Throughput and Energy Tradeoffs for Retransmission-based Random Access Protocols

Derya Malak
Communication Systems Department, EURECOM
derya.malak@eurecom.fr

***Abstract*—The fifth-generation of wireless communication networks is required to support a range of use cases such as enhanced mobile broadband (eMBB), ultra-reliable, low-latency communications (URLLC), massive machine-type communications (mMTCs), with heterogeneous data rate, delay, and power requirements. The 4G LTE air interface is designed to support fewer devices with large payloads, and uses extra overhead to enable scheduled access, which is not justified for small payload sizes. In this paper, we employ a random access communication model with retransmissions for multiple users with small payloads at the low spectral efficiency regime. The radio resources are split non-orthogonally in the time and frequency dimensions. Retransmissions are combined via different Hybrid Automatic Repeat reQuest (HARQ) methods, namely Chase Combining and Incremental Redundancy with a finite buffer size constraint C_{buf}, via a conventional matched filter receiver. We determine the best scaling of the spectral efficiency (SE) versus signal-to-noise ratio (SNR) ρ per bit and the scaling for the user density (number of users per real degrees of freedom, rdof) versus SNR per bit, for the sum-optimal regime and when the interference is treated as noise, using a Shannon capacity approximation. Numerical results show that the scaling results are applicable over a range of η, T, C_{buf}, J, at low received SNR values. The proposed analytical framework can provide insights for resource allocation strategies in general random access systems and in specific 5G use cases for massive URLLC uplink access.**

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

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. The overhead of scheduled access in 4G LTE is not desirable in URLLC applications.

We consider a wireless multiple access communication 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 frequency bins which are shared by the users in a non-orthogonal manner via non-orthogonal multiple access (NOMA), and the time of duration τ second (sec) uniformly divided into T time slots, i.e., transmit opportunities. For a given blocklength n, the total dimension in a given channel use is $m = n/T$. We consider different forms of Hybrid Automatic Repeat reQuest (HARQ): (i) HARQ with Chase Combining of NOMA transmissions, CC-NOMA, (ii) HARQ with chase combining of OMA transmissions, CC-OMA, and (iii) HARQ with Incremental Redundancy, 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.

A. Related Work

Random access models and scaling of throughput. Random-access protocols have been pioneered with the emergence of ALOHA [1], which later yielded the development of carrier sense multiple access (CSMA). 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 [2]. In general, the resource being shared is on a time-frequency grid, and each transmission costs one time-frequency slot [3]. Other models include sparse coding for grant-free multiple access [4], multi-user detectors (MUDs) that improve the performance of random-CDMA [5] for spread spectrum systems (in different bases), e.g., orthogonal multiple access (OMA) coded OFDM, or CDMA with random non-orthogonal spreading, and generally NOMA [6]. Other works have focused on the capacity of Gaussian multiple access channels (MAC) [7], many-access channels with user identification [8], and quasi-static fading MAC [9].

Interference management and resource sharing. Interference management techniques have been studied under different spectral efficiency models. To accommodate massive random access, interference cancellation [10], collision resolution [2], load control [11], and interference cancellation given a target outage rate [12] have been proposed.

In [13], we characterized the scaling of throughput (user density) with deadline under outage constraints for a sub-optimal but more practical random access system where the time and frequency domains are slotted. The receiver uses conventional SUD, which decodes a desired user's data by treating other users' interfering signals as noise, subject to an SINR-based outage constraint. However, the main limitation is the fixed per-user power, which causes a linear scaling between the received SNR and the number of users.

HARQ models and coding sequences. HARQ is a combination of Automatic Repeat reQuest (ARQ) and forward error

correction (FEC) [14]. In particular, there are three models known as HARQ with Selective Repeat, HARQ with CC, and HARQ with IR [15]. This one is a salient variant of HARQ that captures puncturing via parity bits [16] and effects of HARQ buffer sizes [17].

Coding sequences have been devised for massive access, including Walsh sequences [18], or almost affinely disjoint subspaces [19], 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 [20]. 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 [21]. Sequence design for grant-free MAC has been studied in [22], where the authors devised uniquely-decodable multi-amplitude sequences. However, this approach induces a high E_b/N_0, which is not desirable in a practical massive access scenario with small payloads.

B. Overview, Contributions and Organization

The goal of this paper is to analyze a retransmission-based general random 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 describe the system model for random access and detail the key performance metrics, SE (bits/rdof), the SNR per bit (E_b/N_0), and user density (users/rdof). We delineate the retransmission-based random access schemes in Section III and analyze their SE and the SNR per bit for the sum-optimal and TIN cases. More specifically, we consider (i) the classical transmission scheme with no retransmissions, and the retransmission-based schemes using different combining techniques, namely (ii) CC-NOMA, (iii) CC-OMA, and (iv) IR-OMA. In Section IV, we numerically evaluate the scaling results, namely the SE versus SNR per bit and user density versus SNR per bit tradeoffs, and show the behavior with an increasing number of transmissions T, received SNR ρ, the HARQ buffer size C_{buf}, the non-orthogonality factor η, and the total number of users J. The key design insights for the proposed random access framework are as follows:

- **The low ρ regime is relevant.** Our framework exploits the conventional MFR for SUD and is 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, which degrades the performance of TIN-based models.
- **The SE of the sum-optimal strategy improves with NOMA.** Compared to OMA-based transmissions, CC-NOMA has a better SE vs E_b/N_0 performance. The performance of IR-OMA approaches the performance of the classical model as C_{buf} at the receiver increases.
- **The SE of the TIN strategy is optimal at low ρ.** If 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 smaller η can significantly improve the SE of TIN.
- **User density is sensitive to retransmissions.** For the sum-optimal model, although the performances of CC-OMA, IR-OMA, and the classical techniques degrade with increasing T, CC-NOMA does not sacrifice the number of users per rdof as much. A higher number of users J, when the per user power is kept fixed, results in a lowered received SNR ρ per user, which improves the SE for the sum-optimal CC-NOMA model, yet for the TIN-based model, the SINR drops due to the higher effective interference, which degrades the SE for a given SNR per bit. For TIN, CC-OMA and CC-NOMA perform well at low ρ values and CC-OMA can effectively combine retransmissions even at higher values of ρ. However, the SNR per bit requirement of CC-NOMA is very sensitive to ρ, which deteriorates the performance at high ρ values.
- **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 other models that rely on OMA) in the infinite blocklength (IBL) regime, which gives an upper bound to the actual user density scaling. As n increases, these upper bounds become tighter.

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, which could strictly improve performance of random-CDMA.

II. System Model

We consider a random access communication model where a set 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 to communicate its payload. The users use non-orthogonal signatures (kept identical at each retransmission attempt) to transmit their payloads, as shown in Fig. 1-(a).

Frame structure. A frame has a total bandwidth of ω Hertz (Hz) and the time of duration τ second (sec), which 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 frequency bin and time slot as a time-frequency slot (TFS). For the proposed frame structure, the number of resources or rdof in a frame is $N = \omega\tau$. The total number of rdof N is evenly split into T retransmissions. The time-frequency resources 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 symbol.

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 be the number of users at slot $t \in \{1, \dots, T\}$, $\mathcal{J}_t$ be the set of users at slot t such that $J = \sum_{t=1}^{T} J_t$, and $\mathcal{J}$ be the set of all users in the frame.

Let $\mathbf{U}_j = (U_{j1}, U_{j2}, \dots, U_{jK})$ be the K dimensional source vector corresponding to user $j \in \mathcal{J}$. In the case of no feedback, we model each retransmission from user j with a blocklength m by $\mathbf{X}_{tj} = (X_{tj1}, X_{tj2}, \dots, X_{tjm})$, where we assume that the transmitted signal from user j is given by $\mathbf{X}_{tj} = \phi_{tji}(\mathbf{U}_j) = a_{tj}\mathbf{S}_j$, where $a_{tj} \in \mathbb{C}$ denotes the complex amplitude of the transmitted symbol of user j at slot t, and $\mathbf{S}_j$ denotes its signature (spreading sequence). Let $\phi_{tji} : \mathcal{U}_j^K \to \mathcal{X}_j$ for $i = 1, \dots, m$ be the encoder function of $j \in \mathcal{J}$ for retransmission attempt $t \in \{1, \dots, T\}$.

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, as indicated in Fig. 1-(a). We assume that $\mathbf{S}_j$ are unitary, $\|\mathbf{S}_j\| = 1$, and $|\langle \mathbf{S}_j, \mathbf{S}_{j'} \rangle| = \eta$ 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 [20] 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 illustrate the frame structure with overlapping NOMA traffic in Fig. 1-(b).

Received signal and matched filter decoding. We denote the received signal vector during transmission $t \in \{1, \dots, T\}$ by $\mathbf{Y}_t = (Y_{t1}, Y_{t2}, \dots, Y_{tm})$. We also let $\mathbf{Y}_t^i = (Y_{t1}, Y_{t2}, \dots, Y_{ti})$. The channel is additive such that the received signal vector during transmission t is

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

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 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 is desired as it allows many users per dof, where the other users' signals are treated as additive white Gaussian noise (AWGN) [5]. On the other hand, when 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 [5]. In the case that $\{\mathbf{S}_j\}_{j \in \mathcal{J}}$ 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 [5].

Maximum ratio combining. The receiver's HARQ buffer size equals the number of coded symbols per coded packet, where 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 vector $\mathbf{Y}_t$ during transmission $t \in \{1, \dots, T\}$ from the transmitted signals $\{\mathbf{X}_{tj}\}_{j \in \mathcal{J}_t}$. 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.

