The Interplay of Spectral Efficiency, User Density, and Energy in Grant-based Access Protocols

Derya Malak

***Abstract*—We employ grant-based access with retransmissions for multiple users with small payloads, particularly at low spectral efficiency (SE). The radio resources are allocated via non-orthogonal multiple access (NOMA) in the time into T slots and frequency dimensions, with a measure of non-orthogonality η. Retransmissions are stored in a receiver buffer with a finite size C_{buf} and combined via Hybrid Automatic Repeat reQuest (HARQ), using Chase Combining (CC) and Incremental Redundancy (IR). We determine the best scaling for the SE (bits/rdof) and for the user density J/n, for a given number of users J and a blocklength n, versus signal-to-noise ratio (SNR, ρ) per bit, i.e., the ratio E_b/N_0, for the sum-rate optimal regime and when the interference is treated as noise (TIN), using a finite blocklength analysis. Contrasting the classical scheme (no retransmissions) with CC-NOMA, CC-OMA, and IR-OMA strategies in TIN and sum-rate optimal cases, the numerical results on the SE demonstrate that CC-NOMA outperforms, almost in all regimes, the other approaches. For high C_{buf} and small η, IR-OMA could surpass CC-NOMA. At low E_b/N_0, the SE of CC-OMA with TIN, as it exploits CC and offers lower interference, can approach the trend of CC-NOMA and outperform the other TIN-based methods. In the sum-rate optimal regime, the scalings of J/n versus E_b/N_0 deteriorate with T, yet from the most degraded to the least, the ordering of the schemes is as (i) classical, (ii) CC-OMA, (iii) IR-OMA, and (iv) CC-NOMA, demonstrating the robustness of CC-NOMA. Contrasting TIN models at low ρ, the scalings of J/n for CC-based models improve the best, whereas, at high ρ, the scaling of CC-NOMA is poor due to higher interference, and CC-OMA becomes prominent due to combining retransmissions and its reduced interference. The scaling results are applicable over a range of η, T, C_{buf}, and J, at low received SNR. The proposed analytical framework provides insights into resource allocation in grant-based access and specific 5G use cases for massive ultra-reliable low-latency communications (URLLC) uplink access.**

***Index Terms*—Multiple access, NOMA, HARQ, Chase combining, Incremental redundancy, spectral efficiency, SNR per bit, user density, matched filter decoding, maximum ratio combining, and HARQ receiver buffer.**

I. Introduction

The fifth-generation (5G) communication networks will support a wide range of use cases beyond high data rate applications, including Ultra-reliable, low-latency communication (URLLC) settings with small payload sizes transmitted by a large number of users with stringent power requirements. 4G LTE cannot effectively handle the heterogeneity because it ensures interference-free transmission via scheduled access and is designed to support fewer devices with large payloads. On the other hand, the overhead of scheduled access in 4G LTE is not desirable in URLLC applications.

Motivated by the challenges in scheduled access, we consider a wireless multiple access channel (MAC) model where a set of users sends their fixed payloads (in bits) given a preallocation of uplink resources. A given set of shared spectral resources of bandwidth ω Hertz (Hz) is partitioned into B non-overlapping frequency bins shared by the users via non-orthogonal multiple access (NOMA), and the time of duration τ second (sec) is divided into T transmit opportunities. Note that in HARQ, erroneous packets are not discarded and are instead stored in a buffer with a finite size C_{buf} and combined with retransmitted packets [2]. Keeping this situation in mind, we consider different forms of Hybrid Automatic Repeat reQuest (HARQ): (i) HARQ with Chase Combining (CC) of NOMA transmissions, CC-NOMA, (ii) HARQ with Chase combining of OMA transmissions, CC-OMA, and (iii) HARQ with Incremental Redundancy (IR), IR-OMA. The general challenge is to design a random access protocol to maximize the scaling of the density of users versus the SNR per bit.

Via the proposed retransmission-based grant-based access scheme, we aim to address the following central questions for 5G wireless networks and beyond:

- The spectral efficiency (SE, which is the total number of data bits per total real number of degrees of freedom, rdof) versus signal-to-noise ratio (SNR) per bit (or equivalently E_b/N_0) tradeoff for different HARQ schemes with retransmissions via Chase combining or Incremental redundancy. What are the gains in the sum-rate optimal technique[1] that describes an upper bound to the sum rate of the users, versus the per-user rate approach based on TIN only?
- How sensitive is the scaling of the SE versus E_b/N_0 to the number of retransmissions, T, different SNR, ρ, regimes, different uplink load J regimes, where we keep the total power fixed?

[1]The sum-rate optimal rate is achieved via successive interference cancellation (SIC) by treating interference as noise (TIN) [3].

D. Malak is with the Commun. Systems Dept., EURECOM, Biot Sophia Antipolis, 06904 FRANCE (email: derya.malak@eurecom.fr).

The material in this paper was presented in part at the 20th Int. Symp. Modeling and Optim. in Mobile, Ad Hoc, and Wireless Netw. (WiOpt 2022), Turin, Italy, and received the Best Paper Award [1].

Derya Malak's research is partially supported from a Huawei France-funded Chair towards Future Wireless Networks. This work was supported in part by the National Science Foundation under Grant CNS 2008639.

- The blocklength, n, versus the number of retransmissions, T. We assume that each time slot $t \in \mathcal{T} = \{1, \ldots, T\}$ accommodates the transmission of m symbols. How does signal-to-interference-plus-noise ratio (SINR) change with $n = mT$, where m is the blocklength per retransmission?
- The impact of a finite HARQ buffer size on the throughput of HARQ. The size of the buffer available at the receiver, denoted by C_{buf}, to store previously received packets impacts the throughput of retransmission-based schemes. How does C_{buf} affect the scaling performances?
- NOMA-based signaling and the effect of non-orthogonal user signature correlations, denoted by the non-orthogonality factor η, on the scaling of the user density J/n (users/rdof) versus E_b/N_0 given a number of users, J, and as a function of T. How should we design[2] η for the conventional matched filter receiver (MFR) for single-user detection (SUD) for decoding of random signatures?

We next review the connections to the state-of-the-art and summarize the bottlenecks.

A. Related Work

Channel access models and throughput scaling. Random-access protocols have been pioneered with the emergence of ALOHA [4] and slotted or reservation-based ALOHA [5], [6] schemes, which later yielded the development of carrier sense multiple access. However, these contention-based schemes do not have desirable throughput and delay performances and do not guarantee a deterministic load. Recently, different uplink schemes have been proposed to accommodate massive access [7]. In general, the resource being shared is on a time-frequency grid, and each transmission costs one time-frequency slot (TFS). The throughput – incorporating the user identification – has been characterized for massive user connectivity with orthogonal access in [8] from a DoF perspective. Other models include sparse code multiple access for grant-free access [9], multi-user detectors (MUDs) to improve performance of random-CDMA [10], [11] for spread spectrum systems, e.g., orthogonal multiple access (OMA), coded OFDM, and NOMA [12]. Others have focused on the capacity of Gaussian MACs [13], MAC with user identification [14], quasi-static fading MAC [15], Gaussian MAC with feedback [16]. Finite blocklength (FBL) achievability bounds for the Gaussian MAC and random access channel under average-error and maximal-power constraints have been devised in [17], with single-bit decoder feedback.

Random-access versus grant-based access. We have studied grant-free access in [18] and [19] to maximize the rate of users simultaneously accessing the channel given a common outage constraint, with no upper bound on the transmitted energy. Grant-free access protocols are better suited for scenarios where only a specific subset of users share the resources at any given time, e.g., applications of the Internet of Things (IoT), sideline, or 5G New Radio. These protocols are convenient for a broad range of IoT use cases of enhanced Machine-Type Communication (eMTC) and narrowband-IoT [20]. On the other hand, our proposed approach focuses on the scaling of the user density while allowing the grouping of multiple users to share the resources and at the same time tolerating interference and collisions between the users. Hence, our protocol lies between grant-free and grant-based techniques because it allows (i) multiple users to share the resources, and (ii) a collision-aware resource allocation, by grouping users to maximize the scaling of the user density per SNR per bit. Therefore, it is better tailored to capture the tradeoff between SE and SNR per bit for eight (re)transmission models.

Interference management and resource sharing. Different interference management techniques have been studied under different spectral efficiency models. To accommodate massive random access, interference cancellation [21], collision resolution [7], load control [22], and interference cancellation given a target outage rate [23] have been proposed. From the perspective of fundamental limits, the best achievable rate region for two user Gaussian interference channel is given by Han-Kobayashi [24]. While interference alignment is a good technique for specific channel parameters [25]–[28], the capacity region for a large number of users is unknown because ideal interference cancellation is not practical. In [18], we characterized the scaling of throughput (user density) with a deadline for a suboptimal but practical random access system where the time and frequency domains are slotted, and the receiver uses conventional SUD under an SINR-based outage constraint. However, the fixed per-user power in [18] causes a linear scaling between the received SNR and the number of users.

Critical performance metrics. Using different power levels to reduce E_b/N_0 has been considered in [29]. Random linear coding with approximate message passing decoding for many-user Gaussian MAC has been studied in [30], where the authors derive the asymptotic error rate achieved for a given user density, user payload in bits, and user energy. Cognitive radio and NOMA have been blended to maximize the achievable rate of the secondary user without deteriorating the outage performance of primary user [31]. Dynamic power allocation and decoding order at the base station for two-user uplink cooperative NOMA-based cellular networks has been studied in [32], where the authors demonstrated the superior performance over traditional two-user uplink NOMA (without cooperation).

HARQ models and generation of coding sequences. HARQ is a combination of Automatic Repeat reQuest (ARQ) and forward error correction (FEC) [33]. In particular, there are three models known as HARQ with Selective Repeat, HARQ with CC, and HARQ with IR [34], [35], [36]. This one is a salient variant of HARQ that captures puncturing via parity bits, e.g., puncturing with Turbo codes

[2] η can be made sufficiently small for large blocklengths. For a blocklength $m = \frac{n}{T}$ per transmission, $\eta \approx \frac{1}{\sqrt{m}}$ [3].

[37], and effects of different HARQ buffer sizes [2]. In general, it is well known that successive refinement of information can provide an optimal description from a rate-distortion perspective [38], [39], and incremental refinements and multiple descriptions with feedback have been explored [40].

Coding sequences have been devised for massive access, including Walsh sequences and decorrelated sequences [41], and Khachatrian-Martirossian construction to enable $K > n$ users signal in n dimensions simultaneously, where $K \approx \frac{1}{2} n \log_2 n$ is the optimal scaling [3, Slides 57-59]. Furthermore, it has been shown that when the inputs are constrained to ± 1, it is possible to have $K \gg n$. Zadoff-Chu sequences provide low complexity and constant-amplitude output signals, and have been widely used in 3GPP LTE air interface, including the control and traffic channels [42]. Multi-amplitude sequence design for grant-free MAC has been contemplated in [43]. However, inducing a high E_b/N_0, this approach is not desirable in a practical massive access scenario.

B. Overview, Contributions and Organization

The goal of this paper is to analyze a retransmission and grant-based access framework that unifies the properties of NOMA-based transmissions with HARQ-based protocols that rely on CC and IR to provide insights on uplink resource allocation strategies for future 5G wireless communication networks. In Section II, we detail the system model for grant-based access and the key performance metrics, SE (bits/rdof), the SNR per bit (E_b/N_0), and user density (users/rdof) for a given blocklength, total received power constraint, and a total number of retransmissions. We delineate the retransmission and grant-based access schemes in Section III and analyze their SE and the SNR per bit for the sum-rate optimal and TIN cases. More specifically, we consider retransmission-based models where the receiver jointly decodes transmissions via (i) the classical transmission scheme with no retransmissions and the retransmission-based schemes with combining, namely (ii) CC-NOMA, (iii) CC-OMA, and (iv) IR-OMA. Our analysis incorporates channel power gains and the capacity for the FBL channel model. In Section IV, we numerically evaluate the SE versus SNR per bit tradeoff and the user density versus SNR per bit tradeoff, and show their behaviors with respect to the number of transmissions T, received SNR ρ, HARQ buffer size C_{buf}, non-orthogonality factor η, and the total number of users J.

