Fundamental Limits of Non Orthogonal Multiple Access (NOMA) for the Massive Gaussian Broadcast Channel in Finite Block-Length

Jean-Marie S Gorce, Philippe Mary, Dadja Anade, Jean-Marc Kélif

MDPI

Article

Fundamental Limits of Non Orthogonal Multiple Access (NOMA) for the Massive Gaussian Broadcast Channel in Finite block-length

Jean-Marie Gorce [1], Philippe Mary [2], Dadja Anade [1], and Jean-Marc Kélif [3]

[1] Laboratoire CITI, a Joint Laboratory between INRIA, the Université de Lyon and the Institut National de Sciences Appliquées (INSA) de Lyon. 6 Av. des Arts, 69621 Villeurbanne, France; jean-marie.gorce@insa-lyon.fr (J.-M.G.); dadja.anade-akpo@inria.fr (D.A.)

[2] IETR and the Institut National de Sciences Appliquées (INSA) de Rennes, 20 Avenue des Buttes de Coësmes, CS 70839, 35708 Rennes, France; philippe.mary@insa-rennes.fr

[3] Orange Labs, Châtillon sur Seine, France; jeanmarc.kelif@orange.fr.

* Correspondence: jean-marie.gorce@insa-lyon.fr

Version January 20, 2021 submitted to Sensors

Abstract: Superposition coding (SC) has been known to be capacity achieving for the Gaussian memoryless broadcast channel for more than 30 years. However, SC regained interest in the context of non orthogonal multiple access (NOMA) in 5G. From an information theory point of view, SC is capacity achieving in the broadcast Gaussian channel, even when the number of users tends to infinity. But using SC has two drawbacks: decoders complexity increases drastically with the number of simultaneous receivers, and the latency is unbounded since SC is optimal only in the asymptotic regime. To evaluate these effects quantitatively in terms of fundamental limits, we introduce a finite time transmission constraint imposed at the base station and we evaluate fundamental trade-offs between the maximal number of superposed users, the coding block-length and the block error probability. The energy efficiency loss due to these constraints is evaluated analytically and by simulation. Orthogonal sharing appears to outperform SC for hard delay constraints (equivalent to short block-length) and in low spectral efficiency regime (below one bit per channel use). These results are obtained by the association of stochastic geometry and finite block-length information theory.

Keywords: Superposition Coding, Many-user Gaussian Broadcast Channel, Non Orthogonal Multiple Access, Massive access, Finite block-length, Information Theory

1. Introduction

The Internet of things (IoT), connecting objects instead of humans, is one of the major applications of 5G and future generations of communications systems. Moreover, the transition towards machine to machine communications induces an important shift from the theoretical modeling of these systems. Indeed, the IoT paradigm relies on bursty but massive distributed communications to comply with the transmission requests of billions of communicating objects spread over a large area, while transmitting only few packets per day, month or even per year. In such a scenario, the classical fundamental limits of communication systems derived using the tools introduced by Claude E. Shannon [1] need to be revised. From this perspective the capacity or the capacity region in case of multi-user communications, becomes less important in regard to other metrics [2,3].

The seminal Shannon's second theorem established the capacity in additive white Gaussian noise (AWGN) channel, which can also be expressed as the fundamental trade-off between energy efficiency (EE) η_E and spectral efficiency (SE) η_S [4]. Let us define $\eta_S := R/W$ where R and W are respectively

the rate in bits/s and the channel bandwidth in Hz, respectively. In an ideal system, the number of channel uses (symbols in a narrow-band transmission) is $1/W$ and η_S can be expressed in bits per channel use (bpcu). Now, letting the energy E be normalized with respect to the noise power density N_0, the energy efficiency is defined as $\eta_E := \frac{RT}{E/N_0}$, where T corresponds to the transmission duration. By using the definition of η_S, the following relation holds $\eta_E = \eta_S \frac{\sigma^2}{P}$, with the average power $P = E/T$ and $\sigma^2 = WN_0$. Therefore the energy efficiency can be alternatively thought as a power efficiency metric. In the following, by a slight abuse of notation, we will give to η_E the dimension of bits per relative power unit (bppu).

Using this notation, Shannon's channel coding theorem can be written as follows:

$$\eta_E \leq \frac{\eta_S}{\left(2^{2\eta_S} - 1\right)}. \tag{1}$$

This trade-off is achievable only in the asymptotic regime, i.e. when the encoding time spreads over an infinite number of channel uses (c.u.).

