, when intra-frame spatial coupling and ACKs are enabled. CSA using repetition code with $r = 3$ is employed, maximum latency $\Omega = 50$ ms, $M = 256$ antennas, $N_{\rm P} = 64$, $N_{\rm D} = 256$, $N_{\rm s} = 78$, and $N_{\rm extra} = 33$.

sufficiently small, we can subtract interference of all replicas using the channel estimates of singleton users, approaching ideal cancellation performance of PRCE. Despite we are not using the payload information, we report this scheme as PAB with $T_{\rm c} = r\,T_{\rm s}$ because it adopts iterative subtractions in (23).

In Fig. 6 and Fig. 7 we remark the notable gap between PRCE and the logical curve using SIC. This gap is essentially due to the fact that singleton replicas (either the ones that arrived alone in a resource or those becoming singleton ones during the SIC process) are not decoded with probability one and, thus, it is strictly related to Fig. 1. The analytical derivation developed in Section III-B, and in particular the expression of $P_{\rm fail}$ in (22), suggests possible solutions to narrow this gap: for example, we can increase the number of antennas M, or increase the error correction capability t of the channel code (at the cost, however, of reducing the code rate and therefor the sum rate presented next). Some of these solutions are intuitively obvious, but the conducted analysis allows precisely quantifying the effect of a variation of each system parameter. Another important factor which should be considered is the noise level. Nevertheless, since we have used $\sigma_n^2 = 0.1$ in the numerical evaluation, having a smaller noise level does not improve significantly the performance.

In Fig. 8 we show the sum rate in terms of information bits per channel use, defined as

$$\gamma = (1 - P_{\rm L})\, K_{\rm a}\, \frac{N_{\rm D}\, \log_2(\mathsf{M})\, R_{\rm c} - N_{\rm extra}}{N_{\rm s}\,(N_{\rm P} + N_{\rm D})} \tag{30}$$

where $N_{\rm extra} = 33$, $R_{\rm c} = 421/511$, $\mathsf{M} = 4$, and other parameters are the same used in Fig. 6 and Fig. 7. The parameter $N_{\rm extra}$ accounts for payload bits which are not used for information data as CRC and zero padding bits. In particular, we report the sum rates of some schemes using intra-frame spatial coupling packet scheduling with ACKs. In this plot we observe that there exists an optimal $K_{\rm a}$ which maximizes the sum rate γ. However, the values of $K_{\rm a}$ yielding the largest γ may correspond to values of reliability not fulfilling the requirements of next generation MMA systems. On the other hand, the maximum value of the sum rate in information bits per second $\gamma_b = \gamma\, B_{\rm s}$ can be useful to design the backhaul communication network.

V. Conclusion

Interference in grant-free access protocols poses a serious challenge for next generation MMA systems. The use of CRA with multi-packet reception capabilities enabled by massive MIMO and (randomly-chosen) orthogonal pilots improves system scalability, while allowing fulfillment of relatively-tightening latency and reliability constraints. In this paper we showed how, for a given target reliability (i.e., packet loss rate), scalability is heavily reliant on the processing adopted at PHY layer to perform interference subtraction. The main conclusions of this paper can be summarized as: i) the interference cancellation algorithm plays a very significant role in CRA; ii) it is important to efficiently schedule the subtraction operations due to cancellation imperfections; iii) system design and analysis relying on collision-like channels may turn inaccurate. For these reasons, we have proposed an interference cancellation algorithm and a scheduling strategy aiming at improving the overall performance. For example, considering a target packet loss rate $P_{\rm L} = 10^{-3}$ and a requirement on maximum latency of 50 ms, we found out that, employing both techniques, it is possible to achieve a $2.5\times$ scalability gain compared to the state-of-the-art and an $8\times$ gain compared to schemes without SIC.