Per user received SNR. The noise power each user sees is assumed to be additive and constant with value σ_t^2, $t \in \{1, \dots, T\}$ per dimension, i.e., $\mathbb{E}[\langle \mathbf{Z}_t, \mathbf{Z}_t \rangle] = m\sigma_t^2$, where $m\sigma_t^2$ is the total noise power across the number of frequency bins, which is B. The energy constraint for each source $j \in \mathcal{J}$ at any given $t \in \{1, \dots, T\}$ is

$$\frac{1}{T}\mathbb{E}[\mathbf{U}_j^\intercal \mathbf{U}_j] = \mathbb{E}[\mathbf{X}_{tj}^\intercal \mathbf{X}_{tj}] = m\sigma_t^2\rho_{tj} \leq \frac{KE_j}{T} \ , \tag{3}$$

i.e., the total power of channel input linearly scales with the message size K, where the received power of user $j \in \mathcal{J}$ normalized with respect to K is E_j. We assume that $\mathbb{E}[X_{tji}] = 0$ and X_{tji}'s across $j \in \mathcal{J}$ are not independent such that $\mathbb{E}[\mathbf{X}_{tj}^\intercal \mathbf{X}_{tj'}] \leq \frac{KE_{jj'}}{T}$ for $\{(j, j') : \ j \neq j'\}$.

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

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

We further assume that ρ_{tj} are identical and denoted by ρ. Given a constant received power of $\sigma^2\rho$, the received SNR equals ρ. The relation between ρ and σ^2 is given as $\rho = \frac{1}{\sigma^2}\mathbb{E}[\mathbf{X}_{tj}^\intercal \mathbf{X}_{tj}]$. The total power spent by all users is

$$P_{tot} = J\sigma^2\rho \ . \tag{4}$$

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

a) **Spectral efficiency:** *The spectral efficiency*, SE, is the maximum number of bits per channel use (bits/s/Hz):

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

where rdof represents the total number of *real dof*, n.

Definition 1. *(Achievable rate [23].) A rate R is achievable with complete feedback for a discrete memoryless channel (DMC) $p(y \,|\, x)$ if for any $\epsilon > 0$, there exists for sufficiently large n an (n, M) code such that $\frac{1}{n}\log M > R - \epsilon$.*

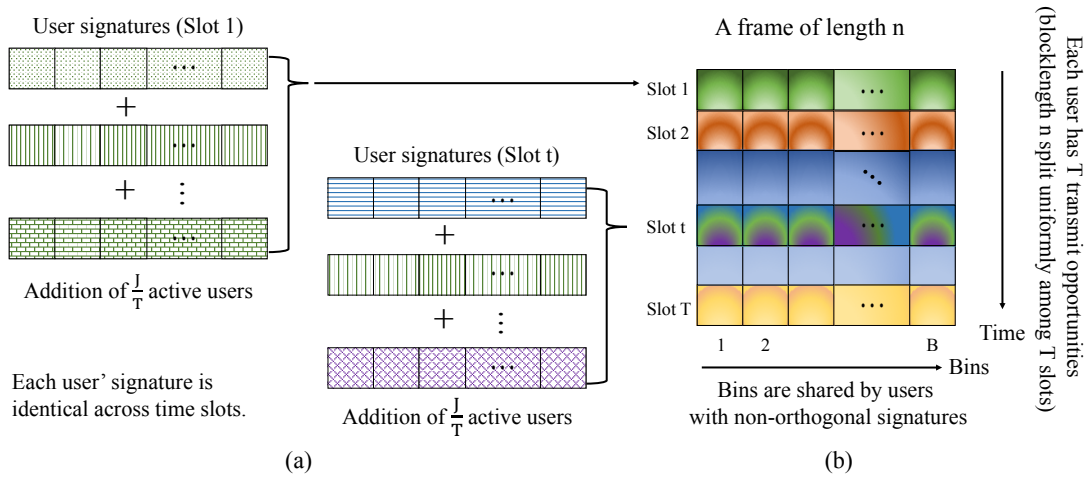

Fig. 1: (a) Non-orthogonal user signatures at time slots 1 and t. Each user uses the same signature across all time-frequency resources. The second user in slot 1 is repeated in slot t (same signature). (b) The frame structure where time is partitioned into T transmit opportunities, and the time-frequency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

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:

$$\frac{L}{n} \leq C = \frac{1}{2}\log_2(1+\text{SINR}) \quad bit/rdof, \quad n \leq N \ , \tag{6}$$

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

For finite n, the rate achievable is approximated by [24]

$$R(n,\epsilon) = \frac{1}{n}\log M(n,\epsilon) \approx C - \sqrt{\frac{V}{2n}}Q^{-1}(\epsilon) \ , \tag{7}$$

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 and Q^{-1} is its inverse. Furthermore, in (7), V is the channel dispersion given by $V = 1 - \frac{1}{(1+\text{SINR})^2}$, and for the AWGN channel model by TIN, C is Shannon's channel capacity, which satisfies (6).

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

$$\frac{E_b}{N_0} = \frac{\text{Total energy spent}}{2 \times \text{Total number of bits}} \ . \tag{8}$$

c) **User density:** *User density* (users/rdof), J/n, represents the total number of users per rdof that can transmit within the same frame (of which J_t/n denotes the density of users that can simultaneously transmit in slot t), for a given total blocklength n for the given frame duration. From (5), (6), and (8) the achievable user density for a given n is affected by the SE vs SNR per bit tradeoffs of the retransmission schemes, which we detail in Section III.

III. Combining NOMA-based Retransmissions

We focus on the scaling behaviors of the SE, the SNR per bit, and the user density for the retransmission-based random 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 schemes for the sum-optimal regime and via TIN.