This asymptotic result relies on two assumptions which are no longer valid in the context of the IoT paradigm:

- The traffic is characterized by a continuous data flow.
- The encoding length is over an infinite number of channel uses, and hence without any latency constraint.

While modeling IoT packets consisting of a few information bits under ultra low latency constraint (ULLC), the asymptotic regime becomes irrelevant. The first attempts to derive fundamental limits in the non-asymptotic regime dates back to Feinstein and Shannon in 50's [5,6]. They provided an achievability bound on the rate considering maximal and average decoding error probability respectively. A refinement on these results including cost constraints on codewords has been provided by Gallager [7]. The problem of achievability and converse bounds on the rate in the finite block-length (FBL) regime has recently received a renewed interest with the work of Polyanskiy *et al* [8], who studied the fundamental limits of the point to point AWGN channel. This has paved the way to study latency and reliability constraints from a fundamental point of view. One of the major results in [8] is the asymptotic expansion of the achievable rate R

$$R = C\left(\gamma\right) - \sqrt{\frac{V\left(\gamma\right)}{n}} Q^{-1}\left(\epsilon\right) + O\left(\frac{\log n}{n}\right), \tag{2}$$

where $C\left(\gamma\right) = \frac{1}{2}\log_2\left(1+\gamma\right)$ is the channel capacity in bpcu per dimension under the signal to noise ratio γ, $V(\gamma) = \frac{\gamma(\gamma+2)}{2(\gamma+1)^2}\left(\log_2 e\right)^2$ is the channel dispersion defined as the variance of the information density between emitted and received codewords, n the number of c.u., ϵ being the error probability (average or maximal) and $Q(x) = 1/\sqrt{2\pi}\int_x^\infty \exp(-y^2/2)dy$. This work has been extended to the multi-antenna case and fading channels in [9–11].

The work initiated in [8] has an impact far beyond theory and is of great interest for practical specifications for IoT networks, since the expression in (2) links three fundamental constraints, identified as critical for IoT; i.e. reliability, latency or spectral efficiency, and energy efficiency [12].

1.1. Multi-user finite block-length: state of the art

Information theory has been proved to be a powerful tool to establish fundamental limits of point-to-point (P2P) or multi-user communication systems including the multiple access channel (MAC) and the broadcast channel (BC). These models fit well with the uplink and the downlink in a wireless cell, respectively. In the asymptotic regime, the exact characterization of achievable rate regions has been obtained [2].

According to [13], Gaussian MAC and BC capacity regions are dual of each other, under transferable power hypothesis. This hypothesis means that in the uplink, the sum-power is constrained

but not the individual powers. This MAC with transferable power represents an optimistic model, but guarantees that BC bounds constitute outer bounds for the MAC, while providing a more tractable expression.

FBL information theory has been initially extended to MAC and BC scenarios in [14]. Gaussian MAC has been particularly investigated in [15,16] among others and achievable rate regions have been characterized. Interesting results can also be reported on the dispersion of Gaussian BC in [17]. Unsal *et al.* have also investigated the Gaussian BC dispersion with superposition coding (SC) [18] leading to an achievable bound. However, several issues may limit the applicability of these bounds to the IoT context starting by the decoding error probability definition which is often a joint probability and thus not suitable for a massive connectivity in an IoT context. Moreover, the achievability bound defined with joint-rate region also limits the insights that an IoT operator may extract from these expressions and existing results are often limited in the number of users considered.

Fundamental bounds with many users MAC have recently been investigated from complementary perspectives. The authors in [19] gave bounds on the joint decoding error probability and capacity region when the number of users grows exponentially or sub-exponentially fast with respect to the number of c.u. and when their communications are asynchronous. The asynchronism and the number of users are linked exponentially with the number of c.u. The main conclusion is that reliable transmission (i.e. vanishing error probability) is impossible when the asynchronism is much more important than the number of users but remains possible when the number of users is sub-exponential with respect to the number of c.u. However, the authors focused on joint decoding error probability and typicality-based decoders. The authors of [20] studied a similar problem to that in [19] but when the number of users K grows linearly with the number of c.u. n, i.e. $K = \mu n$ where μ is the user density. Moreover, the authors considered the per-user decoding error probability criterion, which is a much more relevant metric than the joint decoding error probability, when the number of users is large. The authors gave achievable and converse bounds on the minimal energy per bit for which reliable communication is possible; i.e. vanishing error probability, in the many-user MAC. However, they did not consider second-order expansions as introduced by [8] for Gaussian channels. In [21] the authors defined the many-access channel, which considers a large number of users in the MAC, and they studied the performance when the number of users grows. This work has been further extended in [22] and provided a fundamental limit for the sum-rate. However this model is not connected to the radio cell physical parameters. In [23], the authors explored the fundamental limit of the massive access, taking into account random packet arrivals and decoding error probabilities. The model is quite realistic and complementary to our work because the finite block-length regime is not considered and specific random access policies are evaluated. The impact of random policies has also been investigated in [24] where the information aging is controled.

