Exploiting Pilot Mixtures in Coded Random Access

Exploiting Pilot Mixtures in Coded Random Access

Lorenzo Valentini, *Graduate Student Member, IEEE,* Elena Bernardi, *Graduate Student Member, IEEE,* Enrico Paolini, *Senior Member, IEEE*

***Abstract*—The construction of preamble sequences for channel estimation by superposition of orthogonal pilots can improve performance of massive grant-free uplink from machine-type devices. In this letter, a technique is proposed to obtain full benefit from these "pilot mixtures" in presence of a base station with a massive number of antennas. The proposed technique consists of combining pilot mixtures with an intra-slot successive interference cancellation (SIC) algorithm, referred to as inner SIC, to increase the number of decoded messages per slot. In framed systems, the synergic effect of inner SIC and of an outer SIC algorithm across slots, typical of coded random access protocols, allows achieving a very high reliability with a low number of packet replicas per active user.**

***Index Terms*—Coded random access, grant-free access, massive MIMO, massive machine-type communication, successive interference cancellation.**

I. Introduction

Grant-free multiple access protocols have recently gained an increasing interest for the uplink of next-generation massive machine-type communication (mMTC) applications, where an extremely large number of machine-type devices contend to deliver short packets to a common base station (BS) [1], [2]. By eliminating handshake resource allocation procedures, that are typical of grant-based access, grant-free uplink protocols drastically reduce the amount of control signalling, making channel access very efficient in presence of a massive number of devices with low duty cycle and unpredictable activity. The resulting uplink protocol is very light on the device side, at the cost of an increased complexity in the receiver. Next-generation mMTC use cases will impose more severe reliability and latency requirements, with respect to the ones typical of 5G, with scalability (i.e., the ability to support a very large number of devices) remaining the main key performance indicator [3], [4]. In this context, coded random access (CRA) schemes, combining resource diversity with successive interference cancellation (SIC), are emerging as candidates to achieve different scalability, reliability, and latency tradeoff points in a flexible manner (e.g., [5]–[9]).

To increase the number of decoded packets per slot (i.e., scalability) at given reliability and latency, some form of multi-packet reception (MPR) needs to be exploited. While a naive approach consists of investing in frequency resources, more spectrally efficient ones are indeed possible. An example of them consists of letting machine-type devices share the same time-frequency resources, exploiting massive multiple input multiple output (MIMO) processing at the BS to achieve MPR. Techniques in this area may be roughly classified based on the type of preambles used for channel estimation, namely, orthogonal [8]–[10] versus non-orthogonal [11], [12]. Moreover, a new preamble construction was recently proposed that exploits the superposition of orthogonal pilots [13]. Although the resulting preambles are non-orthogonal with respect to each other, they still allow exploiting orthogonality of the building sequences. It turns out that mixing pilots helps reliability of grant-free access at single slot level. We point out that pilot superposition may also be interpreted as a form of diversity ALOHA [14] obtained through pilots instead of slots.

Repetition based CRA schemes gain in reliability and scalability by letting devices repeat their packets within a frame, in a way similar to diversity ALOHA, and then perform SIC across resources (e.g., time slots) to recover interfered users. This approach can be generalized to packet fragmentation and erasure coding, leading to coded slotted ALOHA [15]. Sticking to repetition-based CRA, it was shown in [16] that, as opposed to simpler (yet very popular) channel models such as the binary adder channel [17], under more realistic channel models and physical layer processing, the best CRA configuration is the one with a constant packet repetition degree 2. From an energy-saving perspective, this result reveals that the optimal scheme in terms of performance is also optimal in terms of energy consumption. Unfortunately, this optimum scheme shows a high error floor in the packet loss rate (PLR) curve which hinders its use at high target reliability values.

Motivated by this result, in this letter we propose a new CRA scheme that effectively exploits pilot mixtures and massive MIMO processing to improve reliability, by means of a nested SIC technique working both across pilots (at slot level) and across slots (at frame level). Remarkably, the proposed technique makes the use of packet repetition degree 2 appealing not only at high PLR values but also at low ones, allowing full exploitation of the most energy efficient CRA configuration. The key contribution of this letter can be summarized as follows:

- A new nested SIC procedure, able of exploiting pilot mixtures in the framework of CRA schemes, is proposed;
- The CRA configurations in which pilot mixtures can provide benefits are discussed;
- An analytical performance bound is derived to benchmark the actual performance.