The key design insights for the proposed grant-based access framework are as follows:

- ***The low ρ regime is relevant.*** We exploit the conventional MFR for SUD suitable at low SNRs ρ. We show that the user density of NOMA-based models scales significantly better at low ρ versus high ρ. The interference cannot be exploited at high ρ, degrading the performance of TIN-based models. The minimum SNR per bit to achieve a non-zero user density grows with ρ.
- ***The SE of the sum-rate optimal strategy improves with NOMA.*** The scalings of SE versus E_b/N_0 for various schemes show that for any given value of E_b/N_0, the best performance is attained by $\mathsf{SE}_{\mathsf{sum}}^{\mathsf{CC,NOMA}}$, and mainly for small T. Compared to OMA-based transmissions, CC-NOMA has a better SE versus E_b/N_0 performance. The performance of IR-OMA approaches that of the classical model as C_{buf} at the receiver increases. The numerical results indicate that $\mathsf{SE}_{\mathsf{sum}}^{\mathsf{CC,NOMA}}$ outperforms the other strategies almost in all regimes. While $\mathsf{SE}_{\mathsf{sum}}^{\mathsf{CC,NOMA}}$ significantly improves with increasing J, and $\mathsf{SE}_{\mathsf{sum}}^{\mathsf{CC,OMA}}$ is less sensitive to C_{buf}, at high C_{buf}, $\mathsf{SE}_{\mathsf{sum}}^{\mathsf{IR,OMA}}$ performs, in general, better than $\mathsf{SE}_{\mathsf{sum}}^{\mathsf{CC,OMA}}$, and it could outperform $\mathsf{SE}_{\mathsf{sum}}^{\mathsf{CC,NOMA}}$ for small η.
- ***The SE of the TIN strategy is optimal at low ρ.*** Provided that C_{buf} is sufficiently large, TIN is good at low SE. If not, a higher T is required. The scaling results are sensitive to η for CC-NOMA, and a codebook with a smaller η can significantly improve the SE of TIN. At low E_b/N_0, the performance of $\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{CC,OMA}}$ can outperform $\mathsf{SE}_{\mathsf{sum}}^{\mathsf{CC,NOMA}}$ and other TIN-based methods because it exploits CC and offers lower interference than $\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{CC,NOMA}}$. At large C_{buf} and T, $\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{CC,OMA}}$ can be superior to $\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{Clas.}}$.
- ***User density is sensitive to retransmissions.*** For the sum-rate optimal model, although the performances of CC-OMA, IR-OMA, and the classical techniques degrade with T, CC-NOMA does not sacrifice the number of users per rdof as much. A higher number of users J, under fixed per-user power, results in a lowered received SNR ρ per user, which improves the SE for the sum-rate optimal CC-NOMA model, yet for the TIN-based model, the SINR drops due to the higher interference. For TIN, CC-OMA, and CC-NOMA perform well for low ρ, and CC-OMA can effectively combine retransmissions even for high ρ. However, the SNR per bit demand for CC-NOMA is sensitive to ρ, deteriorating the performance at high ρ. The SE of the classical model and IR-OMA do not scale as well as CC-OMA because the former models cannot compensate for the interference at high ρ and hence cannot leverage retransmissions. The J/n versus E_b/N_0 performances of sum-rate optimal models deteriorate in T. The ordering of the models in the sum-rate optimal regime, from the most to the least sensitive to degradation, as an increasing function of T, is (i) classical, (ii) CC-OMA, (iii) IR-OMA, and (iv) CC-NOMA, demonstrating the robustness of CC-NOMA to retransmissions.
- ***User density scales up with SNR per bit.*** The user density J/n can superlinearly scale with E_b/N_0 (where the scaling does not necessarily degrade with T in the case of CC-NOMA versus the OMA-based models) in the FBL and the infinite blocklength (IBL) regimes, where IBL gives an upper bound to the scaling, which becomes tighter as n increases. Both for sum-rate optimal and TIN-

This article has been accepted for publication in IEEE Transactions on Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2024.3361550

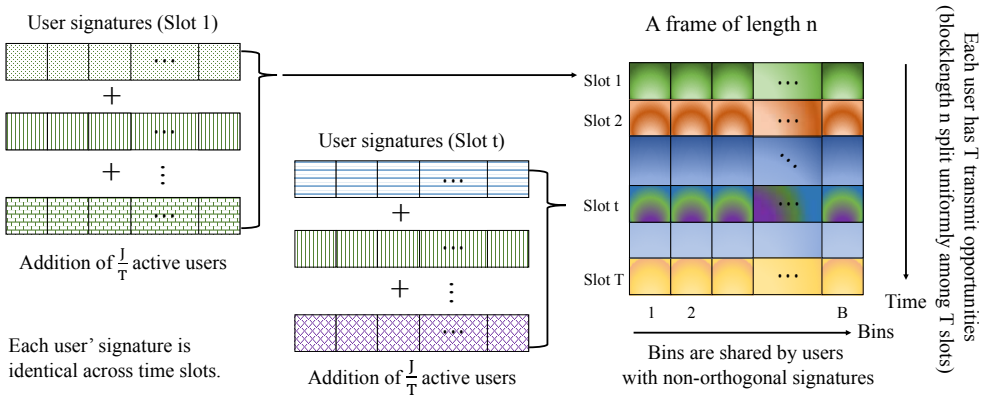

Fig. 1: (a) Non-orthogonal user signatures at time slots 1 and t. Each user uses the same signature across all time-frequency resources. The second user in slot 1 is repeated in slot t (same signature). (b) The frame structure where time is partitioned into T transmit opportunities, and the time-frequency resources are shared in a non-orthogonal manner by the users.

based models, the scaling of J/n versus SNR per bit does not improve with ρ due to the increase in the SNR per bit. Comparing different TIN models, at low ρ, the scalings of J/n versus E_b/N_0 for CC-NOMA and CC-OMA improve similarly, whereas the schemes that do not promote CC do not perform as well. With increasing ρ, the scaling of CC-NOMA deteriorates due to high interference, whereas CC-OMA performs the best among all as it combines retransmissions and provides reduced interference. In the TIN-based CC-NOMA and CC-OMA models, the scalings of J/n improve with increasing T, and the scalings for the IR-OMA and the classical models are not sensitive to T.

Our insights could be applied to 5G wireless system design with delay and resource-constrained communications, which is critical in use cases such as URLLC or mMTC. Nevertheless, the scaling results in our framework provide an upper bound on the achievable SE and the user density because of the following additional assumptions: ideal negative acknowledgment with no error or delay, the IBL regime capacity-achieving encoding, perfect power control, perfect synchronization among users, and decoding via a suboptimal receiver, through matched filtering and SUDs, versus MUD, which could strictly improve performance of random-CDMA [3, slide 146]. A more general framework that allows studying the above key points as well as path loss and outage capacity-based models both for the IBL and the FBL regimes will be considered as future work, as detailed in Section V.

II. System Model

We consider a wireless grant-based access communication model where a collection of users transmits over shared radio resources to a common receiver. The goal of each user is to transmit its payload of fixed size (L bits) within a latency constraint (blocklength n). A user is granted T retransmission attempts, i.e., time slots, to communicate its payload. The users use non-orthogonal signatures to transmit their payloads, as shown in Figure 1-(a). The signatures are kept identical at each attempt.

a) Frame structure: A frame has a total bandwidth of ω Hz and the time of duration τ sec, and is partitioned into B frequency bins of equal width, and T time slots, i.e., transmit opportunities, of equal duration. We refer to a given time slot and frequency bin as a TFS. For the proposed frame structure, the total number of resources or real degrees of freedom (rdof) in a frame is $N = \omega\tau$, which is evenly split into T retransmissions. The TFSs in a frame are shared by a collection of users in a non-orthogonal manner. While in OMA-based approaches, the rdof is split orthogonally among the users, in NOMA-based transmissions, the TFSs in a frame are shared by a collection of users in a non-orthogonal manner. Each user attempts to transmit its payload of fixed-size L bits over shared resources. Given ω, τ, m, and T, the number of symbols in a TFS is $\omega\tau/(BT)$. Under the orthogonal division of the resources, the coding rate is $LBT/(\omega\tau)$ bits per transmitted symbol.

b) User (source) model: Given T (re)transmission attempts, the total blocklength n per-user is split uniformly across T attempts to accommodate the retransmission of a packet. Hence, the blocklength per transmission at each time slot is $m = n/T$. Let J_t and $\mathcal{J}_t$ be the number and set of users at slot $t \in \mathcal{T}$, respectively, such that $J = \sum_{t=1}^{T} J_t$, and $\mathcal{J}$ be the set of all users in the frame.

Let $\mathbf{U}_j = (U_{j1}, U_{j2}, \ldots, U_{jK})$ be the K dimensional source vector corresponding to user $j \in \mathcal{J}$. In the case of no feedback, let $\phi_{tji} : \mathcal{U}_j^K \to \mathcal{V}_j$ for $i \in \{1, \ldots, m\}$ be the encoder function for $j \in \mathcal{J}$ that captures the mapping from $\mathbf{U}_j$ to the channel input $\mathbf{V}_{tj} = \phi_{tji}(\mathbf{U}_j) = b_{tj}\mathbf{S}_j$, i.e., the product of the complex amplitude $b_{tj} \in \mathbb{C}$ of the transmitted symbol and selected signature sequence $\mathbf{S}_j$, for attempt $t \in \mathcal{T}$, where each retransmission $\mathbf{V}_{tj} = (V_{tj1}, V_{tj2}, \ldots, V_{tjm})$ from user j has a blocklength m.

c) User signatures: The number of rdof N in a frame can be thought of as the total length of the signature sequences of the active users over B frequency bins. Each user has the same signature across all time-frequency resources. The TFSs are shared in a non-orthogonal manner, where each waveform at a given time slot is a sum of non-orthogonal signatures, which is shown in Figure 1-(a). We assume that the signature sequences $\mathbf{S}_j$ are unitary, $\|\mathbf{S}_j\| = 1$, i.e., each signature has unit variance, $\mathbb{E}[\mathbf{S}_j^\intercal \mathbf{S}_j] = 1$, and $|\langle \mathbf{S}_j, \mathbf{S}_{j'} \rangle| = |\mathbb{E}[\mathbf{S}_j^\intercal \mathbf{S}_{j'}]| = \eta$ for any $\{(j, j') \in \mathcal{J}_t : j \neq j'\}$. The maximum value of J_t to ensure that all $j \in \mathcal{J}_t$ is decoded with zero-error is given by the Khachatrian-Martirossian construction [3] allows $J_t > m$ users. Under this setup, when $\mathbf{S}_j$'s are random and m is large, $\eta \approx \frac{1}{\sqrt{m}}$ with high probability. We sketch the frame structure with overlapping NOMA traffic in Figure 1-(b).

d) Received signal and conventional matched filter decoding: The transmitted signal from user $j \in \mathcal{J}_t$ multiplied by the channel gain determines the received signal, given by $\mathbf{X}_{tj} = a_{tj}\mathbf{S}_j$, where $a_{tj} \in \mathbb{C}$ is the complex amplitude of the product of the values of the transmitted symbol b_{tj}, and the channel gain H_{tj} of user j at slot t, accounting for fading. Hence, $|a_{tj}|^2 = |b_{tj}|^2 \cdot |H_{tj}|^2$ represents the transmitted signal power times channel power gain variable (see Appendix A in [44]). We denote the received signal vector during transmission $t \in \mathcal{T}$ by $\mathbf{Y}_t = (Y_{t1}, Y_{t2}, \ldots, Y_{tm})$. The channel is additive such that the received signal vector during transmission $t \in \mathcal{T}$ is[3]

$$\begin{aligned}\mathbf{Y}_t &= \mathbf{X}_{tj} + \sum_{j' \in \mathcal{S}_{t,-j}} \mathbf{X}_{tj'} + \mathbf{Z}_t \\ &= a_{tj}\mathbf{S}_j + \sum_{j' \in \mathcal{S}_{t,-j}} a_{tj'}\mathbf{S}_{j'} + \mathbf{Z}_t, \quad t \in \mathcal{T} \ , \end{aligned} \tag{1}$$

where $\mathcal{S}_{t,-j}$ is the collection of the interferers of $j \in \mathcal{J}_t$ in the same time slot t, i.e., $\mathcal{S}_{t,-j} = \{j' \in \mathcal{J}_t : j' \neq j\}$, and $\mathbf{Z}_t \sim \mathcal{CN}(0, \sigma_t^2 I_m)$ is a complex Gaussian random variable. We assume perfect channel knowledge at the receiver, whereas the receiver has no access to $\mathbf{S}_j$.

We consider the *conventional matched filter receiver* (MFR) for decoding, which performs approximately optimal when the target SINR is low. In this case, the effective bandwidth required by the conventional approach is small versus the linear decorrelator receiver, which allows many users per DoF, where the other users' signals are treated as additive white Gaussian noise (AWGN) [11]. If the target SINR is high, both the linear minimum mean-square error (MMSE) and the linear decorrelator receiver decorrelate a user from the rest, yielding no more than one DoF per interferer [11]. When $\{\mathbf{S}_j\}_{j \in \mathcal{J}}$ are known to the receiver, an MMSE-based receiver provides a better signal-to-interference ratio (SIR) per-user via exploiting the structure of the interference [11].

[3]In the case with feedback, $V_{tji} = \phi_{tji}(\mathbf{U}_j, \mathbf{Y}_t^{i-1})$ is the channel input from user $j \in \mathcal{J}$ at time $i \in \{1, \ldots, m\}$ for attempt $t \in \mathcal{T}$, where $\mathbf{Y}_t^i = (Y_{t1}, Y_{t2}, \ldots, Y_{ti})$, and $\phi_{tji} : \mathcal{U}_j^K \times \mathcal{Y}^{i-1} \to \mathcal{V}_j$ is the encoder function. We leave the feedback setting for future work.

The maximum number of supported users for MFR derived in [11] as a function of a target SIR, and the received power under power control. In [11], different from our approach, the characterization of the scaling results for the users is based on (i) a target SIR requirement at the receiver for all users, and (ii) the asymptotic regime in the number of users, contrary to the FBL regime analyzed in the current work.

e) Maximum ratio combining: We assume that the receiver's HARQ buffer size equals the number of coded symbols per coded packet, where the retransmitted packets are summed up with previously received erroneous packets via maximum ratio combining (MRC) of retransmissions prior to decoding.

The common receiver has the decoder function $\Phi_T : \mathcal{Y}^n \to \{\mathcal{U}_j^K\}_j$ that combines T retransmissions to decode the individual source vectors $\{\mathbf{U}_j\}_j$ from the received signal vectors $\mathbf{Y}_t$, $t \in \mathcal{T}$. Using (1), the MRC of T transmissions results in the following combined signal:

$$\mathbf{Y} = \mathbf{U}_j + \sum_{t=1}^{T} a_{tj}^* \sum_{j' \in \mathcal{S}_{t,-j}} a_{tj'}\mathbf{S}_{j'} + \mathbf{Z} \ , \tag{2}$$

where $\mathbf{Y} = \sum_{t=1}^T a_{tj}^* \mathbf{Y}_t$, and $\mathbf{U}_j = \sum_{t=1}^T |a_{tj}|^2 \mathbf{S}_j$, and $\mathbf{Z} = \sum_{t=1}^T a_{tj}^* \mathbf{Z}_t$ are m dimensional vectors. We assume that the coefficients a_{tj} are known. These coefficients can be estimated using the least mean square (LMS) algorithm and then utilized by the MRC for generating the decision variable.

f) Per-user received SNR: The noise power each user sees is assumed to be additive and constant with value σ_t^2, $t \in \mathcal{T}$ per dimension, i.e., $\langle \mathbf{Z}_t, \mathbf{Z}_t \rangle = m\sigma_t^2$, where $m\sigma_t^2$ is the total noise power across the number of frequency bins, which is B. The average received power of user $j \in \mathcal{J}$ during transmission $t \in \mathcal{T}$, which is the total noise power times the received SNR, under unit channel power gain, is

$$\begin{aligned}\sum_{i=1}^{m} \mathbb{E}[X_{tji}^2] &= m\mathbb{E}[X_{tj1}^2] = \mathbb{E}[\mathbf{X}_{tj}^\intercal \mathbf{X}_{tj}] \\ &= |a_{tj}|^2 \mathbb{E}[\mathbf{S}_j^\intercal \mathbf{S}_j] = |a_{tj}|^2 = |b_{tj}|^2 = m\sigma_t^2 \rho_{tj} \ , \end{aligned}$$

where ρ_{tj} denotes the received SNR from $j \in \mathcal{J}$ during transmission $t \in \mathcal{T}$, noting that $\mathbb{E}[\mathbf{S}_j^\intercal \mathbf{S}_j] = 1$.