Compared to these contributions, our work introduces the use of the spatial continuum broadcast and multiple access Gaussian channels (SCBC and SCMAC) [25] to model a spatial density of users and physical channel parameters, associated with the finite block-length analysis, to introduce latency and decoding error probability constraints. This model allows one to obtain an achievable bound for the symmetric rate case in the finite block-length regime.

1.2. Contributions and related work

In the context of IoT, the dense deployment of a large number of nodes in a finite area implies reconsidering the BC/MAC with a spatial distribution of the nodes leading to SCBC/SCMAC models [25,26], well adapted to represent NOMA cellular systems. This new model provided the fundamental EE-SE trade-off under equal-rate conditions in the asymptotic regime. This trade-off can be interpreted as an equivalent of the asymptotic Shannon capacity for a wireless cell with an ultra-dense distribution of users, when every user requests the same rate. The minimal power requested to satisfy a continuum of users using NOMA has been derived with SC, in the asymptotic regime.

While these results provide interesting insights, on the maximal load of dense cells, latency and reliability were kept off the study. Hence, the critical question for IoT networks relates on estimating the price of latency and reliability constraints. As we shall see using FBL formalism, latency and reliability constraints come essentially at the cost of a reduction of the EE-SE region.

To measure this cost, two complementary issues are investigated in this paper. Firstly, to avoid the infinite time transmission induced by the asymptotic regime, a finite time constraint is introduced in the model, following [27] where the idea was introduced. The formal proof is provided with a discussion on its tightness based on simulation results. Secondly, transmission errors associated to transmitting small packets over finite time slots are modeled rigorously in the FBL regime. These two contributions allow to establish an achievable latency-reliability trade-off.

The core of our contribution lies in deriving the minimal requested power to serve a large number of users when the number of channel uses does not tend to infinity. That is only a finite number of users can be superposed contrarily to [25,26]. This approach introduces a scheduling problem that can be reduced to a simpler splitting problem. Moreover, we show that for a given number of superposition levels, i.e. a finite number of splits of the cell, the scheduling order for two users belonging to the same level does not have any influence on the minimal requested power to serve the requested rate density.

The remaining of the paper is organized as follows. Section 2 introduces notations and the system model. Section 3 reviews our previous results on asymptotic SCBC. In Section 4, a finite time constraint (FTT) is introduced. Section 5 deals with decoding error probabilities associated to the FBL regime and estimates the impact on the achievable EE-SE trade-off. Finally, Section 6 draws conclusions and future works.

2. System model

The model described below, was first presented in [25].

2.1. Model and parameters

A unique cell area denoted by $\Omega \subset \mathbb{R}^2$ is served by a unique base station (BS)[1]. $(\Omega, \mathcal{A}, m)$ denotes the corresponding measurable space with $\mathcal{A}$ the Borel $\sigma-$ algebra in Ω and m the Lebesgue measure. Without loss of generality, the BS is assumed to be located at point $(0, 0)$.

The measurable space $(\Omega, \mathcal{A}, m)$ can be extended to $(\Omega \times \mathcal{T}, \mathcal{A}', m')$ where $\mathcal{T} = \mathbb{R}^+$ represents the time and with $\mathcal{A}'$ the Borel $\sigma-$ algebra in $\Omega \times \mathcal{T}$ and m' the associated Lebesgue measure.

Let $U(x, t)$ be defined as a Poisson point process (PPP) on $\Omega \times \mathcal{T}$, which represents the packet request arrivals at position x and time t. Thanks to the stationary properties of PPP, for any subset $B \in \mathcal{A}$, the average number of user requests per time unit for one realization $\tilde{u}(x, t)$ is

$$\lim_{T\to\infty} \frac{1}{T} \int_{t\in T} \int_{x\in B} \tilde{u}(x, t) \cdot \mathrm{d}m'(x, t) = U(B), \tag{3}$$

The global average number of requests per time unit associated with the whole cell service area Ω is denoted by U_T.