Notation: 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. Furthermore, we denote the probability that a random variable X takes the value x, $\Pr(\mathsf{X}=x)$, as $P(x)$. Similarly, we write $P(x,y|z)$ in lieu of $\Pr(\mathsf{X}=x,\mathsf{Y}=y\,|\,\mathsf{Z}=$

The authors are with CNIT/WiLab, DEI, University of Bologna, Italy. Email: {lorenzo.valentini13, elena.bernardi14, e.paolini}@unibo.it. Supported by the European Union under the Italian National Recovery and Resilience Plan of NextGenerationEU, partnership on "Telecommunications of the Future" (PE00000001 - "RESTART").

z), and $P(\mathcal{E})$ to denote the probability that an event $\mathcal{E}$ occurs.

II. Preliminaries and Background

In synchronous CRA, the time is organized in frames, each with N_s slots. Each user, say user k, is slot- and frame-synchronous and contends for transmission of one information message. Time synchronization can be achieved exploiting a beacon broadcast by the BS at the beginning of each frame. It forms a data payload $\boldsymbol{x}(k) \in \mathbb{C}^{1\times N_\mathrm{D}}$ out of its message and sends multiple copies of the payload on the frame. Each of these transmitted packets, features its own preamble and a copy of the payload and fits exactly one slot. Each active user draws randomly, independently of the other users: (i) a repetition degree r with probability distribution Λ_r; (ii) r slot indexes, uniformly and without replacement; (iii) a preamble order p, with probability distribution Ψ_p, independently in each chosen slot. Let $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{N_\mathrm{P}}\}$ be a set of N_P orthogonal pilots. Then, in each chosen slot the user builds a "pilot-mixture" preamble sequence combining p pilots in that set uniformly without replacement, similar to [13]. In the generic such slot, the preamble built by user k is

$$\boldsymbol{p}(k) = \frac{1}{\sqrt{p}} \sum_{j=1}^{p} \boldsymbol{s}(k,j)\,, \tag{1}$$

where $\boldsymbol{s}(k,j) \in \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{N_\mathrm{P}}\}$ is the j-th pilot drawn by user k, and the transmitted packet is $[\boldsymbol{p}(k), \boldsymbol{x}(k)]$. Hereafter, we adopt the probability generating function notation $\Lambda(x) = \sum_r \Lambda_r x^r$ and, similarly, $\Psi(x) = \sum_p \Psi_p x^p$ [5]. For concentrated distributions, we simply use the parameter r or p, e.g., $r = 2$ stands for $\Lambda(x) = x^2$.

Packets are sent over a Rayleigh block fading channel with coherence time equal to the slot time. Perfect power control is assumed. At the receiver side, the BS features a massive number of antennas, M. The signal received by the BS in a slot may be expressed as $[\boldsymbol{P}, \boldsymbol{Y}] \in \mathbb{C}^{M\times(N_\mathrm{P}+N_\mathrm{D})}$ where

$$\boldsymbol{P} = \sum_{k\in\mathcal{A}} \boldsymbol{h}_k \boldsymbol{p}(k) + \boldsymbol{Z}_p \quad \text{and} \quad \boldsymbol{Y} = \sum_{k\in\mathcal{A}} \boldsymbol{h}_k \boldsymbol{x}(k) + \boldsymbol{Z}. \tag{2}$$

In (2): $\mathcal{A}$ is the set of users with a packet in the slot under analysis; $\boldsymbol{h}_k = (h_{k,1}, \ldots, h_{k,M})^T \in \mathbb{C}^{M\times 1}$ is the channel coefficient vector for user k in the slot, whose elements are independent and identically distributed (i.i.d.) random variables with distribution $\mathcal{CN}(0, \sigma_\mathrm{h}^2)$ for all $k \in \mathcal{A}$; $\boldsymbol{Z}_p \in \mathbb{C}^{M\times N_\mathrm{P}}$ and $\boldsymbol{Z} \in \mathbb{C}^{M\times N_\mathrm{D}}$ are matrices of Gaussian noise samples. Due to power control, σ_h^2 is the same for all users and without loss of generality we can assume $\sigma_\mathrm{h}^2 = 1$.