The energy constraint for each transmitted symbol $j \in \mathcal{J}$ at any given $t \in \mathcal{T}$ is

$$\mathbb{E}[\mathbf{X}_{tj}^\intercal \mathbf{X}_{tj}] = m\sigma_t^2 \rho_{tj} \leq \frac{KE_j}{T} \ , \tag{3}$$

where the parameter K is the message size of any source in bits, and E_j denotes an upper bound on the received energy of user $j \in \mathcal{J}$ per source dimension. In (3), the total power of channel input linearly scales with the message size K, yielding a maximum total energy of KE_j per message of user $j \in \mathcal{J}$, and an upper bound on the received energy per slot, denoted by $\frac{KE_j}{T}$. We assume that $\mathbb{E}[X_{tji}] = 0$ and X_{tji}'s across $j \in \mathcal{J}$ are not independent such that $\mathbb{E}[\mathbf{X}_{tj}^\intercal \mathbf{X}_{tj'}] = a_{tj}^* a_{tj'} \mathbb{E}[\mathbf{S}_j^\intercal \mathbf{S}_{j'}] = a_{tj}^* a_{tj'} \eta \leq \frac{KE_{jj'}}{T}$ for $\{(j, j') : j \neq j'\}$, noting that the non-orthogonal user signatures satisfy $|\langle \mathbf{S}_j, \mathbf{S}_{j'} \rangle| = \eta$.

Assuming that $\mathbb{E}[X_{tji}^2]$ does not change with $i \in \{1,\ldots,m\}$, and $\sigma_t^2 = \sigma^2$, from (3) we have

$$\rho_{tj} = \frac{\mathbb{E}[\mathbf{X}_{tj}^\intercal \mathbf{X}_{tj}]}{m\sigma^2} = \frac{1}{m\sigma^2}\sum_{i=1}^{m}\mathbb{E}[X_{tji}^2]$$
$$= \frac{\mathbb{E}[X_{tji}^2]}{\sigma^2} = \frac{|a_{tj}|^2}{m\sigma^2} \ , \quad j \in \mathcal{J} \ .$$

We assume that ρ_{tj} are identical and denoted by ρ. Given a constant received power of $m\sigma^2\rho$, the received SNR is $\rho = \frac{\mathbb{E}[\mathbf{X}_{tj}^\intercal \mathbf{X}_{tj}]}{m\sigma^2}$. For NOMA-based transmissions, the total power spent by all users is

$$P_{tot} = \frac{JTm\sigma^2\rho}{n} = J\sigma^2\rho \ , \tag{4}$$

or equivalently, the total energy spent for a given blocklength n is nP_{tot}. For OMA-based transmissions, the number of users per slot is J/T (versus J for NOMA-based), and $P_{tot} = (J/T)\sigma^2\rho$ [45, Ch. 4-6].

The overall problem is to determine some key performance metrics, which are the spectral efficiency, the SNR per bit, and the user density, and their joint behavior, which we describe in the sequel.

g) The spectral efficiency (SE): It is the maximum number of bits per channel use (bits/s/Hz):

$$\mathsf{SE} = \frac{\text{Total number of data bits}}{\text{rdof}} \ , \tag{5}$$

where rdof is the total number of real DoF, denoted by n.

Definition 1. (Achievable channel coding rate in the IBL regime [46].) *A rate R is achievable for a discrete memoryless channel (DMC) if for rates below capacity C such that*

$$R = \frac{1}{n}\log M < C \ , \tag{6}$$

there exists for sufficiently large n an (M, n) code with complete feedback, with maximal probability of error $\epsilon \to 0$. Conversely, for a sequence of codes (M, n), if $\epsilon > 0$, then it must hold that $R > C$.

Assume that a user attempts to transmit a payload of fixed size L bits over the channel. Hence, the relation between the required codebook size M and L is $L = \log M$. Hence, the blocklength n should be chosen sufficiently large so that the achievable transmit rate, $\frac{L}{n}$, satisfies:

$$\frac{L}{n} \le C = \frac{1}{2}\log_2(1+\text{SINR}) \ bit/rdof, \quad n \le N \ , \tag{7}$$

where C is channel capacity, and SINR represents the signal-to-interference-plus-noise ratio, for an AWGN channel where interference is treated as noise (TIN). The capacity is achievable at an arbitrarily low error rate in the IBL regime, i.e., as $n \to \infty$. However, since N is finite, the ratio L/N is always finite. Hence, given L, the IBL scheme gives an upper bound on R, and a lower bound on n.

In the FBL regime, let $M(n,\epsilon)$ be the maximal code size achievable with a given finite blocklength n, and average error probability ϵ. Then, the maximal rate achievable is approximated by [47].

Definition 2. (Achievable channel coding rate in the FBL regime [47].) *A rate R is achievable with complete feedback for a DMC if for any $\epsilon > 0$, there exists an (M, n) code such that*

$$R(n,\epsilon) = \frac{1}{n}\log M(n,\epsilon) \approx C - \sqrt{\frac{V}{2n}}Q^{-1}(\epsilon) \ , \tag{8}$$

for sufficiently large n, where $M(n,\epsilon)$ is the maximal code size achievable with a given blocklength n and average error probability ϵ, and $Q(x) = \frac{1}{\sqrt{2\pi}}\int_x^\infty e^{-u^2/2}\,\mathrm{d}u$ is the tail probability of the standard normal distribution where Q^{-1} is the inverse Q-function. Furthermore, in (8), $V = 1 - \frac{1}{(1+\text{SINR})^2}$ is the channel dispersion, and $C = \frac{1}{2}\log(1+\text{SINR})$ is the capacity in the units of nats per channel use.

While the Khachatrian-Martirossian construction is designed for the noiseless adder channel [48], it achieves a sum rate that approximates the sum rate of a Gaussian MAC (i) under the assumption of perfect channel inversion power control, such that $\rho = 1$, and (ii) when J is high, justifying (8) in Definition 2. Furthermore, the Gaussian approximation to the FBL regime is tight [47].

In the following, we instead express (8) as $R(n,\epsilon) \approx C - \Delta(C,n,\epsilon)$, where the channel dispersion V can be written as function of $\text{SINR} = \exp(2C) - 1$, and hence, the term $\Delta(C,n,\epsilon) = \sqrt{\frac{V}{2n}}Q^{-1}(\epsilon) = \sqrt{\frac{1}{2n}(1 - \frac{1}{\exp(2C)^2})}Q^{-1}(\epsilon)$ captures the joint behavior of C, n, and ϵ.

We note that at low ρ, the value of SINR is small and the channel dispersion in the FBL regime becomes negligible, yielding from (8) that the IBL approximation is good in the TIN regime.

h) The SNR per bit, E_b/N_0: It represents the ratio of the energy-per-bit to the noise power spectral density, which is a normalized SNR measure:

$$\frac{E_b}{N_0} = \frac{\text{Total energy spent}}{2 \times \text{Total number of bits}} \ , \tag{9}$$

which is dimensionless, and usually expressed in decibels (dB). We note that the scaling 2 in the denominator captures the total number of bits over the entire bandwidth, which is $2 \times rdof$.

i) User density (users/rdof): Given a total count of users $J = \sum_{t=1}^{T} J_t$, where J_t is the count of users active in slot $t \in \mathcal{T}$, and a total blocklength n given a frame duration T, *user density*, J/n, gives the total number of users per rdof that can transmit within the same frame. For the OMA-based schemes, where the blocklength per retransmission is m, the density of users in slot t is J_t/m. For NOMA-based schemes, the ratio J_t/n denotes the maximum density of users that can simultaneously transmit in $t \in \mathcal{T}$. From (5), (7), and (9), the achievable J/n is affected by the SE versus SNR per bit tradeoffs of the retransmission-based protocols for uplink access, which we detail in Section III.

III. Combining NOMA-based Retransmissions in Uplink

We focus on the scaling behaviors of the SE, the SNR per bit, and the user density for the retransmission-based grant-based access schemes. The senders must contend not only with the receiver noise but also with interference from each other. To that end, we next analyze the behavior of the SE and SNR per bit performances of the HARQ-based schemes for first, the *sum-rate optimal regime* that is attainable via SIC, and then the achievable data rate of a single user, i.e., per-user rate via treating the total interference from all other users as noise, i.e., *TIN*. However, our analysis does not capture the joint decoding of the intended user and the strongest interferers, which we leave as future work.

A. The Classical Transmission Scheme with No Multiplexing of Retransmissions

We commence with the classical interference-based model with no multiplexing across different time slots. Each user selects one slot to transmit its message given a blocklength n. The time resources are split uniformly across T slots. There are $J_t = J/T$ users per slot sharing the frame resources. In general, transmissions are exposed to different channel conditions, more specifically, the fading (e.g., Rayleigh fading) or path loss. Incorporating the channel gains $|H_{tj}|^2$, $t \in \mathcal{T}$, $j \in \mathcal{J}$, and assuming that $|H_{tj}|^2$ has unit power and is independent across the slots with a known cumulative distribution function, $F_{|H|^2}$, we can express the SE of the classical sum-rate optimal transmission approach as

$$\mathsf{SE}^{\mathsf{Clas.}}_{\mathsf{sum}} = \frac{1}{2T} \sum_{t \in \mathcal{T}} \log_2 \left(1 + \rho \sum_{j \in \mathcal{J}_t} |H_{tj}|^2\right) \; bit/rdof \; . \tag{10}$$

The SNR per bit of the classical sum-rate optimal transmission model is equal to

$$\frac{E_b}{N_0} = \frac{(J/T)\sigma^2 \rho n}{2m(\mathsf{SE}^{\mathsf{Clas.}}_{\mathsf{sum}} - \Delta(\mathsf{SE}^{\mathsf{Clas.}}_{\mathsf{sum}}, n, \epsilon))} \; , \tag{11}$$

where the total energy spent for a given blocklength n is $nP_{tot} = (J/T)\sigma^2 \rho n$, which is adapted for classical OMA-based transmissions from the relation in (4) for NOMA-based transmissions.

The SE of the classical model via TIN for decoding $j \in \mathcal{J}_t$, where $J_t = J/T$ users per slot, it holds that $\rho = 0$ for the remaining $T-1$ slots for which $j \notin \mathcal{J}_{t'}$, $t' \in \mathcal{T} \backslash \{t\}$, is expressed as

$$\mathsf{SE}^{\mathsf{Clas.}}_{\mathsf{TIN}} = \frac{1}{2T} \sum_{t \in \mathcal{T}} \sum_{j \in \mathcal{J}_t} \log_2 \left(1 + \rho |H_{tj}|^2 \Big/ \left(\rho \sum_{j' \in \mathcal{S}_{t,-j}} |H_{tj'}|^2 + 1\right)\right) \; bit/rdof \; , \tag{12}$$

where the summation $\sum_{j \in \mathcal{J}_t}$ in the front of the logarithm in (12) denotes the achievable SE per time slot for the TIN (versus no coefficient for the sum-rate optimal model in (10)).

Similarly, the SNR per bit for the classical transmission model with TIN is given as

$$\frac{E_b}{N_0} = \frac{(J/T)\sigma^2 \rho n}{2n(\mathsf{SE}^{\mathsf{Clas.}}_{\mathsf{TIN}} - \Delta(\mathsf{SE}^{\mathsf{Clas.}}_{\mathsf{TIN}}, n, \epsilon))} \; . \tag{13}$$

We next provide two results on E_b/N_0 for the IBL regime, for the classical transmission model.

Corollary 1. The classical transmission model. *For the classical transmission model in the IBL regime, for $|H_{tj}| = 1$, $\forall\, t \in \mathcal{T}$, $j \in \mathcal{J}_t$, exploiting the relation between SE and E_b/N_0, we next provide the relations between SE, P_{tot}, and J, for the sum-rate optimal and TIN regimes.*

(i) The classical sum-rate optimal approach. The measures SE and P_{tot} satisfy the relation

$$(2^{2\mathsf{SE}^{\mathsf{Clas.}}_{\mathsf{sum}}} - 1)\sigma^2 = P_{tot} \; .$$

(ii) The classical transmission approach with TIN. The measures SE, P_{tot}, and J satisfy

$$\frac{J}{T}\sigma^2 \frac{(2^{2\mathsf{SE}^{\mathsf{Clas.}}_{\mathsf{TIN}} \cdot \frac{T}{J}} - 1)}{1 - \left(\frac{J}{T} - 1\right)(2^{2\mathsf{SE}^{\mathsf{Clas.}}_{\mathsf{TIN}} \cdot \frac{T}{J}} - 1)} = P_{tot} \; ,$$

where SE is given in (12). We note under fixed total power P_{tot} that $\lim_{\rho \to 0} \frac{E_b}{N_0} = \log 2 \cdot T\sigma^2$ from (13).

From Cor. 1, we note that both $\mathsf{SE}^{\mathsf{Clas.}}_{\mathsf{sum}}$ and $\mathsf{SE}^{\mathsf{Clas.}}_{\mathsf{TIN}}$ increase in P_{tot}. Hence, we can determine the common E_b/N_0 value that leads to the classical sum-rate optimal and the classical TIN-based models to achieve $\mathrm{SE} \to 0$ (which is attained when $J \to \infty$ [3, Slide 69]). The convergence behavior for different T values can be observed in Section IV (see e.g., Figures 3 and 5).

In the case of no retransmissions, TIN is essentially optimal for low SE [3]. However, for strategies combining the retransmissions, TIN may not be optimal even at low SE, see e.g., [3] and [49]. We will next analyze the SE and the SNR per bit by incorporating the channel gains (to accurately capture the SINR) for the HARQ models in Sections III-B, III-C, and III-D, which will be followed by numerical simulations in Section IV to contrast the various HARQ schemes and demonstrate that sum-rate optimal schemes could be more energy efficient via combining of retransmissions versus TIN.