Definition 1 (Requested rate density)**.** *The requested rate density $\rho : \Omega \to \mathbb{R}$ is a Borel measurable function that represents the information rate spatial density $\rho(x)$ requested at point x.*

The rate density is expressed in bit-per-channel-use (bpcu) per m^2. A quantity measured in bpcu can be indeed converted into a physical rate for a real system using the number of channel uses per

[1] In this paper, a unique cell is considered without inter-cell interference. For a multi-cells extension, the reader can be referred to [28] where the inter-cell interference from a Poisson Point Process was considered, or to [29] where a fluid model was used. Further, the impact of cell geometry distribution has been explored in [30] But the association of the spatial continuum multi-cell and FBL is beyond the scope of the paper.

time unit, that relies on system parameters such as bandwidth, slots or frames. Clearly, a channel use can be interpreted as a resource element (RE) in an orthogonal frequency division multiple access system. But its meaning is more general, and may also correspond to one channel unit in any other access technology.

In addition, we are interested in this paper with the transmission of small packets. The first idea behind is the transmission of time-constrained small information quantities with no recurrent flows. This aspect is conventional in most papers related to the massive access for IoT. The second idea that relies on an information theory view considers that one packet is transmitted within a small number of channel uses (typically less than few hundreds). Under this assumption, the classical asymptotic regime used in information theory (e.g. [23,24,31]) does not hold and the finite block-length regime needs to be used [19–22]. This constraint increases the difficulty of the mathematical analysis, but also gives access to the latency versus reliability fundamental trade-off.

For the sake of simplicity, we further assume that all packets transport the same information quantity (in bits), denoted by I_0. This scenario is referred as the symmetric-information, by analogy with the widely used property called symmetric-rates in information theory. This assumption allows to keep the mathematical model tractable and is reasonable in many IoT applications.

Under this assumption the requested rate density of the cell relies directly on the user spatial density:

$$\rho(x) = \frac{I_0 u(x)}{N_{cu}}, \tag{4}$$

where N_{cu} is the number of channel uses per time unit, and represents the bandwidth allocated to the system.

The cell sum-rate per channel use is called the spectral efficiency (SE) of the cell:

$$\eta_s := \int_\Omega \rho(x) \mathrm{d}m(x). \tag{5}$$

For the symmetric-information scenario, one ~~have~~ has $\eta_s = \frac{I_0 U_T}{N_{cu}}$.

In order to connect rate estimates to physical parameters of the cell, let be defined the equivalent noise, as the virtual noise level referenced back to the BS, where it matters for power allocation.

Definition 2 (Equivalent noise distribution)**.** *In a given radio cell in the downlink, for any receiver located at position x, the equivalent noise power is given by*

$$\nu(x) := \sigma^2 / g(x), \tag{6}$$

where σ^2 is the receiver noise power and $g(x)$ is the channel power gain associated to this position.

Without any fading nor shadowing, the maximal equivalent noise ν_M is obtained at the cell edge, while the minimal equivalent noise ν_m is obtained in the near field of the BS[2]*.*

The requested rate density $\rho(x)$ is distributed with respect to the equivalent noise associated to each request. Consider the following functions

$$\begin{aligned} G_\nu(\nu) &:= \frac{1}{\eta_S} \int_\Omega \mathbb{1}\left[\nu(x) \geq \nu\right] \rho(x) \mathrm{d}m(x), \text{ and} \\ f_\nu(\nu) &:= -\frac{\mathrm{d}G_\nu(\nu)}{\mathrm{d}\nu}, \end{aligned} \tag{7}$$

[2] Note that shadowing and fading are removed from the analysis for not to clutter the main output of the study, but the latter is general and can be easily extended considering fading and shadowing.

where $\mathbb{1}[.]$ is the indicator function.

$G_\nu(\nu)$ represents the probability with which a packet request is made with an equivalent noise above ν (the most noisy requests). This is nothing but the complementary cumulated density function (ccdf) of the equivalent noise, with respect to the rate requests. Its derivation is therefore the probability density function (pdf) of ν with respect to the rate distribution. The meaning of $f_\nu(\nu)$ and $G_\nu(\nu)$ is illustrated in Fig.1 for a circular cell.