At the BS, MPR capability can be obtained exploiting massive MIMO and the set of orthogonal pilots, $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{N_\mathrm{P}}\}$, similar to [8]. In each slot, the BS attempts channel and payload estimation for all pilots $\boldsymbol{s}_j$, $j \in \{1, \ldots, N_\mathrm{P}\}$, by computing the vectors $\boldsymbol{\phi}_j \in \mathbb{C}^{M\times 1}$ and $\hat{\boldsymbol{x}}_j \in \mathbb{C}^{1\times N_\mathrm{D}}$ as

$$\boldsymbol{\phi}_j = \frac{\boldsymbol{P}\,\boldsymbol{s}_j^H}{\|\boldsymbol{s}_j\|^2} = \sum_{k\in\mathcal{A}^j} \boldsymbol{h}_k + \boldsymbol{z}_j \quad \text{and} \quad \hat{\boldsymbol{x}}_j = \frac{\boldsymbol{\phi}_j^H \boldsymbol{Y}}{\|\boldsymbol{\phi}_j\|^2}\,, \tag{3}$$

where $\mathcal{A}^j \subseteq \mathcal{A}$ is the subset of users employing pilot j in the current slot and $\boldsymbol{z}_j \in \mathbb{C}^{M\times 1}$ is a noise vector. Due to pilot orthogonality, $\boldsymbol{\phi}_j$ is a maximum likelihood estimate of $\boldsymbol{h}_k$ whenever, in the considered slot, $\boldsymbol{s}_j$ appears only in the pilot-mixture preamble of user k. Moreover, due to statistical quasi-orthogonality of the channel vectors (also known as favorable propagation) we have $\hat{\boldsymbol{x}}_j \simeq \boldsymbol{x}(k)$ for large M, which makes decoding of $\boldsymbol{x}(k)$ possible [18].

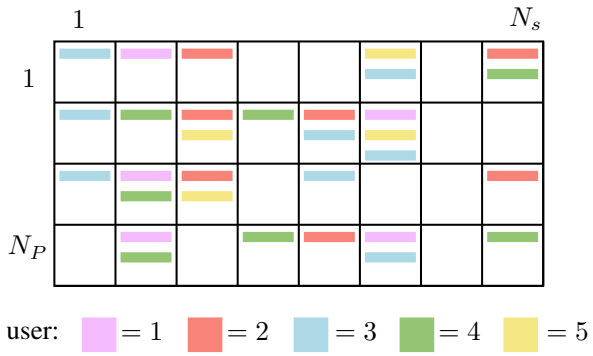

Fig. 1. Pictorial representation of the chosen pilots in a frame. In the example, the frame is composed by $N_\mathrm{s} = 8$ slots; $N_\mathrm{P} = 4$ orthogonal pilots are available to 5 active devices, with $\Lambda(x) = 0.5x^2 + 0.5x^3$ and $\Psi(x) = 0.5x^2 + 0.5x^3$, meaning that each user has 50% probability to pick 2 or 3 slots, and for each slot has a 50% probability to choose 2 or 3 pilots.

III. Coded Random Access with Pilot Mixtures

The access scheme described in Section II can be pictorially represented by a grid, as in Fig. 1. Orthogonality among packets in different slots (or columns) is guaranteed by time division, while packets placed in the same slot but using different pilots (rows) are separable, as per the above processing, only when the total number of active users in the slot, K_s, is small compared to M. It should be noted that slot diversity implies an energy cost, while pilot diversity comes at no extra transmit energy due to the normalization in (1). For clarity, in Fig. 1, placing packets in different column cells has not the meaning of repeating packets, but it means that those pilots were picked by the same user. The interpretation as a repetition, however, is a useful analogy from a CRA perspective.