B. Chase Combining with NOMA-based Retransmissions

For a given payload L, a user transmits T times within a frame of duration τ sec. At the receiver, Chase combining is a common form of HARQ that is used to combine signal energy for a given user's transmissions over T slots. Chase combining has been shown to increase throughput in relatively poor channel conditions [18]. We assume that the transmission is successful, i.e., a user can have its payload decoded, at the end of T attempts when the Chase-combined SINR exceeds the critical threshold. Once a user's

transmission is successfully decoded after T attempts, it stops transmitting. The duration of each slot is τ/T seconds, to ensure that the user will meet the latency constraint τ.

In CC-HARQ, each transmission contains the same data and parity bits. The receiver's HARQ buffer size for CC-HARQ equals the number of coded symbols per coded packet, where the retransmitted packets are summed up at the receiver with previously received erroneous packets via MRC of retransmissions prior to decoding. In Figure 2 (left), we sketch CC-HARQ. We next derive the SE for the Chase combining of NOMA-based retransmissions (CC-NOMA) for the sum-rate optimal model.

Proposition 1. Sum-rate optimal model — Chase combining of non-orthogonal transmissions. *The SE of CC-NOMA for the sum-rate optimal model incorporating channel power gains is given as*

$$\mathsf{SE}_{\mathsf{sum}}^{\mathsf{CC,NOMA}} = \frac{1}{2}\log_2\Big(1+\rho\sum_{t\in\mathcal{T}}|H_{tj}|^2\Big[1+\eta^2\cdot \Big(\sum_{t\in\mathcal{T}}|H_{tj}|^2\Big)^{-2}\Big|\sum_{t\in\mathcal{T}}\sum_{j'\in\mathcal{S}_{t,-j}}H_{tj}H_{tj'}^*\Big|^2\Big]\Big)\ bit/rdof\ , \tag{14}$$

where $|H_{tj}|^2$ is the channel power gain of user $j\in\mathcal{J}$ (the one with the largest SINR) at slot $t\in\mathcal{T}$.

The SNR per bit for CC-NOMA for the sum-rate optimal model, using $\mathsf{SE}_{\mathsf{sum}}^{\mathsf{CC,NOMA}}$ in (14), equals

$$\frac{E_b}{N_0} = \frac{J\sigma^2\rho n}{2n(\mathsf{SE}_{\mathsf{sum}}^{\mathsf{CC,NOMA}} - \Delta(\mathsf{SE}_{\mathsf{sum}}^{\mathsf{CC,NOMA}}, n, \epsilon))}\ . \tag{15}$$

Proof. See Appendix A in [44]. □

We next provide two results on E_b/N_0 for CC-NOMA under the sum-rate optimal IBL model.

Corollary 2. CC-NOMA under the sum-rate optimal model. *In the IBL regime for unit channel power gains, the following relations hold in the limit as $\rho\to 0$ for $|H_{tj}|=1$, $\forall\ t\in\mathcal{T},\ j\in\mathcal{J}_t$.*

(i) A lower bound on E_b/N_0. The SNR per bit satisfies

$$\frac{E_b}{N_0} \geq -1.59dB + 10\log_{10} J\sigma^2 - 10\log_{10}\Big(T\Big[1+\eta^2\Big(\frac{J}{T}-1\Big)^2\Big]\Big)\ . \tag{16}$$

(ii) Sensitivity of the E_b/N_0 limit versus J. The SNR per bit in the limit as $\rho\to 0$, approaches

$$\lim_{\rho\to 0}\frac{E_b}{N_0} = \log 2\cdot P_{tot}\ . \tag{17}$$

Proof. For Part (i), from (14) and (15), we have

$$\begin{aligned}\frac{E_b}{N_0} &= J\sigma^2\cdot\frac{(2^{2\mathsf{SE}}-1)}{2\mathsf{SE}}\cdot 1\Big/\Big(T\Big[1+\eta^2\Big(\frac{J}{T}-1\Big)^2\Big]\Big)\\ &\geq -1.59dB + 10\log_{10} J\sigma^2\\ &\quad - 10\log_{10}\Big(T\Big[1+\eta^2\Big(\frac{J}{T}-1\Big)^2\Big]\Big)\ ,\end{aligned}$$

where the inequality is due to $\frac{2^{2\mathsf{SE}}-1}{2\mathsf{SE}}\geq -1.59dB$ as $\mathsf{SE}\to 0$.

For Part (ii), taking the limit of (15) as $\rho\to 0$, or as $J\to\infty$ for a given finite P_{tot}, we obtain

$$\begin{aligned}\lim_{J\to\infty}\frac{E_b}{N_0} &= \lim_{J\to\infty}\log 2\cdot\frac{\sigma^2\rho\Big(1+\rho T\Big[1+\eta^2\Big(\frac{J}{T}-1\Big)^2\Big]\Big)}{\rho T\eta^2 2\Big(\frac{J}{T}-1\Big)\frac{1}{T}}\\ &= \lim_{J\to\infty}\log 2\cdot\sigma^2\rho T\Big(\frac{J}{T}-1\Big)\ ,\end{aligned}$$

where the first step follows from L'Hôpital's rule, and the last step from $P_{tot}=J\sigma^2\rho$. □

Cor. 2 (Part (ii)) implies that if $P_{tot}=J\sigma^2\rho$ scales by a factor of A, then the SE curve for the sum-rate optimal model moves to the left by $10\log_{10}A$ dB, as indicated in Section IV (see Figure 5).

Proposition 2. TIN model — Chase combining of non-orthogonal transmissions. *The SE for CC-NOMA with TIN incorporating channel power gains is*

$$\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{CC,NOMA}} = \frac{1}{T}\sum_{t\in\mathcal{T}}\sum_{j\in\mathcal{J}_t}\frac{1}{2}\log_2\Big(1+\rho\Big(\sum_{t=1}^{T}|H_{tj}|^2\Big)^2\Big/ \Big(\sum_{t=1}^{T}|H_{tj}|^2+\rho\eta^2\Big|\sum_{t=1}^{T}\sum_{j'\in\mathcal{S}_{t,-j}}H_{tj}H_{tj'}^*\Big|^2\Big)\Big)\ bit/rdof\ , \tag{18}$$

where $|H_{tj}|^2$ is the channel power gain of user $j\in\mathcal{J}$ (the one with the largest SINR) at slot $t\in\mathcal{T}$.

The SNR per bit of CC-NOMA under TIN, using $\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{CC,NOMA}}$ in (18) is given as

$$\frac{E_b}{N_0} = \frac{J\sigma^2\rho n}{2n(\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{CC,NOMA}} - \Delta(\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{CC,NOMA}}, n, \epsilon))}\ . \tag{19}$$

Proof. See Appendix B in [44]. □

We next provide two lower bounds on E_b/N_0 for CC-NOMA under TIN for the IBL model.

Corollary 3. CC-NOMA under TIN. *In the IBL regime, the followings hold in the limit as $\rho\to 0$.*

(i) A lower bound on E_b/N_0. The SNR per bit for $|H_{tj}|=1$, $\forall\ t\in\mathcal{T},\ j\in\mathcal{J}_t$ satisfies

$$\frac{E_b}{N_0}\geq -1.59dB + 10\log_{10}\sigma^2\ . \tag{20}$$

(ii) Sensitivity of E_b/N_0 limit versus J for CC-NOMA under TIN. For a given finite J, the SNR per bit for $|H_{tj}|=1$, $\forall\ t\in\mathcal{T},\ j\in\mathcal{J}_t$ approaches the following lower bound:

$$\lim_{\rho\to 0}\frac{E_b}{N_0} = \log 2\cdot\sigma^2\ . \tag{21}$$

Proof. For Part (i) of the corollary, from (18) and (19), we

have

$$\frac{E_b}{N_0} = \frac{J\sigma^2\rho n}{n\frac{J}{T}\log_2\left(1+\frac{\rho T^2}{T+\rho\eta^2(J-T)^2}\right)} = \frac{J\sigma^2}{2\mathsf{SE}}\cdot\frac{\frac{1}{T}(2^{\frac{2T}{J}SE}-1)}{1-\eta^2\left(\frac{J}{T}-1\right)^2(2^{\frac{2T}{J}\mathsf{SE}}-1)},$$

where the last step follows from using $\frac{2^{2\mathsf{SE}}-1}{2\mathsf{SE}} \geq 10\log_{10}(\log 2) = -1.59dB$ as $\mathsf{SE}\to 0$.

For Part (ii), taking the limit of (19) as $\rho \to 0$, and incorporating that $P_{tot} = J\sigma^2\rho$, we obtain

$$\lim_{\rho\to 0}\frac{E_b}{N_0} = \lim_{\rho\to 0}\frac{J\sigma^2\rho n}{n\frac{J}{T}\log_2\left(1+\frac{\rho T^2}{T+\rho\eta^2(J-T)^2}\right)} = \lim_{\rho\to 0}\log 2\cdot\frac{T\sigma^2\rho}{\frac{\rho T^2}{T+\rho\eta^2\left(\frac{P_{tot}}{\sigma^2\rho}-T\right)^2}} = \lim_{\rho\to 0}\log 2\cdot\frac{\sigma^2}{T}\left(T+\rho\eta^2\left(\frac{P_{tot}}{\sigma^2\rho}-T\right)^2\right). \quad (22)$$

For a given finite J, in the limit as $\rho\to 0$, E_b/N_0 goes to $\log 2\cdot\sigma^2$, where P_{tot} goes to 0. □

From (22), for the IBL model, under a given finite P_{tot}, the value of ρ is inversely proportional to J, and when P_{tot} is held fixed, the E_b/N_0 limit scales with ρ^{-1}. When P_{tot} scales with J, Cor. 3 implies that the SNR per bit limit as $\mathrm{SE}\to 0$ for CC-NOMA under TIN is not sensitive to J and T, whereas from (18), SE for a given E_b/N_0 improves with T and a lower E_b/N_0 is indeed achievable.

C. Chase Combining with OMA-based Retransmissions

In this model, the retransmissions of each user are combined to enhance its received SNR. This scheme is a simplified version of CC-NOMA where the users have orthogonal messages, namely OMA with Chase combining or CC-OMA, which was introduced in [18]. We next provide its SE.

Proposition 3. Sum-rate optimal model — Chase combining of orthogonal transmissions. *The SE of CC-OMA for the sum-rate optimal model* incorporating channel power gains *is*

$$\mathsf{SE}_{\mathsf{sum}}^{\mathsf{CC,OMA}} = \frac{1}{2}\log_2\left(1+\rho\sum_{t\in\mathcal{T}}|H_{tj}|^2\left[1+\left(\sum_{t\in\mathcal{T}}|H_{tj}|^2\right)^{-2}\cdot\left(\sum_{t\in\mathcal{T}}\sum_{j'\in\mathcal{S}_{t,-j}}H_{tj}H_{tj'}^*\right)\right]\right)\ bit/rdof\ , \quad (23)$$

where user j is the one with the largest SINR.

The SNR per bit of CC-OMA for the sum-rate optimal model, using $\mathsf{SE}_{\mathsf{sum}}^{\mathsf{CC,OMA}}$ in (23) is given as

$$\frac{E_b}{N_0} = \frac{(J/T)\sigma^2\rho n}{2n(\mathsf{SE}_{\mathsf{sum}}^{\mathsf{CC,OMA}} - \Delta(\mathsf{SE}_{\mathsf{sum}}^{\mathsf{CC,OMA}}, n, \epsilon))}\ . \quad (24)$$

Proof. See Appendix C in [44]. □

Proposition 4. TIN model — Chase combining of orthogonal transmissions. *The SE of CC-OMA with TIN incorporating channel power gains is given as*

$$\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{CC,OMA}} = \frac{1}{T}\sum_{t\in\mathcal{T}}\sum_{j\in\mathcal{J}_t}\frac{1}{2}\log_2\left(1+\rho\sum_{t\in\mathcal{T}}|H_{tj}|^2\Big/\left(1+\rho\sum_{t\in\mathcal{T}}\sum_{j'\in\mathcal{S}_{t,-j}}|H_{tj}|^2|H_{tj'}|^2/\sum_{t\in\mathcal{T}}|H_{tj}|^2\right)\right)\ bit/rdof\ . \quad (25)$$

The SNR per bit of CC-OMA with TIN, using $\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{CC,OMA}}$ given in (25) is given as

$$\frac{E_b}{N_0} = \frac{(J/T)\sigma^2\rho n}{2n(\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{CC,OMA}} - \Delta(\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{CC,OMA}}, n, \epsilon))}\ . \quad (26)$$

Proof. See Appendix D in [44]. □

In the limit as $J\to\infty$, it holds that $\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{CC,OMA}} \leq \frac{T}{2\log 2}$. The subsequent result follows using the definition (9) for SNR per bit and the SE given in (25) for CC-OMA under the TIN model.

Corollary 4. CC-OMA under TIN. *At IBLs, when $|H_{tj}| = 1$ for all $t\in\mathcal{T}$, $j\in\mathcal{J}_t$, the SNR per bit E_b/N_0 of $\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{CC,OMA}}$ satisfies the lower bound given as*

$$\frac{E_b}{N_0} \geq -1.59dB + 10\log_{10}\sigma^2\ . \quad (27)$$

Proof. The SNR per bit of $\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{CC,OMA}}$ is given as

$$\frac{E_b}{N_0} = \frac{(J/T)\sigma^2\rho n}{2n\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{CC,OMA}}} = \frac{(J/T)\sigma^2\rho n}{n\frac{J}{T}\log_2\left(1+\frac{\rho T}{1+\rho\left(\frac{J}{T}-1\right)}\right)} = \frac{(J/T)\sigma^2}{2\mathsf{SE}}\cdot\frac{\frac{1}{T}(2^{\mathsf{SE}\frac{2T}{J}}-1)}{1-\frac{1}{T}\left(\frac{J}{T}-1\right)(2^{\mathsf{SE}\frac{2T}{J}}-1)} \geq -1.59dB + 10\log_{10}\sigma^2\ , \quad (28)$$

where the second equality follows from using (25) which yields $\rho = \frac{(2^{\mathsf{SE}\frac{2T}{J}}-1)}{T-\left(\frac{J}{T}-1\right)(2^{\mathsf{SE}\frac{2T}{J}}-1)}$, and the last step follows from the same intuition as in (16) and (19). □

From Cor. 4, it is clear that the SNR per bit limit as $\mathrm{SE}\to 0$ for CC-OMA under TIN is not sensitive to the parameters J or T. We refer the reader to Section IV (see e.g., Figures 3 and 5).