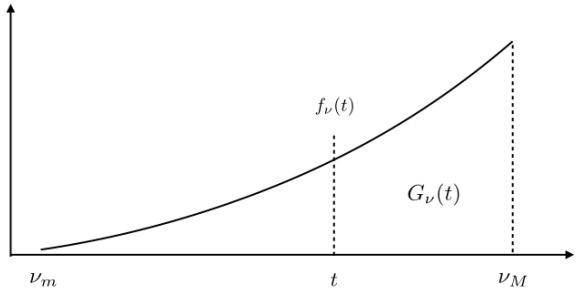


Figure 1. Rate request distribution $f_\nu(.)$ for a regular circular cell.

These definitions provide the key elements to characterize the set of rate distributions that are achievable under some power, latency, spectral efficiency and energy efficiency constraints.

2.2. Reference scenario

Despite the fact that the model is general, the analytical results will be illustrated on a simplified reference scenario, for the sake of clarity, herein described.

The unique cell covers a disk of radius R_c in the downlink mode. Simple power-law pathloss and omnidirectional antennas are considered with no shadowing. Hence, the channel gain is written as

$$g(x) = g_0|x|^{-\alpha}, \tag{8}$$

where g_0 and α represent respectively a reference pathloss and the attenuation slope. $|x|$ is the geometric distance of the point x to the BS in $(0,0)$. For numerical results, the following values are used: $\alpha = 3.65$ and $g_0 = \sigma^2$. Additionally, the transmission power is constrained by a maximal power, i.e. $P \leq P_M$.

The rate demand is uniformly distributed, i.e. $\rho(x) = \rho_0 \; \forall x \in \Omega$ which relies on two assumptions: the symmetric rate hypothesis and a uniform spatial distribution of requests ($u(x) = u_0$). It follows that $\eta_S = m(\Omega)\rho_0 = \pi R_c^2 \rho_0$.

Under these assumptions, the equivalent noise distribution introduced in the former section is given by:

$$f_\nu(\nu) = \frac{2}{\alpha \nu_c}\left(\frac{\nu}{\nu_c}\right)^{2/\alpha - 1}, \quad \forall \nu \in [0, \nu_c] \tag{9}$$

$$G_\nu(\nu) = 1 - \left(\frac{\nu}{\nu_c}\right)^{2/\alpha}, \quad \forall \nu \in [0, \nu_c], \tag{10}$$

with $\nu_c = \frac{\sigma^2}{g_0} R_c^\alpha$ the equivalent noise at the cell edge. The reader may refer to [25] for technical details.

2.3. Superposition coding

SC is capacity achieving for Gaussian BC with successive interference cancellation (SIC) [32]. For a given set of N_u users ordered according to their channel quality, i.e. from the strongest to the weakest, user 1 can only decode its own signal after having decoded the signals sent to users 2 to N_u, user 2

decodes its own signal after having decoded signals from users 3 to N_u and so on. To make the rest of the paper clear, the main steps of SC are herein reviewed.

For two coded messages of length n, X_1^n and X_2^n, assumed to be randomly drawn according to two independent distributions, i.e. $P_{X_1}^n$ and $P_{X_2}^n$, with average powers P_1 and P_2 respectively, the following holds. All decoding steps are done in an equivalent Gaussian channel, where $Z_i^n(k) \sim \mathcal{N}\left(0, \sigma^2 \mathrm{I_d}\right)$, $\forall i \in \{1,2\}$, and according to:

- The second user, with the largest equivalent noise, decodes its own signal in a Gaussian channel given by:

$$Y_2^n = \sqrt{g_2} \cdot X_2^n + \left(\sqrt{g_2} \cdot X_1^n + Z_2^n\right). \tag{11}$$

  The normalized version of this equation is given by:

$$Y_2'^n = Y_2^n / \sqrt{g_2} = X_2^n + \left(X_1^n + \sqrt{\frac{1}{g_2}} \cdot Z_2^n\right). \tag{12}$$

  For this receiver the power of the equivalent additive Gaussian noise is $P_1 + \nu_2$ and its maximum achievable rate in the asymptotic regime, i.e. $n \to \infty$, is $C\left(\frac{P_2}{P_1+\nu_2}\right)$.
- The first user, with the smallest equivalent noise, has two decoding iterations. It first has to decode the second user information in the following channel:

$$Y_1^n = \sqrt{g_1} \cdot X_2^n + \left(\sqrt{g_1} \cdot X_1^n + Z_1^n\right), \tag{13}$$

  with the normalized version given by:

$$Y'^n = X_2^n + \left(X_1^n + \sqrt{\frac{1}{g_1}} \cdot Z_1^n\right). \tag{14}$$

  For this receiver the power of the additive Gaussian noise is $P_1 + \nu_1$ and the achievable data rate is $C\left(\frac{P_2}{P_1+\nu_1}\right)$. Then, after cancelling the second user signal, receiver one decodes its own signal:

$$Y_1''^n = X_1^n + \sqrt{\frac{1}{g_1}} \cdot Z_1^n, \tag{15}$$

  and can achieve its full achievable data rate, i.e. $C\left(\frac{P_1}{\nu_1}\right)$, in the asymptotic regime.