Appendix A
Interference Analysis on General Modulation and Coding Schemes

To generalize the approach to arbitrary modulation and coding schemes, we observe that (19), for a given $\|\boldsymbol{h}_\ell\|^2$, defines an additive Gaussian channel, with a ratio between the average energy per symbol and the one-sided noise power spectral density given by

$$\frac{E_{\rm s}}{N_0} = \frac{\|\boldsymbol{h}_\ell\|^4}{\mathbb{V}\left\{\tilde{I}_j\right\}} = \frac{w^2}{4\mathbb{V}\left\{\tilde{I}_j\right\}}\,. \tag{31}$$

Thus, it is possible to replace (21) by substituting $P_{{\rm fail}|w}$ with the relation between the codeword error probability and $E_{\rm s}/N_0$ for the modulation and coding scheme of interest. For example, we can use the error probability vs. $E_{\rm s}/N_0$ derived for LDPC or Turbo codes, with QPSK modulation. This generalization can be useful to construct analytical designing tools for CRA schemes in realistic scenario, as done in [39]. Note that, if some activity detection algorithm is employed, the decoder will work with the knowledge of the signal-to-noise ratio $E_{\rm s}/N_0$, as this is related to the actual number of active users, $|\mathcal{A}|$. Otherwise, the decoder should be designed to work sub-optimally, with an unknown signal-to-noise ratio. In most cases the codeword error probability vs. $E_{\rm s}/N_0$ function cannot be found analytically, and Monte Carlo simulation should be used.

Appendix B
Analytical Performance Without SIC

In this appendix we derive the average number of successfully decoded users, assuming a collision channel over

resources model, when no SIC is performed. This analysis can be used as a benchmark to evaluate the effectiveness of the proposed SIC strategy. To keep a clean and compact notation, we denote the probability that a random variable A takes the value a, $\mathbb{P}\{\mathsf{A}=a\}$, as $P(a)$. Similarly, we write $P(a,b\,|\,c)$ to indicate the probability $\mathbb{P}\{\mathsf{A}=a,\mathsf{B}=b\,|\,\mathsf{C}=c\}$, and $\mathbb{P}\{\mathcal{E}\}$ to indicate the probability that an event $\mathcal{E}$ holds.

Let us consider the following problem. There are K_a active devices, each of which transmits r replicas of its packet into a frame composed of N_s slots. The device can put no more than one replica in each slot, and in each slot it can choose between N_P possible orthogonal pilots. Therefore we can describe the frame as a grid of $R = N_\mathrm{s}\cdot N_\mathrm{P}$ resources. Defining as *uncollided* a user, any replica of which has arrived alone in a resource, under a collision channel model the number of successful users in the current frame equals the number of uncollided ones. We can write the total number of uncollided users as

$$\mathsf{X} = \mathsf{X}_1 + \mathsf{X}_2 + \cdots + \mathsf{X}_{K_\mathrm{a}} \tag{32}$$

where

$$\mathsf{X}_i = \begin{cases} 1 & \text{if at least one replica of user } i \text{ is uncollided} \\ 0 & \text{otherwise}\,. \end{cases} \tag{33}$$

The average number of uncollided users can therefore be written as

$$\mathbb{E}\{\mathsf{X}\} = \sum_{i=0}^{K_\mathrm{a}} \mathbb{E}\{\mathsf{X}_i\} = K_\mathrm{a}\cdot\mathbb{P}\{\mathsf{X}_i=1\}\,. \tag{34}$$

Denoting by $\mathcal{U}$ the event that the generic replica transmitted by an active user arrives alone in a resource, we have

$$\mathbb{P}\{\mathsf{X}_i=1\} = 1-(1-\mathbb{P}\{\mathcal{U}\})^r\,. \tag{35}$$

Next, let us focus on a single replica from an active device. Let the considered replica be interfered by J replicas transmitted by other devices that have chosen the same slot. By law of total probability we can write

$$\mathbb{P}\{\mathcal{U}\} = \sum_j \mathbb{P}\{\mathcal{U},j\} = \sum_j \mathbb{P}\{\mathcal{U}|j\}\;P(j) \tag{36}$$

where it is immediate to see that

$$\mathbb{P}\{\mathcal{U}|j\} = \left(\frac{N_\mathrm{P}-1}{N_\mathrm{P}}\right)^j\,. \tag{37}$$

To derive $P(j)$, we firstly write the probability that none of the r replicas is transmitted in a specific slot as

$$\frac{(N_\mathrm{s}-1)\ldots(N_\mathrm{s}-r)}{N_\mathrm{s}\ldots(N_\mathrm{s}-r-1)} = 1-\frac{r}{N_\mathrm{s}}\,. \tag{38}$$