A. No Multiplexing of Retransmissions

We commence with the classical interference-based model with no multiplexing across different time slots. The time resources are split uniformly across T slots. There are $J_t = J/T$ users per slot sharing the frame resources. The SE of the classical sum-optimal transmission approach is given as

$$R_{\mathsf{sum}}^{\mathsf{Clas.}} = \frac{1}{2}\log_2\left(1+\rho\frac{J}{T}\right) \quad bit/rdof \ . \tag{9}$$

The SNR per bit of this model is equal to

$$\frac{E_b}{N_0} = \frac{J\sigma^2\rho n}{m\log_2\left(1+\rho\frac{J}{T}\right)} = \frac{J\sigma^2\rho T}{\log_2\left(1+\rho\frac{J}{T}\right)} \ . \tag{10}$$

The SE of the classical model via TIN is expressed as

$$R_{\mathsf{TIN}}^{\mathsf{Clas.}} = \frac{J}{2T}\log_2\left(1+\frac{\rho}{\rho(\frac{J}{T}-1)+1}\right) \quad bit/rdof. \tag{11}$$

Similarly, the SNR per bit for this model is given as

$$\frac{E_b}{N_0} = T\sigma^2\rho\Big/\log_2\left(1+\frac{\rho}{\rho(\frac{J}{T}-1)+1}\right) \ . \tag{12}$$

In general, transmissions are exposed to different channel conditions, more specifically the fading (e.g., exponentially distributed interference power, i.e., Rayleigh fading) or path

loss. The channel gains $|H_{jt}|^2$, as function of $t = 1, \ldots, T$ and $j \in \mathcal{J}$ can be incorporated into the system model, assuming that $|H_{jt}|^2$ has unit power and is independent across the slots with a known cumulative distribution function, $F_{|H|^2}$. Incorporating the channel gains, we can express the SE of the classical sum-optimal model as

$$R_{\mathsf{sum}}^{\mathsf{Clas.}} = \frac{1}{2T} \sum_{t=1}^{T} \log_2 \left(1 + \rho \sum_{j \in \mathcal{J}_t} |H_{jt}|^2\right) \quad bit/rdof \ .$$

The SNR per bit for the classical sum-optimal model is

$$\frac{E_b}{N_0} = J\sigma^2 \rho n \Big/ \frac{m}{T} \sum_{t=1}^{T} \log_2 \left(1 + \rho \sum_{j \in \mathcal{J}_t} |H_{jt}|^2\right) \ .$$

A more comprehensive SINR model that incorporates the channel gains will be considered as a future direction.

In the case of no retransmissions, TIN is essentially optimal for low SE [20]. However, for strategies that combine the retransmissions, the classical TIN may not be optimal even at low SE. We will later demonstrate in Sections III-B, III-C, and III-D, and via simulations (Section IV) that sum-optimal capacity models could be more energy efficient via combining of retransmissions versus TIN.

B. Chase Combining with NOMA-based Retransmissions

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

Proposition 1. *The SE of CC-NOMA for the sum rate optimal model is expressed by the following upper bound:*

$$R_{\mathsf{sum}}^{\mathsf{CC,NOMA}} = \frac{1}{2} \log_2 \left(1 + \rho T \left[1 + \eta^2 \Big(\frac{J}{T} - 1\Big)^2\right]\right) . \quad (13)$$

Proof. See Appendix. □

Using (8), the SNR per bit for CC-NOMA for the sum rate optimal model is given as

$$\frac{E_b}{N_0} = J\sigma^2 \rho \Big/ \log_2 \left(1 + \rho T \left[1 + \eta^2 \Big(\frac{J}{T} - 1\Big)^2\right]\right) . \quad (14)$$

The SE for CC-NOMA with TIN is

$$R_{\mathsf{TIN}}^{\mathsf{CC,NOMA}} = \frac{J}{2T} \log_2 \left(1 + \frac{\rho T^2}{T + \rho\eta^2 (J-T)^2}\right) . \quad (15)$$

The SNR per bit of $R_{\mathsf{TIN}}^{\mathsf{CC,NOMA}}$ is given as:

$$\begin{aligned} \frac{E_b}{N_0} &= J\sigma^2 \rho n \Big/ n \frac{J}{T} \log_2 \left(1 + \frac{\rho T^2}{T + \rho\eta^2 (J-T)^2}\right) \quad (16) \\ &= J\sigma^2 \cdot \frac{\frac{1}{T}(2^{\frac{2T}{J} SE} - 1)}{1 - \eta^2 \big(\frac{J}{T} - 1\big)^2 (2^{\frac{2T}{J}\mathsf{SE}} - 1)} \cdot \frac{1}{2\mathsf{SE}} \\ &\geq -1.59dB + 10 \log_{10} \sigma^2 \ , \end{aligned}$$

where we used $\frac{2^{2\mathsf{SE}}-1}{2\mathsf{SE}} \geq -1.59dB$ as $\mathsf{SE} \to 0$.

C. Chase Combining with OMA-based Retransmissions

Here, 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 [13]. We next provide its SE.