3. Symmetric capacity in the asymptotic regime

3.1. Fundamental trade-off with SC

When SC is used, BS waits for a time T to aggregate a set of packet requests that are transmitted in the next slot to the corresponding nodes in n channel uses. Under no latency constraint, the time T can be taken arbitrarily large allowing to verify $n \to \infty$, corresponding to the asymptotic regime. The study of this regime leads to the access capacity region defined in [25] as the set of rate spatial distributions $\rho(x)$ for which an encoder-decoder pair exists such that the transmission error tends to 0 when T tends to infinity.

In comparison with Shannon's asymptotic regime, our model adds a complementary parameter: when $n \to \infty$, the number of randomly distributed nodes (each node represents a message request) tends to infinity. Since we do not consider individual rates, but individual fixed information quantities

I_0, the sum-rate converges to $U_T I_0$ while individual rates tend to 0 as they are equal to I_0/n. This is illustrated in Fig. 2. It is worth noting that cell's sum-rate tends to its average spectral efficiency. In [25], based on an iterative splitting process, the maximal sum-rate the cell can achieve when a continuum of users is considered has been established. The corresponding fundamental limit is expressed as:

Theorem 1 (GSCBC fundamental limit [25]). *The achievable EE-SE trade-off for a given rate spatial density $\rho(x)$ is given by*

$$\eta_E \leq \left[a \int_{\nu_m}^{\nu_M} t f_\nu(t) \cdot e^{a\eta_S G_\nu(t)} dt \right]^{-1}, \tag{16}$$

where $a = 2\log(2)$ and η_E is the energy-efficiency.

This result can be applied to the reference scenario described in Section 2.2, using (9) in (16), which yields

$$\tilde{P}_m = \nu_M \frac{a\eta_S}{1+\alpha/2} {}_1F_1(1; 2+\alpha/2; a\eta_S), \tag{17}$$

with ${}_1F_1(a;b;x)$ the confluent hypergeometric function [33, sec. 9.21]. $\tilde{P}_m$ is the minimal transmission power required at BS to serve the rate spatial density $\rho(x)$.

The fundamental EE-SE limit of the corresponding cell is provided by (17). Given the power normalized by the equivalent noise at the cell edge $p_r = \tilde{P}_m/\nu_c$, EE in bppu is defined as $\eta_E = \eta_S/p_r$, leading to the fundamental EE-SE limit

$$\eta_E = \frac{1+\alpha/2}{a \cdot {}_1F_1(1; 2+\alpha/2; a\eta_S)}. \tag{18}$$

The EE should be understood as the total number of bits the base station can transmit under a transmission power constraint expressed as the relative sum-power received by an edge user. So, the term $1/\eta_E$ plays for the symmetric SCBC the role of the classical E_b/N_0 for a point-to-point link. Clearly, (18) is equivalent to the Shannon's second theorem in (1), for the symmetric SCBC. The symmetric SCBC capacity $C(\gamma)$ is obtained by inverting (17) with respect to η_S and denoting by $\gamma = \frac{\tilde{P}_m}{\nu_M}$, the SNR at the cell edge.

3.2. Fundamental trade-off with orthogonal sharing

A classical alternative to SC, is to exploit orthogonal multiple access (OMA), e.g. by time division. In this case, to maximize the symmetric-information, the BS allocates a fixed number of channel uses to each packet, and adapt the transmission power to constantly preserve the spectral efficiency. The transmission power used for a node in x is:

$$\frac{1}{a}\log\left(1+\frac{P(x)}{\nu(x)}\right) = \eta_s \tag{19}$$

Lemma 1 (Achievable bounds with OMA). *In a single-cell under the spatial continuum model, the fundamental EE-SE trade-off achievable by orthogonal multiple access is*

$$\eta_E = \frac{1}{2}\frac{2+\alpha}{e^{a\eta_S}-1}\eta_S. \tag{20}$$

Proof. See Appendix A □

The corresponding curve is given on Fig. 4 where the EE-SE limit is represented. The blue curve represents the EE-SE fundamental limit achievable with OMA with $\alpha = 3.65$, and the red curve the fundamental limit established with NOMA (SC), plotted using Theorem 1.