To get full benefit from the resources offered by the frame grid, we propose a *nested* SIC mechanism. We aim at obtaining a performance gain by combining diversity and SIC, not only at a frame level across different slots, but also at a slot level by means of pilot diversity achieved by (1) exploiting massive MIMO. Hereafter, we refer to this latter intra-slot SIC as "inner SIC". With reference again to Fig. 1, this means performing SIC both across columns, as conventionally done in CRA, and across rows, enabled by pilot mixtures. In every slot, the BS processes the received samples (2), by performing (3) for all $j \in \{1, \ldots, N_\mathrm{P}\}$. Successful channel decoding on $\hat{\boldsymbol{x}}_j$, for any j, triggers the inner SIC procedure in the current slot. The whole packet is therefore reconstructed using the information provided by the payload[1], and its interference is subtracted from the signal samples as

$$\boldsymbol{P}^{(i+1)} = \boldsymbol{P}^{(i)} - \boldsymbol{\phi}_j \boldsymbol{p} \quad \text{and} \quad \boldsymbol{Y}^{(i+1)} = \boldsymbol{Y}^{(i)} - \boldsymbol{\phi}_j \boldsymbol{x}\,, \tag{4}$$

[1]Using the payload bits as a seed of a random number generator (shared between users and BS), the receiver can trace back all choices the user has made for preamble construction and packet frame placements.

Algorithm 1: Slot-by-slot processing with inner SIC

forall $n \in \{1, \dots, N_\mathrm{s}\}$ **do**
  forall $j \in \{1, \dots, N_\mathrm{P}\}$ **do**
    Channel Estimation: compute $\boldsymbol{\phi}_j$ as in (3);
    Payload Estimation: compute $\hat{\boldsymbol{x}}_j$ as in (3);
    if $\hat{\boldsymbol{x}}_j$ *contains a valid packet* **then**
      Reconstruct the packet preamble $\boldsymbol{p}$;
      Inner SIC: update $\boldsymbol{P}$ and $\boldsymbol{Y}$ as in (4);
      Store the packet for **outer SIC** processing;
      Set $j = 1$ to restart the slot computation;

where $\boldsymbol{p}$ and $\boldsymbol{x}$ are the reconstructed preamble and payload of the successfully decoded user. The index i indicates the number of SIC operations done in the current slot, with initial conditions $\boldsymbol{P}^{(0)} = \boldsymbol{P}$ and $\boldsymbol{Y}^{(0)} = \boldsymbol{Y}$. Note that (4) relies on the channel estimate $\boldsymbol{\phi}_j$, since $\boldsymbol{\phi}_j$ is not interfered by other users due to orthogonality. It is thus not necessary to re-estimate the channel coefficients in this phase. After any interference subtraction operation, the cleaned signal samples are processed again by (3) and possibly (4), until no more packets can be decoded. Information of any decoded packet is stored in a buffer for outer SIC processing. The proposed slot-by-slot processing is summarized in Algorithm 1. To reduce complexity, activity detection procedures can be adopted to process only slot-pilot pairs where users may be retrieved.

At the end of the frame, once the BS has concluded the inner SIC phase, the receiver enters the outer SIC phase. This is the typical CRA SIC across slots [15]. Here, buffered decoded packets are processed in order. For each such packet, the contribution of replica interference is subtracted from the corresponding slots and decoding is reattempted in these slots for all pilots. Since in block fading channel the channel gains vary in each slot, it is necessary to estimate the channel for each replica to be subtracted. Solutions to this issue have been proposed, e.g., in [8] and [18]. We consider the payload aided based algorithm [18], estimating the channel as

$$\widehat{\boldsymbol{h}}^{(i)} = \frac{\boldsymbol{Y}^{(i)}\,\boldsymbol{x}^H}{\|\boldsymbol{x}\|^2} \tag{5}$$

and subtracting the interference on the stored slot symbols as

$$\boldsymbol{P}^{(i+1)} = \boldsymbol{P}^{(i)} - \widehat{\boldsymbol{h}}^{(i)}\boldsymbol{p} \quad \text{and} \quad \boldsymbol{Y}^{(i+1)} = \boldsymbol{Y}^{(i)} - \widehat{\boldsymbol{h}}^{(i)}\boldsymbol{x}\,, \tag{6}$$

where $\boldsymbol{p}$ and $\boldsymbol{x}$ are the preamble and payload of the buffered packet. Whenever a new packet is successfully decoded, its information is pushed in the buffer. The outer SIC phase is iterated until the buffer becomes empty.