Proposition 2. *The SE for CC-OMA with the sum-optimal capacity model is given as*

$$R_{\mathsf{sum}}^{\mathsf{CC,OMA}} = \frac{1}{2} \log_2 \left(1 + \rho T \Big[1 + \frac{1}{T}\Big(\frac{J}{T} - 1\Big)\Big]\right) . \quad (17)$$

Proof. The final result follows from Prop. 1 in [13]. □

The SNR per bit of $R_{\mathsf{sum}}^{\mathsf{CC,OMA}}$ is given as

$$\frac{E_b}{N_0} = J\sigma^2 \rho \Big/ \log_2 \left(1 + \rho T \Big[1 + \frac{1}{T}\Big(\frac{J}{T} - 1\Big)\Big]\right) . \quad (18)$$

The SE of CC-OMA with TIN is given as

$$R_{\mathsf{TIN}}^{\mathsf{CC,OMA}} = \frac{J}{2T} \log_2 \left(1 + \frac{\rho T}{1 + \rho\big(\frac{J}{T} - 1\big)}\right) . \quad (19)$$

In the limit as $J \to \infty$, it holds that $R_{\mathsf{TIN}}^{\mathsf{CC,OMA}} \leq \frac{T}{2 \log 2}$. The subsequent result follows using (8) and (19).

The SNR per bit of $R_{\mathsf{TIN}}^{\mathsf{CC,OMA}}$ is given as

$$\begin{aligned} \frac{E_b}{N_0} &= J\sigma^2 \rho n \Big/ n \frac{J}{T} \log_2 \left(1 + \frac{\rho T}{1 + \rho\big(\frac{J}{T} - 1\big)}\right) \\ &= J\sigma^2 \cdot \frac{\frac{1}{T}(2^{\mathsf{SE}\frac{2T}{J}} - 1)}{1 - \frac{1}{T}\big(\frac{J}{T} - 1\big)(2^{\mathsf{SE}\frac{2T}{J}} - 1)} \cdot \frac{1}{2\mathsf{SE}} \\ &\geq -1.59dB + 10 \log_{10} \sigma^2, \quad (20) \end{aligned}$$

where the second step follows from using (19), and the last step follows from the same intuition as in (16).

D. Incremental Redundancy with OMA Retransmissions

In this section, we 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. Hence, a finite HARQ buffer has an impact on the throughput of HARQ. Unlike CC-NOMA, in IR-OMA every retransmission contains different information than the previous one. Furthermore, different from CC, where the buffer size is the same as the number of packets per transmission, in IR-OMA, which is 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.

In IR-HARQ, multiple different sets of code bits are generated for the same information bits used in a packet. These sets consist of different redundant flavors obtained by different puncturing configurations. The retransmitted packets provide successive refinement [25] by iteratively improving the rate-distortion as more information is sent.

Expected quantization distortion. Using the refinement-based approach in [25], 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 the relation $\mathbf{Q}_{t,T} \sim \mathcal{CN}\big(0, \frac{2\sigma_q^2(t,T)}{m} I_m\big)$, 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 $1 \leq t < T$ and is successful at attempt T. From (1), $\mathbf{Y}_t$ has a dimension $m = n/T$.

Proposition 3. *The SE for IR-OMA with the sum-optimal capacity model is given as*

$$R_{\mathsf{sum}}^{\mathsf{IR,OMA}} = \sum_{t=1}^{T} \frac{B}{2} \log_2 \Big(1 + \frac{\rho J \sigma^2/B}{\sigma^2 + \sigma_q^2(t,T-1)/B}\Big) \ , \quad (21)$$

which has the units of $bit/rdof/(T$ slots$)$.

Proof. Given the buffer size normalized with respect to the packet lengths, C_{buf}, the transmit rate is given by $\frac{mC_{\mathsf{buf}}}{n} = \frac{C_{\mathsf{buf}}}{T}$. Using (1) at transmission attempt $t < T$, it holds that

$$\frac{C_{\mathsf{buf}}}{T} = I(\mathbf{Y}_t; \hat{\mathbf{Y}}_t) = \frac{B}{2} \log_2 \Big(1 + \frac{J\rho\sigma^2/B + \sigma^2}{\sigma_q^2(t,T)/B}\Big) \ ,$$

which leads to a quantization noise as function of C_{buf}:

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

where in (21) $\sigma_q^2(T,T) = 0$, i.e., at retransmission T, $\mathbf{Y}_T = \hat{\mathbf{Y}}_T$, i.e., the receiver recovers $\mathbf{Y}_T$. We also have the notational convention $\sigma_q^2(T,T-1) = 0$. Combining the retransmissions, each providing an addendum to the first transmission so that the signal as a result of T transmissions achieves the desired distortion, we obtain (21). □

The SNR per bit of IR-OMA for sum-optimal case is

$$\frac{E_b}{N_0} = J\sigma^2\rho \Big/ \frac{B}{T} \sum_{t=1}^{T} \log_2 \big(1 + \frac{\rho J/B}{1 + \sigma_q^2(t,T-1)/(B\sigma^2)}\big) \ .$$