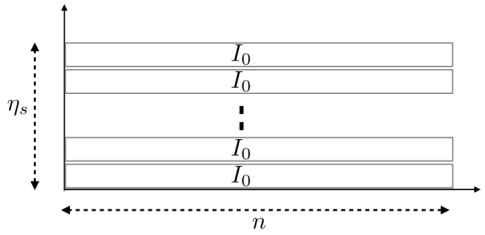


Figure 2. The asymptotic regime is obtained at the limit when $n \to \infty$. The cell spectral efficiency is kept constant, but the number of packets transmitted simultaneously tends to infinity. Each packet I_0 is spread over n channel uses and the individual spectral efficiency tends to 0.

To sum-up, this section reported the fundamental EE-SE limit of the SCBC in the asymptotic regime as derived in [25]. Note that the *asymptotic regime* refers to a doubly asymptotic regime. Indeed, when $n \to \infty$, it follows that $\epsilon \to 0$ but with the SCBC, the number of nodes transmitting simultaneously also tends to infinity with individual rates going to 0. But the sum-rate converges to the SCBC capacity.

4. Finite time transmission constrained model

The objective of this paper is to introduce a transmission time constraint into the former model to obtain achievable bounds of NOMA under more realistic assumptions than the doubly asymptotic regime.

Consider the situation in which each packet has to be transmitted in a finite time $T \in \mathbb{N}$, i.e. within a finite number of channel uses. For the moment, we still consider arbitrary low error probabilities, sustained with γ and n sufficiently high. This hypothesis will be relaxed in Section 5.

4.1. FTT formulation

Definition 3 (Finite-time Transmission Constraint)**.** *A multi-user network with packets of I_0 bits, with $I_0 \in \mathbb{N}$, is said to be FTT constrained when each transmission lasts at most n^* channel uses.*

Note that this constraint only imposes that a packet of I_0 bits is transmitted in at most n^* channel uses but the queuing delay is not controlled. The FTT constraint is then a necessary but not sufficient condition for delay constrained transmissions. The FTT constraint provides interesting insights anyway. For instance, it allows to setup the transmission duration of each packet thereby controlling the activity time of each receiver.

To assess the symmetric rate fundamental limit of the cell under the FTT constraint, let us review that the average spectral efficiency (i.e. the sum-rate) of the cell noted $\bar{\eta}_s$ shall be equal to:

$$\bar{\eta}_s = \frac{I_0 U_T}{N_{cu}}. \tag{21}$$

When the BS transmits a packet of I_0 bits in n^* channel uses, the individual rate for this packet is $\eta_u = \frac{I_0}{n^*}$. In order to achieve the target spectral efficiency $\bar{\eta}_s$, the BS has to use SC to transmit simultaneously several packets. Therefore, the FTT constrained problem is equivalent to the following scheduling problem:

Definition 4 (SC Scheduling policy)**.** *Given:*

- *a frame of N_{cu} channel uses of duration T, itself divided into L slots, each slot $s_l; \forall l \in \{1, 2, \ldots, L\}$ contains n^* channel uses. One has $N_{cu} = Ln^*$,*
- *a BS's queue containing a random number of packets to be transmitted to a set of nodes $\mathcal{N}_U$, selected according to the PPP $U(x, t)$ restricted to the subset $\Omega \times T$,*

A SC scheduling policy selects a subset of users $\mathcal{N}_u(l) \subset \mathcal{N}_U$ for each slot s_l, which are ordered with their increasing equivalent noise, i.e. $\nu_{k+1}(l) > \nu_k(l)$. Decoding is performed at each user, according to the SC technique.

The number of users associated to each slot s_l, is noted $N_u(l) = |\mathcal{N}_u(l)|$ and the corresponding spectral efficiency is :

$$\eta_S(l) = N_u(l)\eta_u. \tag{22}$$

4.2. Optimal scheduling policy

We now propose to determine an optimal scheduling policy in the asymptotic regime.

Definition 5 (Optimal scheduling policy)**.** *A scheduling policy for the PPP $U(x,t)$ over $\Omega \times T$, is asymptotically optimal under a FTT constraint, if all user requests are served within n^* channel uses at most, and if the transmission power is minimal over all possible scheduling policies, when $T \to \infty$.*

Note this asymptotic regime is conditioned on the FTT constraint and is thus more constrained than the regime studied in section III.