D. Incremental Redundancy with OMA-based Retransmissions

We next consider an incremental redundancy model with OMA (IR-OMA). From Sections III-B and III-C, due to the finite HARQ buffer size, the throughput of CC-NOMA is determined by the addition of all active users' signals at any given time slot. Unlike for CC, where the buffer size is the same as the number of packets per transmission, in IR-OMA, also known as HARQ Type III, the buffer size is equal to the number of coded bits of the total transmitted coded packets, where each retransmitted packet is self-decodable, and contains different information than the previous one.

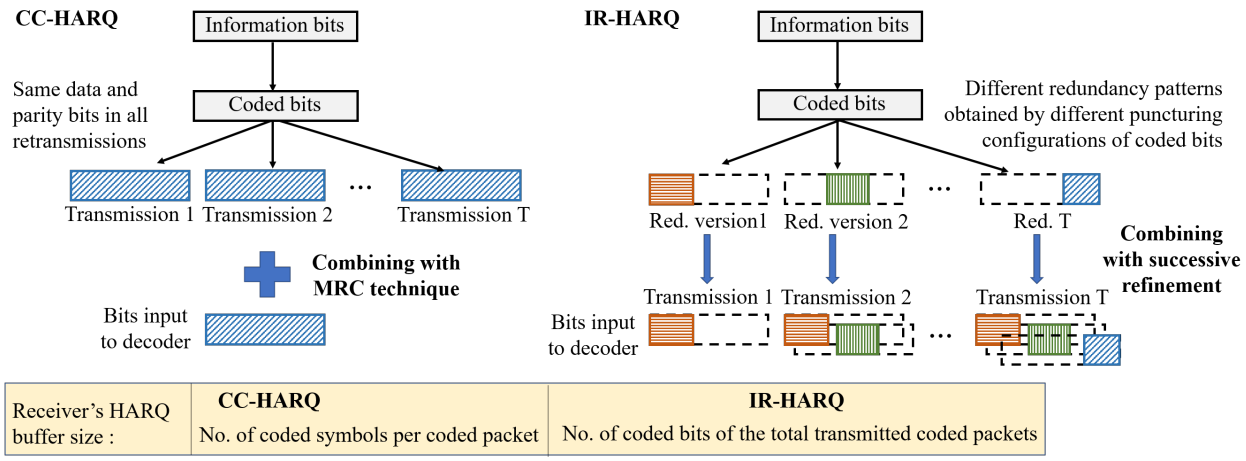

Fig. 2: (Left) CC-HARQ, where the retransmissions contain the same data and parity bits, which are summed at the receiver prior to decoding. (Right) IR-HARQ, where each retransmission provides some additional bits, and is self-decodable.

In IR-HARQ, each transmission consists of new redundancy bits from the channel encoder, which enables IR-OMA to achieve a superior performance over CC-OMA. However, IR has higher complexity due to additional signaling of retransmission numbers and a larger receiver buffer size [37]. IR is better suited for highly time-varying channels where the rate of the error control code is adapted to the current channel state. In IR-HARQ, multiple distinct sets of code bits are generated for the same information bits used in a packet, and transmitted under different channel conditions. These sets consist of distinct redundancy patterns obtained by different puncturing configurations of a common code. The rate adaptation achievable via puncturing reduces the decoder complexity. The transmitter and receiver only share a series of puncturing tables to specify which code bits are to be transmitted for a specific code rate [50]. The receiver then simply inserts erasures for all code bits that are not received. Punctured turbo codes are used for unequal error protection [51] and IR-HARQ, e.g., [52]–[54]. In Figure 2 (right), we sketch IR-HARQ, where each retransmitted packet provides some additional information bits, and is self-decodable, i.e., it provides successive refinement [38] by iteratively improving the rate-distortion tradeoff as more information is transmitted. Since we focus on the fixed-access strategy, the analysis of IR-NOMA would be relatively simple. Due to limited space, we only detail IR-OMA here.

a) Expected quantization distortion: Using the refinement-based approach in [38], the average quantization distortion is characterized as the mean squared error distortion between the quantized signal $\hat{\mathbf{Y}}_{t,T}$ and the received signal $\mathbf{Y}_t$. The quantized m dimensional signals are given by $\hat{\mathbf{Y}}_{t,T} = \mathbf{Y}_t + \mathbf{Q}_{t,T}$. The quantization noise satisfies $\mathbf{Q}_{t,T} \sim \mathcal{CN}\big(0, \frac{2\sigma_q^2(t,T)}{m} I_m\big)$, where $\sigma_q^2(t,T)$ represents the total quantization noise power per rdof (the quantization distortion per frequency bin is $\sigma_q^2(t,T)/B$) for IR-OMA at slot t given a total number of T retransmissions, where attempt t is unsuccessful if $1 \le t < T$ and the retransmission is successful at attempt T. From (1), $\mathbf{Y}_t$ has a dimension $m = n/T$.

Proposition 5. Sum-rate optimal model — Incremental redundancy of orthogonal transmissions. *The SE of IR-OMA for the sum-rate optimal model* incorporating channel power gains *is given as*

$$\mathsf{SE}_{\mathsf{sum}}^{\mathsf{IR,OMA}} = \sum_{t=1}^{T} \frac{B}{2} \log_2 \Big(1 + \rho \sum_{j \in \mathcal{J}_t} |H_{tj}|^2/B \Big/ \Big(1 + \sigma_q^2(t, T-1)/(Bm\sigma^2)\Big)\Big) \; bit/rdof/(T \; slots) \; , \quad (29)$$

where the following relation holds between the quantization noise $\sigma_q(t,T)$ and the buffer size C_{buf}:

$$\sigma_q^2(t,T) = \frac{B(J\rho/B+1)m\sigma^2}{2^{\frac{2C_{\mathsf{buf}}}{TB}} - 1}, \quad t < T \; , \quad (30)$$

and $\sigma_q^2(T,T) = 0$, i.e., at retransmission T, the received signal $\mathbf{Y}_T = \hat{\mathbf{Y}}_T$, i.e., the receiver recovers $\mathbf{Y}_T$. Furthermore, $\sigma_q^2(t,T-1)$ for $t \le T-1$ can be derived from (30), and $\sigma_q^2(T,T-1) = 0$.

The SNR per bit of IR-OMA for sum-rate optimal case, using $\mathsf{SE}_{\mathsf{sum}}^{\mathsf{IR,OMA}}$ given in (29), is given as

$$\frac{E_b}{N_0} = \frac{J\sigma^2\rho n}{2n(\mathsf{SE}_{\mathsf{sum}}^{\mathsf{IR,OMA}} - \Delta(\mathsf{SE}_{\mathsf{sum}}^{\mathsf{IR,OMA}}, n, \epsilon))/T} \; . \quad (31)$$

Proof. See Appendix E in [44]. □

Note that in Prop. 5, $\log_2\Big(1 + \frac{\rho J\sigma^2}{\sigma^2}\Big) \le \frac{B}{2}\log_2\Big(1 + \frac{\rho J\sigma^2/B}{\sigma^2}\Big)$, which follows from employing $\sum_{i=1}^{n} \log(1+x_i) \le n\log(1 + \frac{1}{n}\sum_{i=1}^{n} x_i)$ with $x_1 = 1$ and $x_i = 0$, $i \ne 1$. In this paper, we do not optimize B and the division of total transmit power across B bins, which is left as future work. Instead, in Section IV, we assume $B = 1$ to provide lower bounds on the performance tradeoffs.

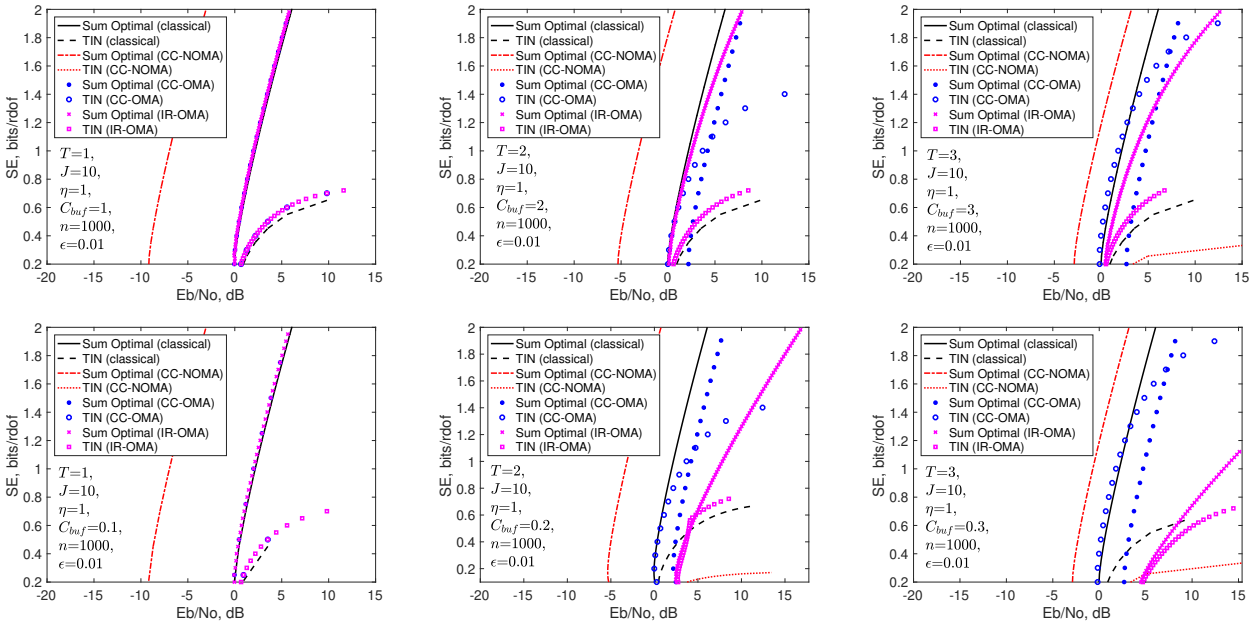

Fig. 3: Scaling of SE versus E_b/N_0 for varying T for $\eta = 1$ and $J = 10$. (Row I) moderate buffer size, $C_{\mathsf{buf}} = T$. (Row II) small buffer size, $C_{\mathsf{buf}} = 0.1T$.

As the buffer size $C_{\mathsf{buf}} \to \infty$, $\frac{E_b}{N_0} \to \frac{J\sigma^2\rho}{\log_2(1+\rho J)}$. Similarly, for smaller C_{buf}, $\frac{E_b}{N_0} > \frac{J\sigma^2\rho}{\log_2(1+\rho J)}$. Hence, it is easy to note that as C_{buf} increases the IR-OMA sum SE matches the sum SE for the classical problem without combining transmissions (sum-rate optimal case). However, when C_{buf} is small the gap between the SE for the classical transmission model and the IR-OMA sum SE grows as T increases.

Proposition 6. TIN model — Incremental redundancy of orthogonal transmissions. *The SE of IR-OMA with TIN* incorporating channel power gains *is*

$$\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{IR,OMA}} = \sum_{t=1}^{T}\sum_{j\in\mathcal{J}_t}\frac{B}{2}\log_2\Big(1+\rho|H_{tj}|^2/B\Big/\Big(\rho/B\sum_{j'\in\mathcal{S}_{t,-j}}|H_{tj'}|^2+1+\frac{\sigma_q^2(t,T-1)}{Bm\sigma^2}\Big)\Big)$$
$$bit/rdof/(T\ slots)\ , \quad (32)$$

where the following relation between the quantization noise $\sigma_q(t,T)$ and the buffer size C_{buf}:

$$\sigma_q^2(t,T) = \frac{B(\rho/B+1)m\sigma^2}{2^{\frac{2C_{\mathsf{buf}}}{TB}}-1} - \Big(\frac{J}{T}-1\Big)\rho m\sigma^2\ . \quad (33)$$

The SNR per bit of IR-OMA for TIN, using $\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{IR,OMA}}$ given in (32), is given as

$$\frac{E_b}{N_0} = \frac{J\sigma^2\rho n}{2^{\frac{n}{T}}(\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{IR,OMA}} - \Delta(\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{IR,OMA}}, n, \epsilon))}\ . \quad (34)$$

Proof. See Appendix F in [44]. □

We next provide two results on E_b/N_0 for the IBL regime for unit channel power gains.

Corollary 5. IR-OMA under TIN. *At IBLs, the followings hold for $|H_{tj}| = 1$, $\forall\ t \in \mathcal{T}$, $j \in \mathcal{J}_t$.*

(i) A limit on E_b/N_0 as $C_{\mathsf{buf}} \to \infty$. The SNR per bit in the limit as $C_{\mathsf{buf}} \to \infty$, approaches

$$\lim_{C_{\mathsf{buf}}\to\infty}\frac{E_b}{N_0} = \frac{T\sigma^2\rho}{\log_2\big(1+\frac{\rho}{\rho(J/T-1)+1}\big)}\ , \quad (35)$$

which is the SNR per bit for the classical transmission model with TIN in (13) for the IBL regime.

(ii) A limit on E_b/N_0 as $\rho \to 0$. For large buffer sizes C_{buf}, as $\rho \to 0$, it holds that

$$\lim_{\rho\to 0}\frac{E_b}{N_0} \approx \log 2 \cdot P_{tot}\ . \quad (36)$$

Proof. For Part (i) of the corollary, from Prop. 6, as $C_{\mathsf{buf}} \to \infty$, $\zeta_t = (J/T - J/(T-1))$ for $t < T$, and $\zeta_T = (J/T-1)$, and the SNR per bit for IR-OMA with TIN approaches

$$\lim_{C_{\mathsf{buf}}\to\infty}\frac{E_b}{N_0} = \frac{T\sigma^2\rho}{\frac{1}{T}\sum_{t=1}^{T}\log_2\big(1+\frac{\rho}{\rho\zeta_t+1}\big)} \le \frac{T\sigma^2\rho}{\log_2\big(1+\frac{\rho}{\rho(J/T-1)+1}\big)}\ , \quad (37)$$

where the inequality in the second step is indeed an equality because $\sigma_q^2(t,T) \to 0$ as $C_{\mathsf{buf}} \to \infty$, where (33) no longer holds, and the desired bit-error-rate (BER) is met from (32).