As $C_{\mathsf{buf}} \to \infty$, $\frac{E_b}{N_0} \to \frac{J\sigma^2\rho}{\log_2(1+\rho J)}$. For smaller C_{buf}, $\frac{E_b}{N_0} > \frac{J\sigma^2\rho}{\log_2(1+\rho J)}$. We note that as C_{buf} increases the IR-OMA sum SE matches the sum SE for the classical problem without combining transmissions (sum-optimal case). 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 4. *The SE for IR-OMA with TIN is given as*

$$R_{\mathsf{TIN}}^{\mathsf{IR,OMA}} = \sum_{t=1}^{T} \frac{JB}{2T} \log_2 \Big(1 + \frac{\rho\sigma^2/B}{\rho\frac{\sigma^2}{B}\big(\frac{J}{T} - 1\big) + \sigma^2 + \frac{\sigma_q^2(t,T-1)}{B}}\Big) \ , \quad (22)$$

which has the units of $bit/rdof/(T$ slots$)$, where the buffer size normalized with respect to the packet lengths, i.e., C_{buf}, more precisely the transmit rate satisfies:

$$\frac{C_{\mathsf{buf}}}{T} = \frac{B}{2} \log_2 \Big(1 + \frac{\rho\sigma^2/B + \sigma^2}{((J/T-1)\rho\sigma^2 + \sigma_q^2(t,T))/B}\Big) \ ,$$

which implies that as $\sigma_q^2(t,T) \to 0$ for $C_{\mathsf{buf}} \geq \frac{BT}{2} \log_2 \big(1 + \frac{\rho\sigma^2/B+\sigma^2}{(J/T-1)\rho\sigma^2/B}\big)$, where the size of C_{buf} could be tuned to the channel capacity (7) in the FBL regime. The relation between C_{buf} and $\sigma_q^2(t,T)$ leads to the following relation between $\sigma_q(t,T)$ and the buffer size:

$$\sigma_q^2(t,T) = \frac{B(\rho/B+1)\sigma^2}{2^{\frac{2C_{\mathsf{buf}}}{TB}} - 1} - (J/T-1)\rho\sigma^2 \ . \quad (23)$$

The proof of next proposition can be obtained from (22) and (23), which we skip due to space limits.

Proposition 5. *The SNR per bit of IR-OMA for TIN equals*

$$\frac{E_b}{N_0} = T\sigma^2\rho \Big/ \frac{B}{T} \sum_{t=1}^{T} \log_2 \big(1 + \frac{\rho/B}{\rho\zeta_t/B + 1}\big) \ , \quad (24)$$

where $\zeta_t = \big(J/T - J/(T-1) + 1/(2^{\frac{2C_{\mathsf{buf}}}{(T-1)B}} - 1)\big) + B/\rho \cdot 1/(2^{\frac{2C_{\mathsf{buf}}}{(T-1)B}} - 1)$ for $t < T$, and $\zeta_T = (J/T - 1)$.

In general as $\rho \to 0$ for large buffer sizes C_{buf}, it is immediate from Prop. 5 that $\frac{E_b}{N_0} \to \log 2 \cdot J\sigma^2\rho$. For smaller C_{buf}, the ratio $\frac{E_b}{N_0}$ is typically larger vs classical TIN.

IV. Numerical Evaluation of Scaling Results

Contrasting SE vs SNR per bit. We first study the SE (bits/rdof) versus the E_b/N_0 (dB) tradeoff for the different HARQ-based retransmission combining models in Section III. The set of chosen parameters for the sum-optimal and the TIN schemes is indicated on the plots. Our numerical results in Fig. 2 are for the IBL regime, providing upper bounds to the actual scaling behaviors for the SE models.

Number of retransmissions T. Increasing T degrades the SE of the sum-optimal model (with CC-NOMA (13) and CC-OMA (17)). The SE for the TIN (with CC-NOMA and CC-OMA) improves. We note for IR-OMA that as T increases, unlike the CC-NOMA and CC-OMA models, each retransmission contains less information for successive refinement of the signal. In terms of the SE vs E_b/N_0 behavior, for $T > 1$, TIN CC-OMA is better than TIN IR-OMA and TIN classical. This trend is also obvious from relations (11), (15), (19), and (22). The gap between TIN CC-OMA and TIN IR-OMA grows with increasing T.

Finite buffer size C_{buf} at the decoder. The SNR per bit for the classical TIN and the IR-OMA models 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}, TIN IR-OMA can perform better than TIN CC-NOMA when interference is high. When C_{buf} is small, i.e., under high quantization noise, the performance of TIN IR-OMA could be worse than TIN CC-NOMA and classical TIN for $T > 1$.