A. Benefits of Exploiting Pilot Mixtures

To describe the benefits of the proposed technique, let us firstly analyze an intrinsic source of error in CRA schemes with power control. In this regime, an unresolvable collision occurs when two or more users pick the same pilot mixtures in the same slots. Focusing only on this source of error allows us computing a lower bound on the PLR, P_L, useful for parameter design when inner SIC is employed. Restricting our analysis on concentrated distributions $\Lambda(x) = x^r$ and $\Psi(x) = x^p$, we have that the probability that at least two users are involved in such a collision event is

$$P_\mathrm{u} = 1 - \prod_{i=0}^{N-1} \frac{C-i}{C}\,, \tag{7}$$

where C is the total number of possible choices that users can make and N is the number of active users. The exact meaning of N depends on the context as explained next. In a slotted but unframed system we have $C = \binom{N_\mathrm{P}}{p}$ and $N = K_\mathrm{s}$, where K_s is the number of active devices per slot. Moreover, in a framed and slotted system with nested SIC we have $C = \binom{N_\mathrm{s}}{r}\binom{N_\mathrm{P}}{p}^r$ and $N = K_\mathrm{a}$, where K_a is the number of active devices per frame. Finally, we can lower bound the PLR as $P_\mathrm{L} \geq 2\,P_\mathrm{u}\,/\,N$, since at least two users out of N are involved in the unresolvable collision. This lower bound results tight for small K_s (or K_a), i.e., in the low traffic regime.

The combination of pilot mixtures with SIC can bring several benefits to the access protocol. The first advantage this technique bears, consists of a reduced error floor. In fact, in [16] it was shown that, differently to what was proven under collision channel, the best performance is obtained with $\Lambda(x) = x^2$, or equivalently for $r = 2$. However, for $p = 1$ this distribution exhibits high error floors which limit its adoption. In contrast, for $p > 1$ (pilot mixtures), consistently with (7), the error floor appears at remarkably lower PLR values. A second advantage is that, adopting the sole inner SIC, the scheme is still able to achieve a remarkable performance as it will be highlighted in Section IV. This can be appealing if we need to process data only in real-time and without storing the whole frame at the BS. This also reduces the processing burden the BS should manage. A third advantage can be achieved in presence of acknowledgement (ACK) messages at the end of each slot as proposed in [9]. The usage of ACK messages permits to save user energy due to transmission interruption. In this case, the inner SIC used in presence of pilot mixtures is able to retrieve more users per slot compared to conventional schemes using $p = 1$. Moreover, more ACKs imply less interference in the frame, which also improves performance.

B. Benchmarks without Successive Interference Cancellation

Let us start with a slotted but unframed scheme where users exploit pilot mixtures but no inner SIC is performed at the receiver. The performance depends essentially on two factors: i) the probability of "singleton" packets, in which at least one pilot of the mixture is chosen by only one user; ii) The successful decoding probability of singleton packets. We can see that i) depends on the access protocol, while ii) is strictly related to the physical (PHY) layer and channel model. For example, in Fig. 1, no user 5 packets are singleton ones, while user 1 has a singleton in slot 2. Assuming that a singleton always triggers a successful packet decoding, given K_s users per slot, the PLR is tightly approximated as

$$P_\mathrm{L}^{(\mathrm{S})} \approx \sum_p \left(1 - \left(1 - \frac{\Psi'(1)}{N_\mathrm{P}}\right)^{K_\mathrm{s}-1}\right)^p \Psi_p\,. \tag{8}$$

To derive (8), let us focus on an active user with preamble order p in the slot. Decoding of the user's packet fails whenever all pilots in its mixture are picked by other users. The probability that any pilot in the mixture is chosen by another device with a preamble order p is $\binom{N_\mathrm{P}-1}{p-1}/\binom{N_\mathrm{P}}{p} = p/N_\mathrm{P}$. This probability, averaged over the preamble order of the interferer, becomes $\Psi'(1)/N_\mathrm{P}$. Then, the probability that one of the chosen pilots is also picked by at least one interferer is $1-(1-\Psi'(1)/N_\mathrm{P})^{K_\mathrm{s}-1}$. Treating the collision events for all p pilots in the mixture as independent, and averaging with respect to the preamble order distribution leads to the approximation in (8). Note that (8) holds with equality for any concentrated distribution $\Psi(x)=x^p$; for irregular distributions it matches very tightly the actual simulated results, with negligible deviations only for small K_s[2]. In Section IV we will show the beneficial impact of the inner SIC on the overall performance, which extends the range of load values for which a small PLR can be attained.