Part (ii) of the corollary is immediate from Prop. 6. □

From Cor. 5, the behavior of SE versus E_b/N_0 is highly affected by C_{buf}, and to achieve the same SE, it is required to have a larger E_b/N_0 when C_{buf} is low. In the regime as $C_{\mathsf{buf}} \to \infty$, we can observe from (35) that the SNR per bit $\frac{E_b}{N_0}$ for IR-OMA under TIN behaves similarly as the classical transmission model with TIN in (13). For smaller

This article has been accepted for publication in IEEE Transactions on Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2024.3361550

SE	$\frac{\partial}{\partial T}$SE	$\frac{\partial}{\partial \eta}$SE
$\mathsf{SE}_{\mathsf{sum}}^{\mathsf{Clas.}} = \frac{1}{2}\log_2\left(1+\rho\frac{J}{T}\right)$	<0	NA
$\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{Clas.}} = \frac{J}{2T}\log_2\left(1+\frac{\rho}{\rho(\frac{J}{T}-1)+1}\right)$	<0	NA
$\mathsf{SE}_{\mathsf{sum}}^{\mathsf{CC,NOMA}} = \frac{1}{2}\log_2\left(1+\rho T\left[1+\eta^2\left(\frac{1}{T}\sum_{t=1}^{T}J_t-1\right)^2\right]\right)$	>0	>0
$\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{CC,NOMA}} = \frac{J}{2T}\log_2\left(1+\frac{\rho T^2}{T+\rho\eta^2(J-T)^2}\right)$	<0 at high SE, >0 at low SE	<0
$\mathsf{SE}_{\mathsf{sum}}^{\mathsf{CC,OMA}} = \frac{1}{2}\log_2\left(1+\rho T\left[1+\frac{1}{T}\left(\frac{J}{T}-1\right)\right]\right)$	>0 at high SE, <0 at low SE	NA
$\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{CC,OMA}} = \frac{J}{2T}\log_2\left(1+\frac{\rho T}{1+\rho\left(\frac{J}{T}-1\right)}\right)$	>0	NA
$\mathsf{SE}_{\mathsf{sum}}^{\mathsf{IR,OMA}} = \frac{B}{2}\log_2(1+\rho J/B)+\sum_{t=1}^{T-1}\frac{B}{2}\log_2\left(1+\frac{\rho J/B}{1+\sigma_q^2(t,T-1)/(Bm\sigma^2)}\right)$	<0	NA
$\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{IR,OMA}} = \frac{JB}{2T}\log_2\left(1+\frac{\rho/B}{\rho/B(J/T-1)+1}\right)+\frac{JB(T-1)}{2T}\log_2\left(1+\frac{\rho/B}{1+\rho/B\zeta_t}\right)$	>0 at high C_{buf}, <0 at low C_{buf}	NA

TABLE I: The SE of the different retransmission-based models with combining under unity channel power gain.

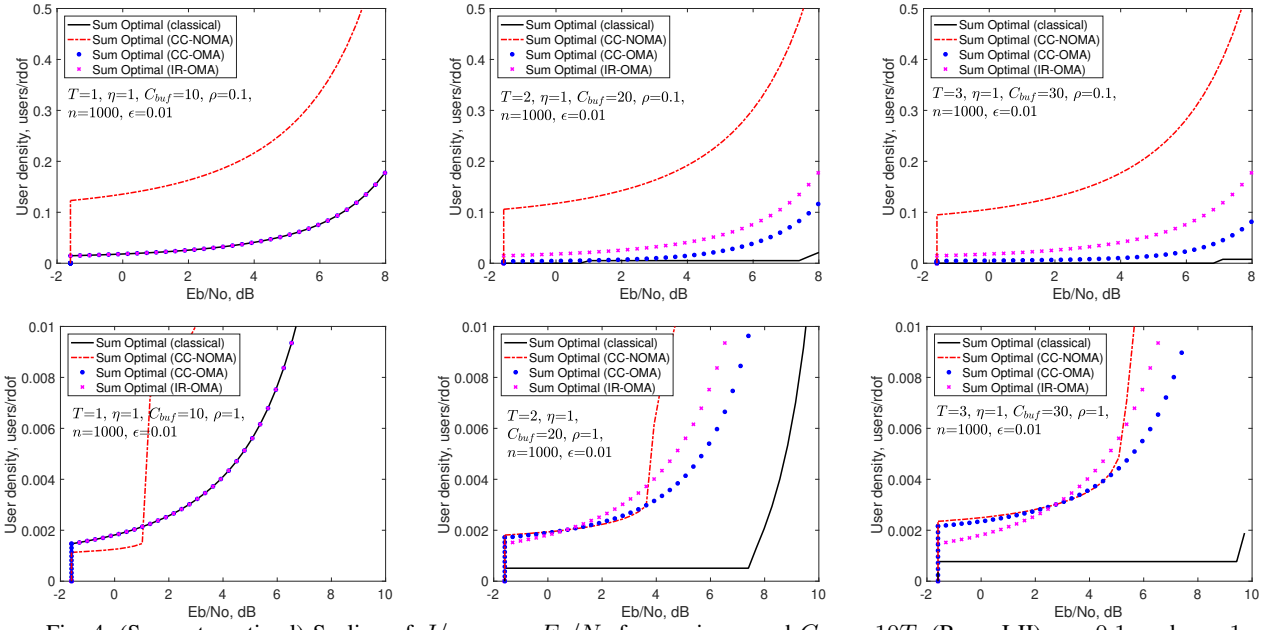

Fig. 4: (Sum-rate optimal) Scaling of J/n versus E_b/N_0 for varying ρ and $C_{\mathsf{buf}} = 10T$. (Rows I-II) $\rho = 0.1$, and $\rho = 1$.

C_{buf}, the ratio $\frac{E_b}{N_0}$ is typically higher than the SNR per bit for the classical TIN case. We will illustrate this behavior in Section IV (see Figures 3 and 5).

In Table I, we summarize the SE of different retransmission-based models (with unit channel power gain), with two additional columns describing the behavior of SE with respect to T and η. Note also that $\frac{\partial}{\partial \rho}\text{SE} > 0$ for all models. For the different HARQ models in hand, next in Section IV, we will study the SE versus E_b/N_0 and J/n versus E_b/N_0 tradeoffs by exploiting the joint behavior of SE, E_b/N_0, and ρ, for the FBL regime (for fixed n and ϵ). For the IBL models, we refer the reader to [1].

IV. Numerical Evaluation of Scaling Results

In this section, exploiting our findings in Sections II and III, we first study the SE (bits/rdof) versus the E_b/N_0 (dB) tradeoff (see Figures 3 and 5) for the different HARQ-based retransmission combining models for the sum-rate optimal and the TIN schemes detailed in Section III, as function of the design parameters T, η, C_{buf}, and J. Our numerical results (in Figures 3 and 5) are for the FBL regime (with $n = 1000$, $\epsilon = 0.01$), approximating the actual scaling behaviors for the SE models for each strategy. We then focus on the scaling behavior of the user density J/n with respect to E_b/N_0 (dB) (see Figures 4, 6, and 7) as function of T, η, C_{buf}, and ρ under unity channel power gain. For various regimes of interest, we indicate the set of chosen parameters of the sum-rate optimal and the TIN schemes in the legend on each plot. Based on the numerical experiments run for different values of T, η, C_{buf}, and J (for fixed values of n, J, and ϵ), we next present our observations (see Figures 3 to 7).

a) Number of retransmissions T: From Cor. 1, we note that both $\mathsf{SE}_{\mathsf{sum}}^{\mathsf{Clas.}}$ and $\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{Clas.}}$ are constant for fixed J/T. From (14) and Cor. 2, increasing T causes degradation in $\mathsf{SE}_{\mathsf{sum}}^{\mathsf{CC,NOMA}}$. From Prop. 2 and Cor. 3, $\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{CC,NOMA}}$ improves

This article has been accepted for publication in IEEE Transactions on Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2024.3361550

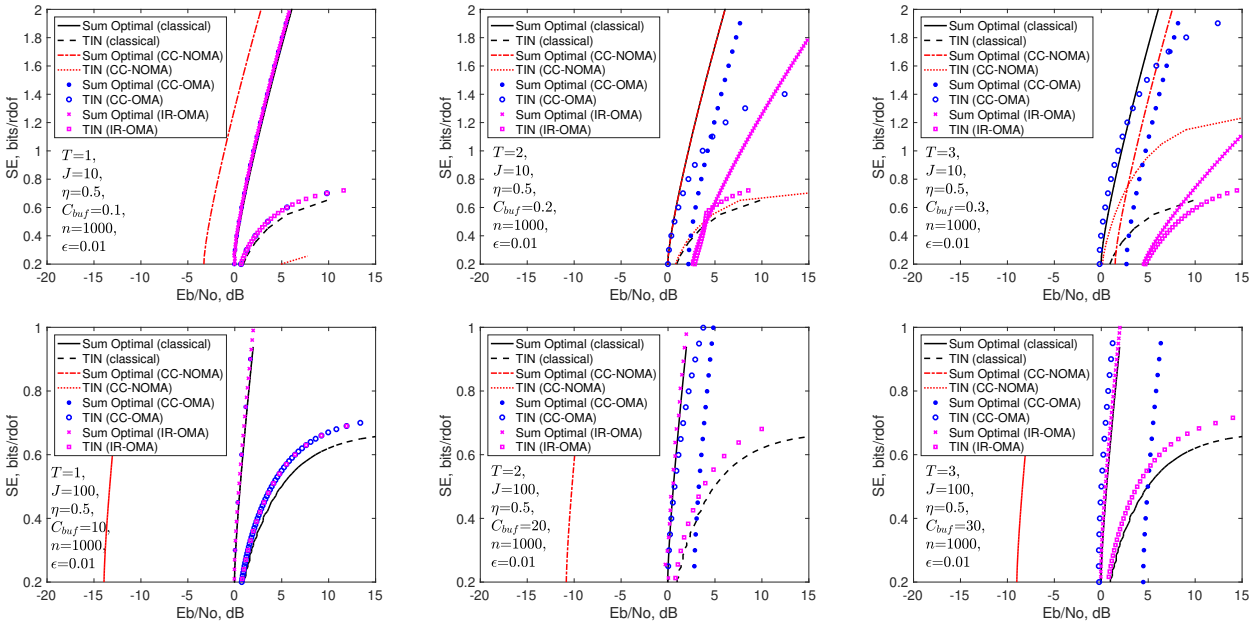

Fig. 5: Scaling of SE versus E_b/N_0. (Row I) $\eta = 0.5$, $J = 10$, $C_{\mathsf{buf}} = 0.1T$. (Row II) $\eta = 0.5$, $J = 100$, $C_{\mathsf{buf}} = 10T$.

with T. From (23) $\mathsf{SE}_{\mathsf{sum}}^{\mathsf{CC,OMA}}$ also decays with T, which we can observe from Figures 3 and 5. A competitor strategy in terms of SE is $\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{CC,OMA}}$, which improves with T and can even outperform $\mathsf{SE}_{\mathsf{sum}}^{\mathsf{CC,OMA}}$, which follows from Prop. 4. By looking at the partial derivatives of $\mathsf{SE}_{\mathsf{sum}}^{\mathsf{CC,OMA}}$ in (23) and $\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{CC,OMA}}$ in (25) with respect to ρ, at low E_b/N_0 values, TIN-based CC-OMA can perform better than sum-rate optimal CC-OMA because the interference term in $\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{CC,OMA}}$ becomes low whereas $\mathsf{SE}_{\mathsf{sum}}^{\mathsf{CC,OMA}}$ decreases in T for large J, as can be observed from Figure 3. From Figures 3 and 5, at high E_b/N_0 values, $\mathsf{SE}_{\mathsf{sum}}^{\mathsf{CC,OMA}}$ is better than $\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{CC,OMA}}$, and $\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{CC,OMA}}$ is higher than $\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{IR,OMA}}$ with increasing gains in T. Overall, these scaling results show that for any given value of E_b/N_0, the best performance is attained by $\mathsf{SE}_{\mathsf{sum}}^{\mathsf{CC,NOMA}}$, in general for small T. Furthermore, at low E_b/N_0, the gap between $\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{CC,OMA}}$ and $\mathsf{SE}_{\mathsf{sum}}^{\mathsf{CC,NOMA}}$ becomes smaller, and hence, the performance of $\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{CC,OMA}}$ can approach or outperform $\mathsf{SE}_{\mathsf{sum}}^{\mathsf{CC,NOMA}}$, and $\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{CC,OMA}}$ also outperforms the other TIN-based approaches because it exploits Chase combining and unlike $\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{CC,NOMA}}$ it has lowered interference.