Scaling of user density J/n versus E_b/N_0. We investigate this scaling behavior in Figs. 3 and 4 as a function of ρ. A high ρ value yields a higher E_b/N_0 to achieve the same user density. As T increases, $m = n/T$ decreases, and the supported user density drops. As the received SNR ρ increases, the user scalings of different models for the sum-optimal strategy become similar under high C_{buf}. This 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-optimal models is $\frac{\rho}{\log_2 \rho}$ for high ρ, which causes a significant drop in J/n. However,

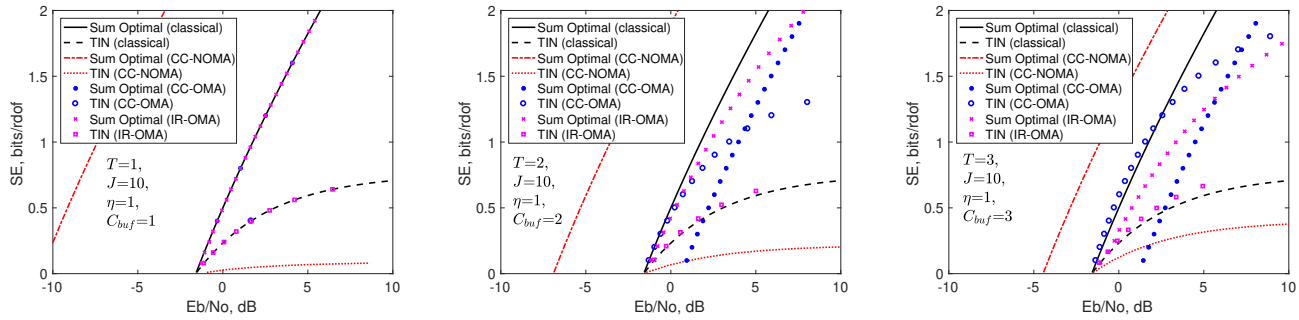

Fig. 2: Scaling of SE vs E_b/N_0 for varying T for $\eta = 1$ and $J = 10$, and moderate buffer size, $C_{\mathsf{buf}} = T$.

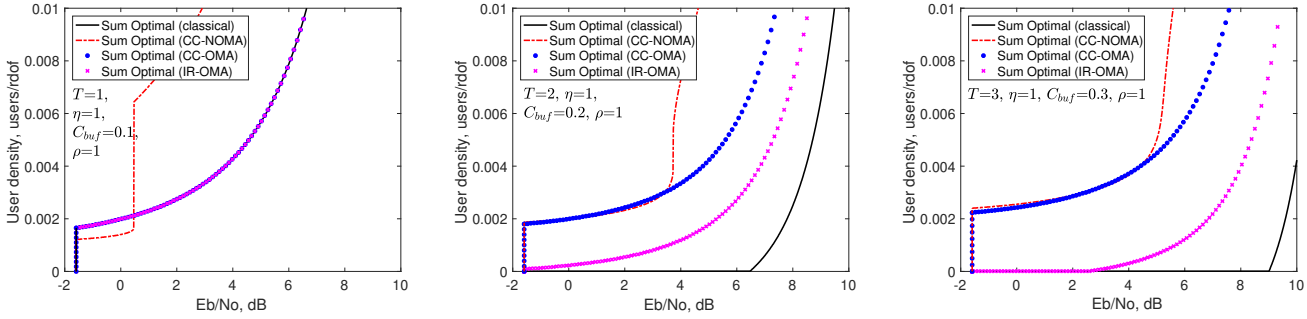

Fig. 3: (Sum-optimal) Scaling of J/n vs. E_b/N_0 for varying ρ and $C_{\mathsf{buf}} = 0.1T$, for $T = 1$, $T = 2$, and $T = 3$.

with the conventional MFR, the optimal SE performance of the sum-rate optimal model cannot be accurately captured when ρ is high [5].

While retransmissions are inevitable in HARQ-based protocols, retransmission-based access schemes in general degrade the SE vs E_b/N_0 performance and similarly for J/n vs E_b/N_0. For the sum-optimal model, in terms of the sensitivity of the J/n versus E_b/N_0 tradeoff with respect to increasing T, we have the ordering as the CC-NOMA model, the IR-OMA model, the CC-OMA model, and the classical model from the less sensitive (best) to the most sensitive (worst). For the TIN-based model, the CC-NOMA and the CC-OMA models improve the user density by increasing T, and the IR-OMA and the classical models are not robust to retransmissions. While the scaling performance improves with T, increasing T causes diminishing returns in gains. According to [7], the scalings for the FBL are linear instead of being exponential (subject to BER constraints). However, unlike the setting in [7], where the total power P_{tot} is kept constant, we have a per user power constraint such that $P_{tot} = J\sigma^2\rho$, which scales linearly with J under fixed ρ.

From Fig. 4, we observe that the different models we considered in this paper perform the best at low spectral efficiency. In other words, increasing ρ decreases the user density scaling performance for IR-OMA, CC-NOMA, and CC-OMA. To compensate for the loss of CC-NOMA, even though better coding signatures (lower η) can be incorporated, this model still requires a higher minimum SNR per bit versus the other models with a higher sensitivity to ρ. As T increases it is possible to achieve a higher number of users per rdof, and similarly, via increasing C_{buf} we can achieve a better scaling for IR-OMA. At high C_{buf} (or high $\rho = 10$), 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 (20) than IR-OMA with an SNR per bit given in (24) (versus vice versa for lowered C_{buf} (or smaller $\rho \leq 1$)).