The above reasoning can be extended to a slotted and framed scheme where users exploit slot diversity and pilot mixtures, but no form of SIC (neither inner nor outer) is performed at the receiver. Given that there are K_a devices active on the frame and assuming a user message is not successfully received when all pilots in its mixture are interfered in all r chosen slots, we obtain the PLR approximation

$$P_\mathrm{L}^{(\mathrm{F})} \approx \sum_r \Lambda_r \left[\sum_p \Psi_p \sum_{j=0}^{p} (-1)^j \binom{p}{j} \left(1 - \frac{\Lambda'(1)}{N_\mathrm{s}} + \frac{\Lambda'(1)}{N_\mathrm{s}} \left(1 - \frac{\Psi'(1)}{N_\mathrm{P}}\right)^j\right)^{K_\mathrm{a}-1}\right]^r . \quad (9)$$

In fact, reasoning as before, the PLR can be approximated as $\sum_r (1-P(\mathcal{U}))^r \Lambda_r$, where $P(\mathcal{U})$ is the probability that a replica has at least one singleton pilot. Using (8), we have that the probability of the event $\mathcal{U}$ given that $K_\mathrm{I} = K_\mathrm{s}-1$ interfering users are present in the slot is $P(\mathcal{U}|K_\mathrm{I}) = 1 - P_\mathrm{L}^{(\mathrm{S})}$, while K_I is binomial distributed with success probability $\Lambda'(1)/N_\mathrm{s}$ and $K_\mathrm{a}-1$ trials. Finally, by the law of total probability we have that $P(\mathcal{U}) = \sum_{K_\mathrm{I}=0}^{K_\mathrm{a}-1} P(\mathcal{U}|K_\mathrm{s})\, P(K_\mathrm{s})$ which yields (9). Similar to (8), independence assumptions among pilot and replica collision events make (9) an approximation on the actual PLR for irregular distributions. On the other hand, for concentrated distributions $\Lambda(x) = x^r$ and $\Psi(x) = x^p$, (9) becomes exact and holds with equality.

IV. Numerical Results

A. Simulation Setup

Simulation results are provided for CRA schemes in which each user transmits data payloads encoded with an $(n = 511, k = 421, t = 10)$ binary Bose–Chaudhuri–Hocquenghem (BCH) code. Part of the k information bits are used to validate the decoded packets via a cyclic redundancy check (CRC).

[2]Equation (8) assumes successful packet decoding when the packet is a singleton one. In this respect, it represents a lower bound on the actual PLR. It is possible to account also for ii) in the analysis, i.e., PHY layer impairments and channel model, as done in [18]. For a large number of antennas, the impact of ii) is however negligible compared to i).

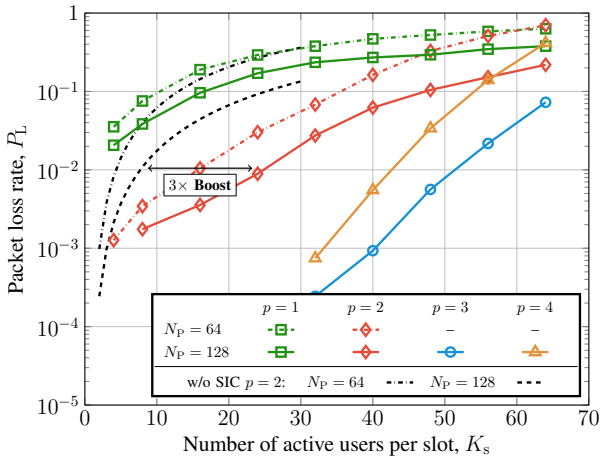

Fig. 2. Packet loss rate comparison between different pilot mixture schemes in a slotted and unframed scenario.