b) Finite buffer size C_{buf} at the decoder: The SNR per bit values under TIN for the classical model in (13) and the IR-OMA model (Prop. 5) have a matching fundamental E_b/N_0 limit when C_{buf} is sufficiently large for $\rho = 0$ for any given T. For large C_{buf}, while $\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{IR,OMA}}$ is higher than $\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{Clas.}}$, the behavior of $\mathsf{SE}_{\mathsf{sum}}^{\mathsf{Clas.}}$ and $\mathsf{SE}_{\mathsf{sum}}^{\mathsf{IR,OMA}}$ schemes are similar, and similarly, for $\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{Clas.}}$ and $\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{IR,OMA}}$, see e.g., Figure 3 (Row I). As C_{buf} decays, implying a lower SE, the performance of IR-OMA degrades both for the sum-rate optimal and TIN scenarios, as explained in Section III-D, see e.g., Figure 3 (Row II). At small C_{buf}, the SNR per bit of IR-OMA under TIN is higher than the classical TIN model. Intuitively, for any given T, at small C_{buf}, the curves for $\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{IR,OMA}}$ and $\mathsf{SE}_{\mathsf{sum}}^{\mathsf{IR,OMA}}$ start to overlap. For large C_{buf}, when $\sigma_q(t,T)$ becomes negligible, $\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{IR,OMA}}$ is approximately the same as $\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{Clas.}}$, and $\mathsf{SE}_{\mathsf{sum}}^{\mathsf{IR,OMA}}$ approaches that of the $\mathsf{SE}_{\mathsf{sum}}^{\mathsf{Clas.}}$, as expected. The evaluations indicate that while $\mathsf{SE}_{\mathsf{sum}}^{\mathsf{CC,NOMA}}$ outperforms the other strategies almost in all regimes, and $\mathsf{SE}_{\mathsf{sum}}^{\mathsf{CC,OMA}}$ is less sensitive to C_{buf}, at high C_{buf}, $\mathsf{SE}_{\mathsf{sum}}^{\mathsf{IR,OMA}}$ competes with $\mathsf{SE}_{\mathsf{sum}}^{\mathsf{CC,NOMA}}$ and $\mathsf{SE}_{\mathsf{sum}}^{\mathsf{CC,OMA}}$, yet $\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{IR,OMA}}$ is only slightly above $\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{Clas.}}$.

c) Contrasting SE versus SNR per bit and non-orthogonality of transmissions measured via η: For small C_{buf}, i.e., under high quantization noise, as T increases, we expect the SE of IR-OMA (both the sum-rate optimal and TIN models), where each retransmission successively refines the information, to be a lower bound to CC-NOMA and CC-OMA and classical models for $T > 1$. On the other hand, for large C_{buf}, $\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{IR,OMA}}$ can perform superior to $\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{CC,NOMA}}$ when interference is high, e.g., if $\eta = 1$ or J is high, from (18) and (35), and $\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{IR,OMA}}$ can perform similarly to $\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{Clas.}}$, from the equivalence of (12) to (32) as $C_{\mathsf{buf}} \to \infty$. For large C_{buf}, $\mathsf{SE}_{\mathsf{sum}}^{\mathsf{IR,OMA}}$ performs inferior to $\mathsf{SE}_{\mathsf{sum}}^{\mathsf{CC,NOMA}}$ and, in general, better than $\mathsf{SE}_{\mathsf{sum}}^{\mathsf{CC,OMA}}$, and could outperform $\mathsf{SE}_{\mathsf{sum}}^{\mathsf{CC,NOMA}}$ for small η, which follows from contrasting (14) and (29). Decreasing η in CC-NOMA reduces the interference and improves $\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{CC,NOMA}}$ via combining retransmissions, as illustrated in Figure 5. On the other hand, $\mathsf{SE}_{\mathsf{sum}}^{\mathsf{CC,NOMA}}$ degrades. For larger T and C_{buf}, which requires a higher ρ as could be inferred from (23) and (29), $\mathsf{SE}_{\mathsf{sum}}^{\mathsf{IR,OMA}}$ could be lower than $\mathsf{SE}_{\mathsf{sum}}^{\mathsf{CC,OMA}}$ (and similarly for $\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{IR,OMA}}$ versus $\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{CC,OMA}}$), and $\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{CC,OMA}}$ can be superior to $\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{Clas.}}$. In general, for the sum-rate optimal models, in the bandwidth-limited regime (high SNR), the SE

This article has been accepted for publication in IEEE Transactions on Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2024.3361550

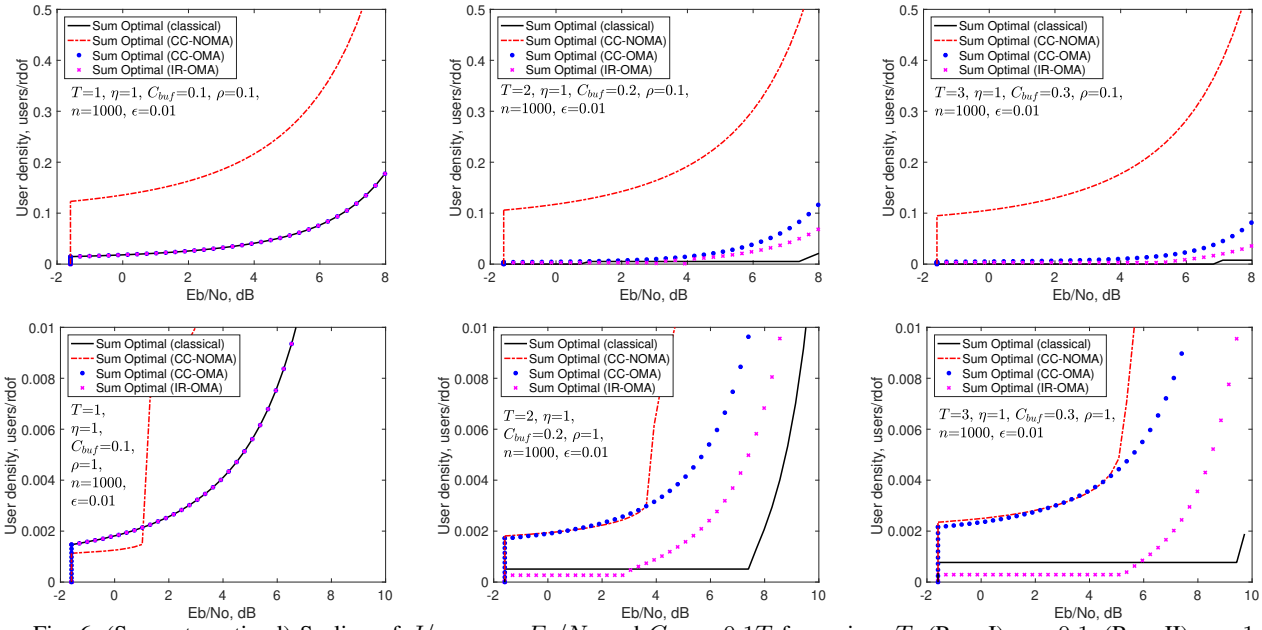

Fig. 6: (Sum-rate optimal) Scaling of J/n versus E_b/N_0 and $C_{\mathsf{buf}} = 0.1T$ for various T. (Row I) $\rho = 0.1$, (Row II) $\rho = 1$.

is less sensitive to the changes in ρ versus the power-limited (low SNR) region [45, Ch. 5].

d) Scaling of the SE with J: Increasing J moves the curve of $\mathsf{SE}_{\mathsf{sum}}^{\mathsf{CC,NOMA}}$ to the left (see Figure 5), which coincides with the discussion after Cor. 2. Similarly, the SEs of the other sum-rate optimal models also become steeper. On the other hand, for the TIN models, the SE drops in J, e.g., with unitary channel gains, from (18), $\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{CC,NOMA}}$ scales as $T/(J\log 2)$ as $J \to \infty$, and $J/T\log_2(1+\rho T)$ as $J \to T$, i.e., no interference. The trend of SE is similar for $\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{Clas.}}$, $\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{CC,OMA}}$, and $\mathsf{SE}_{\mathsf{TIN}}^{\mathsf{IR,OMA}}$.

While retransmissions are inevitable in HARQ-based protocols, they generally degrade the performances of SE and J/n versus E_b/N_0. Our numerical results on J/n versus E_b/N_0 for the sum-rate optimal models in Figures 4 and 6 and the TIN models in Figure 7 at FBLs, provide approximations of the actual scaling behaviors. We next detail them, reminding the reader that $n = 1000$ and $\epsilon = 0.01$.

e) Scaling of user density J/n versus E_b/N_0: We investigate this scaling behavior in Figures 4-7 as a function of ρ. A stricter average probability of error requirement at a decoder is equivalent to a high ρ value, which yields a higher E_b/N_0 to achieve the same J/n. As T increases, the supported density J/n drops. As ρ increases, the user density scalings of different models for the sum-rate optimal strategy become more similar under high C_{buf}. This situation is because the growth of E_b/N_0 is not much sensitive to ρ at low ρ and the approximate growth rate for the sum-rate optimal models is $\frac{\rho}{\log_2 \rho}$ for high ρ, which causes a significant drop in J/n. With the conventional MFR, the optimal SE for the sum-rate optimal model cannot be accurately captured at high ρ [11]. However, we might not observe this behavior for the TIN models under high C_{buf} (see Figure 7 (Row II)). For CC-NOMA under TIN, from (19), E_b/N_0 roughly grows with $\eta^2(J-T)^2/T$ at low ρ, and the scaling is subquadratic at high ρ, for CC-OMA from (28), E_b/N_0 grows with J/T at low ρ, and the scaling becomes sublinear at high ρ. For the classical TIN model from (13), and similarly for IR-OMA with TIN at high C_{buf}, E_b/N_0 grows linearly with J at low ρ, and J/T is not much sensitive to E_b/N_0 at high ρ.

We next compare different TIN models. At low ρ values, the scalings of J/n versus E_b/N_0 for CC-NOMA and CC-OMA improve similarly, whereas the schemes that do not promote combining do not perform well. However, at higher ρ values, the scaling of the CC-NOMA scheme deteriorates due to high interference, whereas CC-OMA performs the best because it combines retransmissions while not being susceptible to interference. The TIN-based CC-NOMA and CC-OMA models improve the user density scaling by increasing T that causes diminishing returns in gains, and scaling for the IR-OMA and the classical models are not robust to retransmissions. From SNR per bit of the classical sum-rate optimal model in (11), and exploiting the SNR per bit for the other sum-rate optimal models, which are given for CC-NOMA in (15), for CC-OMA in (24), and for IR-OMA in (31), when $T = 1$, the classical model, CC-OMA, and IR-OMA behave the same, and CC-NOMA has a lower E_b/N_0 than the classical sum-rate optimal approach, CC-OMA, and IR-OMA for a given J/n. This behavior can be observed in Figure 4 for $T = 1$. For $T = 2$ and $T = 3$, for the classical sum-rate optimal approach, the effective J/n for a given E_b/N_0 decays as a function of T, following from (11), which is similar for the other sum-rate optimal models, namely CC-NOMA, CC-OMA, and IR-OMA. However, the scaling of these three models is less sensitive than the

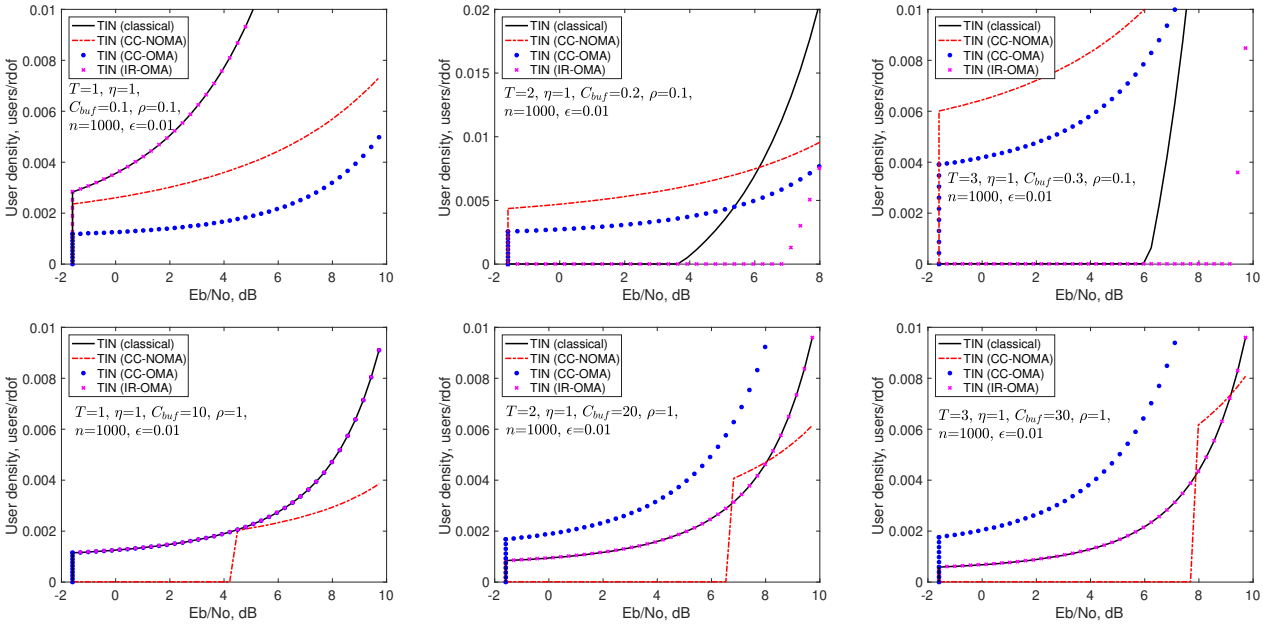

Fig. 7: (TIN) J/n versus E_b/N_0 for varying ρ and C_{buf}. (Row I) $\rho = 0.1$, $C_{\mathsf{buf}} = 0.1T$, (Row II) $\rho = 1$, $C_{\mathsf{buf}} = 10T$.

classical model, indicating their robustness. From the most to the least sensitive as an increasing function of T, the ordering for the user density scalings of the models in the sum-rate optimal regime is classical, CC-OMA, IR-OMA, and CC-NOMA.

We observe from Figure 7 for TIN-based models that the different models we considered in this paper perform the best at low SE. Increasing ρ decreases the scaling performance of J/n for IR-OMA, CC-NOMA, and CC-OMA. To compensate for the loss of CC-NOMA, even though we can incorporate better coding signatures to enable lower η, this model still requires a higher minimum SNR per bit versus the other models with a higher sensitivity to ρ. From Part (ii) of Cor. 3, for CC-NOMA with the TIN model, the SNR per bit E_b/N_0 limit to ensure a nonzero user density increases with $\rho\eta^2$, and it is easy to notice from Figure 7 (Row II) that this limit could indeed be very high (i.e., > 8 dB). The conventional MFR approach is ideal for the low SINR regime, and with MFR, the characterization for the TIN-based might be suboptimal at high ρ [11]. As T increases, achieving a superior number of users per rdof and a better scaling for IR-OMA via increasing C_{buf} is possible. At higher C_{buf} (or when $\rho \geq 10$), IR-OMA yields a better performance over CC-OMA, where CC-OMA scales better due to the combining of transmissions as given by the SNR per bit in the first step of (28) than IR-OMA with an SNR per bit in (34) versus vice versa for lower C_{buf} (or when $\rho \leq 1$).