V. Conclusions

We considered HARQ-based random access models for 5G wireless communication networks, where the receiver jointly decodes retransmissions via different combining techniques, namely CC-NOMA, CC-OMA, and IR-OMA. We characterized the SE vs SNR per bit, and the user density vs SNR per bit tradeoffs, and demonstrated via numerical simulations that retransmissions can improve the scaling behaviors of SE and the user density. We further showed that the SE of the sum-optimal strategy improves with NOMA, and the SE via TIN is optimal at low SNR. In CC-NOMA, the user density is not affected much by retransmissions, and for IR-OMA the sensitivity decreases with C_{buf}.

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. It is crucial to support heterogeneous traffic type requirements on one platform where distinct classes of users are under different SINR requirements. The generalization of the classical capacity models to the FBL regime is of primary interest via incorporating channel gain, fading, or path loss,

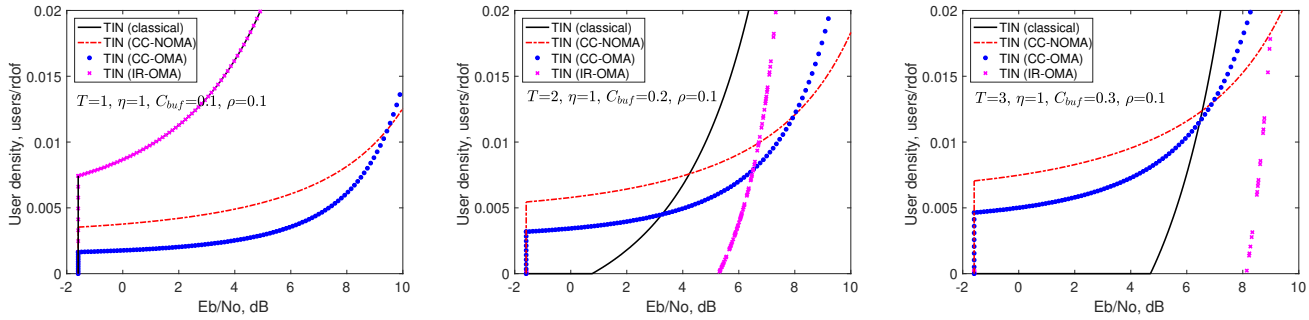

Fig. 4: (TIN) Scaling of J/n vs. E_b/N_0 for $\rho = .1$ and varying C_{buf}, $C_{\mathsf{buf}} = 0.1T$.

as well as techniques to achieve optimal performance for the SU and the MU settings.

APPENDIX

Total received signal power (for decoding j) is given as $P_S = \left(\sum_{t=1}^{T} |a_{tj}|^2\right)^2$. Computing the expected value of the total noise (by TIN), we obtain

$$P_N = \eta^2 \Big| \sum_{t=1}^{T} \sum_{j' \in \mathcal{S}_{t,-j}} a_{tj} a_{tj'}^* \Big|^2 + \sum_{t=1}^{T} |a_{tj}|^2 m\sigma_t^2 \ .$$

Using P_S and P_N, the total received power is $(Tm\sigma^2\rho)^2 + \eta^2(J-T)^2(m\sigma^2\rho)^2$. Rescaling this by the noise power, the SINR equals $T\rho\big[1+\eta^2\Big(\frac{J}{T}-1\Big)^2\big]$. Incorporating (3), the total capacity $R_{\mathsf{sum}}^{\mathsf{CC,NOMA}}$ (in bit/rdof) of the $J = \sum_{t=1}^{T} J_t$ user Gaussian MAC is given as in (13).

The SU decoder sees an effective SINR given as

$$\text{SINR} = \frac{P_S}{P_N} = \frac{(T\rho)^2}{\eta^2\big(\sum_{t=1}^{T}(J_t-1)\rho\big)^2 + T\rho} \ , \qquad (25)$$

which follows from dividing both the numerator and the denominator terms by $(m\sigma^2)^2$, and letting $\rho_{tj} = \rho$, and defining $\sum_{j' \in \mathcal{S}_{t,-j}} 1_{a_{tj'} \neq 0} = J_t - 1$.

REFERENCES

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

[2] 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.

[3] W. Yu, "On the fundamental limits of massive connectivity," in *Proc., Inf. Theory and Apps. Wkshp (ITA)*, Feb. 2017, pp. 1–6.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] S. S. Kowshik and Y. Polyanskiy, "Quasi-static fading MAC with many users and finite payload," in *Proc., IEEE ISIT*, Jul. 2019.

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

[11] 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.

[12] 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.

[13] 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.

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

[15] 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.

[16] 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.

[17] M. Danieli, S. Forchhammer, J. D. Andersen, L. P. Christensen, and S. S. Christensen, "Maximum mutual information vector quantization of log-likelihood ratios for memory efficient HARQ implementations," in *Proc., Data Compression Conf.*, Mar. 2010.

[18] 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.

[19] H. Liu, N. Polianskii, I. Vorobyev, and A. Wachter-Zeh, "Almost affinely disjoint subspaces," *Finite Fields and Their Applications*, vol. 75, p. 101879, Oct. 2021.

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

[21] H.-J. Zepernick and A. Finger, *Pseudo random signal processing: theory and application*. John Wiley & Sons, Jul. 2013.

[22] 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.

[23] R. W. Yeung, *Information Theory and Network Coding*. Springer Science & Business Media, 2008.

[24] 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.

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