After padding the BCH codeword with a final zero bit, the encoded bits are mapped onto a quadrature phase-shift keying (QPSK) constellation with Gray mapping, yielding a data payload of $N_\mathrm{D} = 256$ symbols. Simulation results are given for $M = 256$ BS antennas. Pilots are constructed using Hadamard matrices [19]. Both schemes with and without feedback are considered for the numerical analysis. In case a feedback channel is used, it is assumed ideal (i.e., all ACKs are always successfully received). We plot the PLR P_L against the number of active users, representing the system scalability parameter. This latter parameter is equal to K_s for a slotted and unframed system, while it is equal to K_a for a slotted and framed one. In the framed case, we use $N_\mathrm{s} = 62$ slots.

B. Performance Evaluation

1) Slotted and Unframed System Analysis: In Fig. 2, we compare schemes using pilot mixtures and inner SIC ($p > 1$) varying the number of available orthogonal pilot sequences $N_\mathrm{P} \in \{64, 128\}$, for a slotted but unframed system. The figure also shows the comparison between different concentrated distributions $\Psi(x) \in \{x, x^2, x^3, x^4\}$. The gain attained over the scheme with $\Psi(x) = x$ can be interpreted as the performance boost given by pilot mixtures and proposed inner SIC. In the figure, we emphasize that increasing the preamble order p has a positive effect for small values of p (e.g., moving from $p = 1$ to $p = 3$ in the example). In contrast, when p increases beyond a certain value, pilot overload in the slot deteriorates the PLR due to impossibility of performing accurate channel estimation, an effect that becomes predominant. A second result we emphasize is the effectiveness of the inner SIC in comparison to the analytical and ideal curve without SIC derived in (8). For example, considering $N_\mathrm{P} = 128$ and $p = 2$, inner SIC increases by a factor of 3 the number of served users per slot at a target PLR $P_\mathrm{L}^* = 10^{-2}$. We also note that increasing p in absence of inner SIC does not significantly improve scalability at low PLR. Finally, we can see that increases N_P, the performance improves as expected.

2) Slotted and Framed System Analysis: In Fig. 3 we analyze the role of the preamble order p in a slotted and framed system.

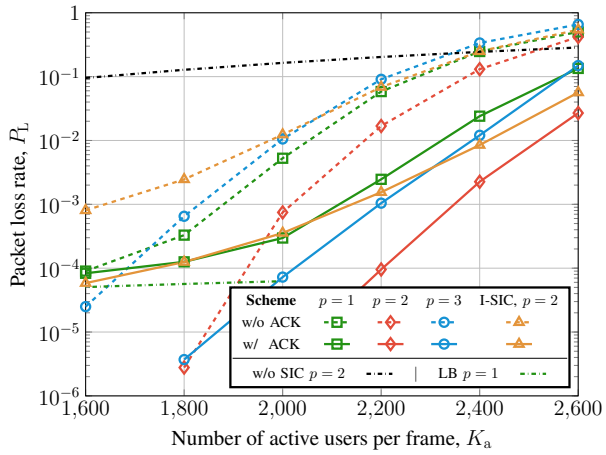

Fig. 3. Packet loss rate comparison between different pilot mixture schemes in a framed scenario. I-SIC: inner SIC.