V. Conclusions

In this paper, we proposed eight HARQ-based grant-based access models for 5G wireless communication networks: (i) the classical scheme with no retransmissions, and the retransmission-based schemes using different combining techniques at the receiver, namely (ii) CC-NOMA, (iii) CC-OMA, and (iv) IR-OMA, both for the sum-rate optimal and TIN-based strategies. For each model, we characterized the tradeoffs for SE versus SNR per bit, and the user density versus SNR per bit, and demonstrated through numerical simulations that retransmissions can improve the scaling behaviors of SE and the user density. Our results indicate that sum-rate optimal CC-NOMA provides the best scaling almost in all regimes, and at low SNR per bit, the performance of TIN-based CC-OMA, which outperforms the TIN-based classical and IR-OMA approaches with increasing T via exploiting CC and providing reduced interference, can attain the best performance. Furthermore, at high C_{buf}, the SE performance of IR-OMA approximates CC-NOMA and CC-OMA under the sum-rate optimal models. At low ρ values, the user densities of CC-NOMA and CC-OMA improve similarly, whereas the schemes that do not promote combining do not perform well. At high ρ values, as interference is more dominant, the scaling of the CC-NOMA-based scheme deteriorates, whereas CC-OMA performs the best because it combines retransmissions and has less interference. The ordering of the J/n versus E_b/N_0 performances of the models, from the most to the least sensitive as an increasing function of T – which degrade in T – in the sum-rate optimal regime, is classical, CC-OMA, IR-OMA, and CC-NOMA. Comparing different TIN models at low ρ values, the scalings of J/n versus E_b/N_0 for CC-NOMA and CC-OMA improve similarly, whereas the schemes without combining do not perform well.

Critical future directions include incorporating feedback and optimizing the number of retransmissions T and the number of frequency bins B. From a resource-allocation

perspective, handling the issues of identification of user IDs, asynchrony, and traffic burstiness are of critical importance and left as future work. Another direction is the joint design of the physical and network layer aspects. It is crucial to support heterogeneous traffic type requirements on one platform where distinct classes of users are under different SINR requirements. Power allocation for the cell edge versus cell center users could be different to mitigate the interference, and capacity model could be revisited under general power control mechanisms. Furthermore, the MMSE receiver is superior than the conventional MFR over a wide range of SIRs [11], which makes it more suitable under multiple traffic types.

The generalization of the classical capacity models is of primary interest through incorporating path loss, and investigating the outage capacity exploiting the fading distribution for the asymptotic (IBL) and FBL models, as well as techniques to achieve optimal performance for the SU and the MU settings, and for joint decoding of users. This will pave the way for understanding the 3-way tradeoff between SE, E_b/N_0, and L. In addition, the BER performance for the MU NOMA model depends on the modulation and coding scheme. The study of the probability of error (per-user or for all users) achieved for a given user density, payload, and energy, is left as future work.

REFERENCES

[1] D. Malak, "Throughput and energy tradeoffs for retransmission-based random access protocols," in *Proc., IEEE WiOpt*, Turin, Italy, Sep. 2022.

[2] M. Danieli *et al.*, "Maximum mutual information vector quantization of log-likelihood ratios for memory efficient HARQ implementations," in *Proc., Data Compression Conf.*, Snowbird, UT, Mar. 2010.

[3] Y. Polyanskiy, "Information theoretic perspective on massive multiple-access," *Short Course (slides) Skoltech Inst. of Tech., Moscow, Russia*, Jul. 2018.

[4] N. Abramson, "The ALOHA system: Another alternative for computer communications," in *Proc., AFIPS*, Nov. 1970, pp. 281–285.

[5] L. G. Roberts, "ALOHA packet system with and without slots and capture," *ACM SIGCOMM Computer Commun. Review*, vol. 5, no. 2, pp. 28–42, Apr. 1975.

[6] W. Crowther, R. Rettberg, D. Walden, S. Ornstein, and F. Heart, "A system for broadcast communication: Reservation-ALOHA," in *Proc., Hawaii Int. Conf. Syst. Sci*, Honolulu, HI, Jan. 1973, pp. 596–603.

[7] G. C. Madueno, Č. Stefanović, and P. Popovski, "Efficient LTE access with collision resolution for massive M2M communications," in *Proc., IEEE Globecom Wkshps*, Dec. 2014, pp. 1433–1438.

[8] W. Yu, "On the fundamental limits of massive connectivity," in *Proc., Inf. Theory and Apps. Wkshp*, San Diego, CA, Feb. 2017.

[9] A. Bayesteh, E. Yi, H. Nikopour, and H. Baligh, "Blind detection of SCMA for uplink grant-free multiple-access," in *Proc., Int. Symp. Wireless Commun. Systems*, Aug. 2014, pp. 853–857.

[10] S. Verdú and S. Shamai, "Spectral efficiency of CDMA with random spreading," *IEEE Trans. Inf. Theory*, vol. 45, pp. 622–640, Mar. 1999.

[11] D. N. C. Tse and S. V. Hanly, "Linear multiuser receivers: Effective interference, effective bandwidth and user capacity," *IEEE Trans. Inf. Theory*, vol. 45, no. 2, pp. 641–657, Mar. 1999.

[12] M. Vaezi, R. Schober, Z. Ding, and H. V. Poor, "Non-orthogonal multiple access: Common myths and critical questions," *IEEE Wireless Commun.*, vol. 26, no. 5, pp. 174–180, Sep. 2019.

[13] S. S. Kowshik and Y. Polyanskiy, "Fundamental limits of many-user MAC with finite payloads and fading," *IEEE Trans. Inf. Theory*, vol. 67, no. 9, pp. 5853–5884, Jun. 2021.

[14] X. Chen, T.-Y. Chen, and D. Guo, "Capacity of Gaussian many-access channels," *IEEE Trans. Inf. Theory*, vol. 63, pp. 3516–39, Feb. 2017.

[15] S. S. Kowshik and Y. Polyanskiy, "Quasi-static fading MAC with many users and finite payload," in *Proc., IEEE ISIT*, Paris, France, Jul. 2019, pp. 440–444.

[16] L. Ozarow, "The capacity of the white Gaussian multiple access channel with feedback," *IEEE Trans. Inf. Theory*, vol. 30, no. 4, pp. 623–629, Jul. 1984.

[17] R. C. Yavas, V. Kostina, and M. Effros, "Gaussian multiple and random access channels: Finite-blocklength analysis," *IEEE Trans. Inf. Theory*, vol. 67, no. 11, pp. 6983–7009, Sep. 2021.

[18] D. Malak, H. Huang, and J. G. Andrews, "Throughput maximization for delay-sensitive random access communication," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 709–723, Dec. 2018.

[19] ——, "Fundamental limits of random access communication with retransmissions," in *Proc., IEEE ICC*, May 2017, pp. 1–7.

[20] H. Xu, "LTE IoT is starting to connect the massive IoT today, thanks to eMTC and NB-IoT," https://www.qualcomm.com/news/onq/2017/06/lte-iot-starting-connect-massive-iot-today-thanks-emtc-and-nb-iot, Jun. 2017.

[21] K. Dovelos, L. Toni, and P. Frossard, "Finite length performance of random MAC strategies," in *Proc., IEEE ICC*, Paris, France, May 2017.

[22] M. Koseoglu, "Pricing-based load control of M2M traffic for the LTE-A random access channel," *IEEE Trans. Commun.*, vol. 65, no. 3, pp. 1353–1365, Dec. 2016.

[23] H. S. Dhillon, H. C. Huang, H. Viswanathan, and R. A. Valenzuela, "Power-efficient system design for cellular-based machine-to-machine communications," *IEEE Trans. Wireless Commun.*, vol. 12, no. 11, pp. 5740–5753, Oct. 2013.

[24] T. Han and K. Kobayashi, "A new achievable rate region for the interference channel," *IEEE Trans. Inf. Theory*, vol. 27, no. 1, pp. 49–60, Jan. 1981.

[25] V. R. Cadambe, S. A. Jafar, and S. Shamai, "Interference alignment on the deterministic channel and application to fully connected Gaussian interference networks," *IEEE Trans. Inf. Theory*, vol. 55, no. 1, pp. 269–274, Dec. 2008.

[26] C. Huang, V. R. Cadambe, and S. A. Jafar, "Interference alignment and the generalized degrees of freedom of the X channel," *IEEE Trans. Inf. Theory*, vol. 58, no. 8, pp. 5130–5150, May 2012.

[27] R. H. Etkin and E. Ordentlich, "The degrees-of-freedom of the K-user Gaussian interference channel is discontinuous at rational channel coefficients," *IEEE Trans. Inf. Theory*, vol. 55, no. 11, pp. 4932–4946, Oct. 2009.

[28] V. R. Cadambe and S. A. Jafar, "Interference alignment and degrees of freedom of the k-user interference channel," *IEEE Trans. Inf. Theory*, vol. 54, no. 8, pp. 3425–3441, Jul. 2008.

[29] M. J. Ahmadi and T. M. Duman, "Random spreading for unsourced MAC with power diversity," *IEEE Commun. Letters*, vol. 25, no. 12, pp. 3995–99, Oct. 2021.

[30] K. Hsieh, C. Rush, and R. Venkataramanan, "Near-optimal coding for many-user multiple access channels," *IEEE J. Sel. Areas Inf. Theory*, vol. 3, no. 1, pp. 21–36, Mar. 2022.

[31] H. Liu *et al.*, "A new rate splitting strategy for uplink CR-NOMA systems," *IEEE Trans. Vehicular Tech.*, Apr. 2022.

[32] M. Elhattab, M. A. Arfaoui, C. Assi, A. Ghrayeb, and M. Qaraqe, "On optimizing the power allocation and the decoding order in uplink cooperative NOMA," *arXiv preprint arXiv:2203.13100*, Mar. 2022.

[33] S. Sesia, "Techniques de codage avancées pour la communication sans fil, dans un système point à multipoint," Ph.D. dissertation, Télécom ParisTech, 2005.

[34] W. Lee, O. Simeone, J. Kang, S. Rangan, and P. Popovski, "HARQ buffer management: An information-theoretic view," *IEEE Trans. Commun.*, vol. 63, no. 11, pp. 4539–4550, Aug. 2015.

[35] W. Yafeng, Z. Lei, and Y. Dacheng, "Performance analysis of type III HARQ with turbo codes," in *Proc., IEEE Vehicular Tech. Conf.*, vol. 4, Apr. 2003, pp. 2740–2744.

[36] D. N. Rowitch and L. B. Milstein, "On the performance of hybrid FEC/ARQ systems using rate compatible punctured turbo (RCPT) codes," *IEEE Trans. Commun.*, vol. 48, no. 6, pp. 948–959, Jun. 2000.

[37] P. Frenger, S. Parkvall, and E. Dahlman, "Performance comparison of HARQ with chase combining and incremental redundancy for HSDPA," in *Proc., IEEE Vehicular Tech. Conf.*, Oct. 2001.

[38] W. H. Equitz and T. M. Cover, "Successive refinement of information," *IEEE Trans. Inf. Theory*, vol. 37, pp. 269–275, Mar. 1991.

[39] V. Kostina and E. Tuncel, "Successive refinement of abstract sources," *IEEE Trans. Inf. Theory*, vol. 65, pp. 6385–98, Jun. 2019.

[40] J. Østergaard, U. Erez, and R. Zamir, "Incremental refinements and multiple descriptions with feedback," *IEEE Trans. Inf. Theory*, vol. 68, pp. 6915–40, May 2022.

[41] T. Helleseth, D. J. Katz, and C. Li, "The resolution of Niho's last conjecture concerning sequences, codes, and Boolean functions," *IEEE Trans. Inf. Theory*, vol. 67, no. 10, pp. 6952–62, Jul. 2021.

[42] H.-J. Zepernick and A. Finger, *Pseudo random signal processing: theory and application*. John Wiley & Sons, Jul. 2013.
[43] Q. Yu and K. Song, "Uniquely decodable multi-amplitude sequence for massive grant-free multiple-access adder channels," *arXiv preprint arXiv:2110.11827*, Oct. 2021.
[44] D. Malak, "The interplay of spectral efficiency, user density, and energy in grant-based access protocols," *arXiv preprint arXiv:2207.11756*, Jul. 2022. [Online]. Available: https://arxiv.org/pdf/2207.11756.pdf
[45] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge University Press, 2005.
[46] R. W. Yeung, *Information theory and network coding*. Springer Science & Business Media, 2008.
[47] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–59, Apr. 2010.
[48] G. Khachatrian and S. Martirossian, "Code construction for the t-user noiseless adder channel," *IEEE Transactions on Information Theory*, vol. 44, no. 5, pp. 1953–1957, Sep. 1998.
[49] C. Geng *et al.*, "On the optimality of treating interference as noise," *IEEE Trans. Inf. Theory*, vol. 61, pp. 1753–67, Feb. 2015.
[50] E. Uhlemann, L. K. Rasmussen, and F. Brännström, "Puncturing strategies for incremental redundancy schemes using rate compatible systematic serially concatenated codes," in *Proc., Int. Symp. on Turbo Codes and Related Topics*, Apr. 2006.
[51] A. S. Barbulescu and S. Pietrobon, "Rate compatible turbo codes," *Electronics letters*, vol. 31, no. 7, pp. 535–536, Mar. 1995.
[52] P. Jung, J. Plechinger, M. Doetsch, and F. Berens, "A pragmatic approach to rate compatible punctured turbo-codes for mobile radio applications," in *Proc., Int. Conf. on Advances in Commun. and Control, Corfu, Greece*, Jun. 1997.
[53] J. Li and H. Imai, "Performance of hybrid-ARQ protocols with rate compatible turbo codes," in *Proc., Int. Symp. on Turbo Codes, Brest, France*, Sep. 1997, p. 188.
[54] D. N. Rowitch, "Rate compatible punctured turbo (RCPT) codes in a hybrid FEC/ARQ system," in *IEEE Global Telecommun. Mini-Conf., 1997*, 1997.