Consequentially, we can derive $P(j)$ as

$$P(j) = \binom{K_\mathrm{a}-1}{j}\left(\frac{r}{N_\mathrm{s}}\right)^j\left(1-\frac{r}{N_\mathrm{s}}\right)^{K_\mathrm{a}-1-j} \tag{39}$$

and conclude that

$$\mathbb{P}\{\mathcal{U}\} = \sum_{j=0}^{K_\mathrm{a}-1}\binom{K_\mathrm{a}-1}{j}\left(\frac{r}{N_\mathrm{s}}\,\frac{N_\mathrm{P}-1}{N_\mathrm{P}}\right)^j\left(1-\frac{r}{N_\mathrm{s}}\right)^{K_\mathrm{a}-1-j}$$

$$= \left(1-\frac{r}{N_\mathrm{s}N_\mathrm{P}}\right)^{K_\mathrm{a}-1}\,. \tag{40}$$

Finally, in absence of SIC the packet loss probability is

$$P_{\mathrm{L,noSIC}} = 1-\frac{\mathbb{E}\{X\}}{K_\mathrm{a}} = \left(1-\left(1-\frac{r}{N_\mathrm{s}N_\mathrm{P}}\right)^{K_\mathrm{a}-1}\right)^r\,. \tag{41}$$

ACKNOWLEDGMENT

The authors wish to thank Alberto Faedi for his work on software simulator implementation.

REFERENCES

[1] J. Sachs et al., "Machine-type communications," in *5G Mobile and Wireless Communications Technology*, A. Osseiran, J. F. Monserrat, and P. Marsch, Eds. Cambridge, U.K.: Cambridge Univ. Press, 2016, ch. 4, pp. 77–106.

[2] H. Shariatmadari et al., "Machine-type communications: Current status and future perspectives toward 5G systems," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 10–17, Sep. 2015.

[3] C. Bockelmann et al., "Massive machine-type communications in 5G: Physical and MAC-layer solutions," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 59–65, Sep. 2016.

[4] S.-Y. Lien, K.-C. Chen, and Y. Lin, "Toward ubiquitous massive accesses in 3GPP machine-to-machine communications," *IEEE Commun. Mag.*, vol. 49, no. 4, pp. 66–74, Apr. 2011.

[5] 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.

[6] J. Wolf, "Coding techniques for multiple access communication channels," in *New Concepts in Multi-User Communication*, J. Skwirzynski, Ed. Alphen an de Rijn, The Netherlands: Sijthoff & Noordhoff, 1981, pp. 83–103.

[7] R. Gallager, "A perspective on multiaccess channels," *IEEE Trans. Inf. Theory*, vol. IT-31, no. 2, pp. 124–142, Mar. 1985.

[8] P. Mathys, "A class of codes for a T active users out of N multiple-access communication system," *IEEE Trans. Inf. Theory*, vol. 36, no. 6, pp. 1206–1219, Nov. 1990.

[9] G. Durisi, T. Koch, and P. Popovski, "Toward massive, ultrareliable, and low-latency wireless communication with short packets," *Proc. IEEE*, vol. 104, no. 9, pp. 1711–1726, Sep. 2016.

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

[11] X. Chen, T. Chen, and D. Guo, "Capacity of Gaussian many-access channels," *IEEE Trans. Inf. Theory*, vol. 63, no. 6, pp. 3516–3539, Jun. 2017.

[12] K. Ngo, A. Lancho, G. Durisi, and A. G. I. Amat, "Massive uncoordinated access with random user activity," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2021, pp. 3014–3019.

[13] E. Paolini, L. Valentini, V. Tralli, and M. Chiani, "Irregular repetition slotted ALOHA in an information-theoretic setting," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2022, pp. 3019–3024.

[14] M. Hasan, E. Hossain, and D. Niyato, "Random access for machine-to-machine communication in LTE-advanced networks: Issues and approaches," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 86–93, Jun. 2013.

[15] L. Liu, E. G. Larsson, W. Yu, P. Popovski, C. Stefanovic, and E. De Carvalho, "Sparse signal processing for grant-free massive connectivity: A future paradigm for random access protocols in the Internet of Things," *IEEE Signal Process. Mag.*, vol. 35, no. 5, pp. 88–99, Sep. 2018.

[16] X. Chen, D. W. K. Ng, W. Yu, E. G. Larsson, N. Al-Dhahir, and R. Schober, "Massive access for 5G and beyond," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 3, pp. 615–637, Mar. 2021.