We report the PLR for schemes using: i) pilot mixtures and the proposed nested SIC; ii) pilot mixtures and only inner SIC; iii) no pilot mixture (i.e., $p = 1$) and outer SIC. Results are carried out also allowing the possibility to employ ACK messages at the end of each slot, as suggested in [9]. The reception of an ACK by an active user enables the possibility to preemptively stop transmissions of subsequent replicas by the user, that saves energy and does not generate unnecessary interference in future slots. A comparison is provided varying the number of active users in the frame, K_{a}, for different concentrated distributions $\Psi(x) \in \{x, x^2, x^3\}$. For all schemes we adopt $r = 2$ (i.e., $\Lambda(x) = x^2$) and $N_{\mathrm{p}} = 128$. Using (9) we observe that we cannot achieve good scalability at low PLR without SIC. This is due to the fact that we are exploiting only time diversity, while SIC permits to retrieve several collided users. As an example, for $p = 2$ we obtain a target $P_{\mathrm{L}}^* = 10^{-3}$ at $K_{\mathrm{a}} \approx 400$ without SIC. Enabling only inner SIC, we significantly improve scalability to $K_{\mathrm{a}} \approx 1600$ for the same P_{L}^*. Next, enabling ACK messages we obtain a further scalability gain, reaching $K_{\mathrm{a}} \approx 2150$. Lastly, with the proposed nested SIC we end up with $K_{\mathrm{a}} \approx 2350$. As reviewed in Section III-A, the state-of-the-art schemes using $p = 1$ may suffer from intolerably high error floors under the asymptotically optimal distribution $\Lambda(x) = x^2$ [16]. Pilot mixtures represent a valid solution to lower the error floor without sacrificing the waterfall performance. In this example, fixing $K_{\mathrm{a}} = 1800$ and using the analysis in Section III-A, the PLR of the scheme with $p = 1$ is lower bounded by $P_{\mathrm{L}} \geq 5.7 \cdot 10^{-5}$, while that of the scheme with $p = 2$ is lower bounded by $P_{\mathrm{L}} \geq 1.4 \cdot 10^{-8}$. Unexpectedly, as opposed to the unframed case, we highlight that in the framed one the choice $p = 2$ outperforms $p = 3$. This reveals that focusing on single slot analysis may provide inaccurate results and conclusions in framed system, which motivates CRA design and analysis.

V. Conclusions

A grant-free CRA-type access scheme has been proposed that exploits both massive MIMO and pilot mixture preambles through a nested SIC algorithm. As a remarkable result, the proposed processing is able to substantially lower the high error floor typical of energy-efficient CRA configurations featuring two replicas per user. Analytical performance benchmarks have been derived to assess the improvement due to SIC, as well as a lower bound useful for error floor estimation and system design. The proposed approach provides good performance even in slotted but unframed systems, where only the inner SIC algorithm is applied, which is useful in situations where low-complexity BS processing is required.

Acknowledgements

We would like to thank M. Chiani for helpful discussions.

References

[1] 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.

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

[3] J. Gao, W. Zhuang, M. Li, X. Shen, and X. Li, "MAC for machine-type communications in industrial IoT—Part I: Protocol design and analysis," *IEEE Internet of Things J.*, vol. 8, no. 12, pp. 9945–9957, Jun. 2021.

[4] S. R. Pokhrel, J. Ding, J. Park, O.-S. Park, and J. Choi, "Towards enabling critical mMTC: A review of URLLC within mMTC," *IEEE Access*, vol. 8, pp. 131 796–131 813, Jul. 2020.

[5] 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.

[6] E. Paolini, Č. Stefanović, 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.

[7] 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.

[8] J. H. Sørensen, E. De Carvalho, Č. 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.

[9] 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.

[10] K. Senel, H. V. Cheng, E. Björnson, and E. G. Larsson, "What role can NOMA play in massive MIMO?" *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 3, pp. 597–611, Jun. 2019.

[11] 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.

[12] 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.

[13] X. Dai, T. Yan, Q. Li, H. Li, and X. Wang, "Pattern division random access (PDRA) for M2M communications with massive MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 70, no. 12, Dec. 2021.

[14] G. Choudhury and S. Rappaport, "Diversity ALOHA-a random access scheme for satellite communications," *IEEE Trans. Commun.*, vol. 31, no. 3, pp. 450–457, Mar. 1983.

[15] 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.

[16] L. Valentini, M. Chiani, and E. Paolini, "A joint PHY and MAC layer design for coded random access with massive MIMO," in *Proc. 2022 IEEE Global Commun. Conf.*, Rio di Janeiro, Brazil, Dec. 2022.

[17] K.-H. Ngo, A. G. i Amat, and G. Durisi, "Irregular repetition slotted ALOHA over the binary adder channel," in *Proc. 2023 IEEE Int. Conf. Commun.*, Rome, Italy, May 2023.

[18] L. Valentini, M. Chiani, and E. Paolini, "Interference cancellation algorithms for grant-free multiple access with massive MIMO," *IEEE Trans. Commun.*, 2023, to appear.

[19] G. L. Stüber and G. L. Steuber, *Principles of Mobile Communication*. Springer, 2001, vol. 2.