[17] G. Gui, M. Liu, F. Tang, N. Kato, and F. Adachi, "6G: Opening new horizons for integration of comfort, security, and intelligence," *IEEE Wireless Commun.*, vol. 27, no. 5, pp. 126–132, Oct. 2020.

[18] C. Kalalas and J. Alonso-Zarate, "Massive connectivity in 5G and beyond: Technical enablers for the energy and automotive verticals," in *Proc. 2nd 6G Wireless Summit*, Levi, Finland, Mar. 2020, pp. 1–5.

[19] S. R. Pokhrel, J. Ding, J. Park, O. Park, and J. Choi, "Towards enabling critical mMTC: A review of URLLC within mMTC," *IEEE Access*, vol. 8, pp. 131796–131813, 2020.
[20] 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, pp. 2933–2946, Jun. 2018.
[21] J. H. Sørensen, E. De Carvalho, C. Stefanovic, and P. Popovski, "Coded pilot random access for massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 12, pp. 8035–8046, Dec. 2018.
[22] A. Fengler, S. Haghighatshoar, P. Jung, and G. Caire, "Grant-free massive random access with a massive MIMO receiver," in *Proc. 53rd Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, USA, Nov. 2019, pp. 23–30.
[23] H. Han, Y. Li, W. Zhai, and L. Qian, "A grant-free random access scheme for M2M communication in massive MIMO systems," *IEEE Internet Things J.*, vol. 7, no. 4, pp. 3602–3613, Apr. 2020.
[24] A. T. Abebe and C. G. Kang, "MIMO-based reliable grant-free massive access with QoS differentiation for 5G and beyond," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 3, pp. 773–787, Mar. 2021.
[25] J. Choi, J. Ding, N. Le, and Z. Ding, "Grant-free random access in machine-type communication: Approaches and challenges," *IEEE Wireless Commun.*, vol. 29, no. 1, pp. 151–158, Feb. 2022.
[26] 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.
[27] J. Liu and X. Wang, "Unsourced multiple access based on sparse Tanner graph—Efficient decoding, analysis, and optimization," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 5, pp. 1509–1521, May 2022.
[28] E. Casini, R. De Gaudenzi, and O. Del Rio Herrero, "Contention resolution diversity slotted ALOHA (CRDSA): An enhanced random access schemefor satellite access packet networks," *IEEE Trans. Wireless Commun.*, vol. 6, no. 4, pp. 1408–1419, Apr. 2007.
[29] 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.
[30] 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.
[31] E. Paolini, C. Stefanovic, G. Liva, and P. Popovski, "Coded random access: Applying codes on graphs to design random access protocols," *IEEE Commun. Mag.*, vol. 53, no. 6, pp. 144–150, Jun. 2015.
[32] F. Clazzer, C. Kissling, and M. Marchese, "Enhancing contention resolution ALOHA using combining techniques," *IEEE Trans. Commun.*, vol. 66, no. 6, pp. 2576–2587, Jun. 2018.
[33] M. Berioli, G. Cocco, G. Liva, and A. Munari, "Modern random access protocols," *Found. Trends Netw.*, vol. 10, no. 4, pp. 317–446, 2016.
[34] A. Munari, "Modern random access: An age of information perspective on irregular repetition slotted ALOHA," *IEEE Trans. Commun.*, vol. 69, no. 6, pp. 3572–3585, Jun. 2021.
[35] L. Valentini, M. Chiani, and E. Paolini, "Massive grant-free access with massive MIMO and spatially coupled replicas," *IEEE Trans. Commun.*, vol. 70, no. 11, pp. 7337–7350, Nov. 2022.
[36] N. H. Mahmood, H. Alves, O. A. López, M. Shehab, D. P. M. Osorio, and M. Latva-Aho, "Six key features of machine type communication in 6G," in *Proc. 2nd 6G Wireless Summit (6G SUMMIT)*, Mar. 2020, pp. 1–5.
[37] M. Ghanbarinejad and C. Schlegel, "Irregular repetition slotted ALOHA with multiuser detection," in *Proc. 10th Annu. Conf. Wireless On-Demand Netw. Syst. Services (WONS)*, Mar. 2013.
[38] C. Stefanovic, E. Paolini, and G. Liva, "Asymptotic performance of coded slotted ALOHA with multipacket reception," *IEEE Commun. Lett.*, vol. 22, no. 1, pp. 105–108, Jan. 2018.
[39] L. Valentini, M. Chiani, and E. Paolini, "A joint PHY and MAC layer design for coded random access with massive MIMO," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2022, pp. 2505–2510.
[40] L. Valentini, A. Faedi, M. Chiani, and E. Paolini, "Impact of interference subtraction on grant-free multiple access with massive MIMO," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Seoul, South Korea, May 2022, pp. 1318–1323.
[41] E. Björnson, J. Hoydis, and L. Sanguinetti, "Massive MIMO networks: Spectral, energy, and hardware efficiency," *Found. Trends Signal Process.*, vol. 11, nos. 3–4, pp. 154–655, 2017.
[42] A. Conti, M. Z. Win, and M. Chiani, "Invertible bounds for M-QAM in Rayleigh fading," *IEEE Trans. Wireless Commun.*, vol. 4, no. 5, pp. 1994–2000, Sep. 2005.

Lorenzo Valentini (Graduate Student Member, IEEE) received the B.S. degree (summa cum laude) in electronics engineering and the M.S. degree (summa cum laude) in electronics and telecommunications engineering from the University of Bologna, Italy, in 2017 and 2019, respectively, where he is currently pursuing the Ph.D. degree in electronics, telecommunications and information technologies engineering. From 2019 to 2020, he was with the Interdepartmental Centre for Industrial ICT Research, University of Bologna, working on the Internet of Things. His research interests include communication theory, wireless sensor networks, massive multiple access protocols, and quantum information.

Marco Chiani (Fellow, IEEE) received the Dr.Ing. degree (summa cum laude) in electronic engineering and the Ph.D. degree in electronic and computer engineering from the University of Bologna, Italy, in 1989 and 1993, respectively. Since 2003, he has been a Frequent Visitor with the Massachusetts Institute of Technology (MIT), Cambridge, where he holds a Research Affiliate appointment. He is currently a Full Professor of telecommunications with the University of Bologna. His research interests include information theory, wireless systems, statistical signal processing, and quantum information. He received the 2011 IEEE Communications Society Leonard G. Abraham Prize in the Field of Communications Systems, the 2012 IEEE Communications Society Fred W. Ellersick Prize, and the 2012 IEEE Communications Society Stephen O. Rice Prize in the Field of Communications Theory. He served as the elected Chair (2002–2004) of the Radio Communications Committee of the IEEE Communication Society and as an Editor (2000–2007) for Wireless Communication for the IEEE TRANSACTIONS ON COMMUNICATIONS.

Enrico Paolini (Senior Member, IEEE) received the Dr.Ing. degree (summa cum laude) in telecommunications engineering and the Ph.D. degree in electrical engineering from the University of Bologna, Italy, in 2003 and 2007, respectively. While working towards the Ph.D. degree, he was a Visiting Research Scholar with the Department of Electrical Engineering, University of Hawai'i at Manoa, Honolulu, HI, USA. He was a Visiting Scientist with the Institute of Communications and Navigation, German Aerospace Center, in 2012 and 2014, under DLR-DAAD Fellowships. He is currently an Associate Professor with the Department of Electrical, Electronic, and Information Engineering, University of Bologna. His research interests include digital communication systems, error correcting codes, massive multiple access protocols, and detection and tracking in radar systems. He served as the Co-Chair for the ICC 2014, ICC 2015, and ICC 2016 Workshop on Massive Uncoordinated Access Protocols (MASSAP), the VTC 2019-Fall Workshop on Small Data Networks, the 2018 IEEE European School of Information Theory (ESIT), and the 2020 IEEE Information Theory Workshop (ITW 2020). He served as the TPC Co-Chair for the IEEE GLOBECOM 2022–Communication Theory Symposium and the IEEE GLOBECOM 2019–Communication Theory Symposium. He is the Chair of the ITSoc Italy Section Chapter and the Vice Chair of the IEEE ComSoc Radio Communications Committee. He was an Editor of IEEE COMMUNICATIONS LETTERS from 2012 to 2015 and the IEEE TRANSACTIONS ON COMMUNICATIONS (in coding and information theory) from 2015 to 2020.

Open Access funding provided by 'Alma Mater Studiorum - Università di Bologna' within the CRUI CARE Agreement