Let $\gamma_k(l)$ be the effective signal-to-interference-plus-noise-ratio (SINR) for node $u_k(l)$, defined as :

$$\gamma_k(l) = \frac{P_k(l)}{\nu_k(l) + \sum_{i<k} P_i(l)}. \tag{23}$$

This SINR is effective when the appropriate decoding is used and thanks to the superposition coding principle.

Since each user needs to get a reliable individual rate $\eta_u = I_0/n^*$, its effective SINR $\gamma_k(l)$ needs to verify:

$$\gamma_k(l) \geq \gamma^*, \forall(k,l), \tag{24}$$

where $l \in \{1,2,\ldots,L\}$ and $k \in \{1,2,\ldots,N_u(l)\}$ and with $\gamma^* = e^{a\eta_u} - 1$, according to the channel capacity theorem.

It is then straightforward to say that the optimal power used for each packet in slot s_l are given by:

$$\begin{aligned} P_1(l) &= \gamma^* \cdot \nu_1(l), \\ P_2(l) &= \gamma^* \cdot (\nu_2(l) + P_1(l)), \\ &\vdots \\ P_{N_u(l)} &= \gamma^* \cdot \left(\nu_{N_u}(l) + P_1(l) + \cdots + P_{N_{u,l}-1}(l)\right). \end{aligned} \tag{25}$$

The BS transmission power for slot l is then $P_m(l) = \sum_k P_k(l)$.

Then, in this symmetric rate setup, where all nodes require the same SINR, the following Lemma holds, where the slot numbering is omitted for the sake of clarity.

Lemma 2 (Minimum sum-power in the symmetric Gaussian BC)**.** *Given a set of users indexed by $k \in \{1,\ldots,N_u\}$, ordered such that $\forall k, \nu_k \geq \nu_{k-1}$, the minimum sum-power necessary to transmit reliably to each node independent information I_0, under effective SINR γ^*, is :*

$$P_m = \sum_{\bar{k}=1}^{N_u} d_{\bar{k}} \cdot \nu_{\bar{k}}, \tag{26}$$

with $d_{\bar{k}} := \gamma^*(1+\gamma^*)^{\bar{k}-1}$, *and* $\bar{k} = N_u - k + 1$.

Proof. The proof relies on the decomposition of P_k according to (25), i.e. $P_k = \sum_{j=1}^{k} c(k,j) \cdot \nu_j$, where the $c(k,j)$, represented in Table 1, are given by:

$$\begin{aligned} c(k,j) &= 0; \quad \forall j > k, \\ c(k,k) &= \gamma^*, \\ c(k,j) &= \gamma^* \sum_{i=j}^{k-1} c(i,j); \quad \forall j < k. \end{aligned} \tag{27}$$

	ν_1	ν_2	ν_3	$\dots$	ν_{N_u}
$P_1:$	γ^*			$\dots$	
$P_2:$	γ^{*2}	γ^*		$\dots$	
$P_3:$	$\gamma^{*3}+\gamma^{*2}$	γ^{*2}	γ^*	$\dots$	
$P_4:$	$\gamma^{*4}+2\gamma^{*3}+\gamma^{*2}$	$\gamma^{*3}+\gamma^{*2}$	γ^{*2}	$\dots$	
$\vdots$					

Table 1. Power series coefficients $c(k,j)$ (k^{th} row, j^{th} column) representing the coefficient of the j^{th} equivalent noise into the k^{th} power term.

These coefficients $c(k,j)$, can be computed recursively, with $c(k,k) = \gamma^*$ and $c(k,k-1) = \gamma^{*2}$, and the following recursion for $j < k-1$:

$$\begin{aligned} c(k,j) &= \gamma^* \sum_{i=j}^{k-1} c(i,j), \\ &= \gamma^{*2}(1+\gamma^*)^{k-j-1}. \end{aligned} \tag{28}$$

In Table 1, each row represents a decomposition of the power of one message, with respect to the equivalent noise $\nu(k)$ of all users. In parallel, the sum-power can be computed column-wise first, leading to:

$$P_m = \sum_{k=1}^{N_u} d(k) \cdot \nu_k, \tag{29}$$

with $d(k) = \sum_{i=k}^{N_u} c(i,k)$. Using (27), these coefficients can be straightforwardly rewritten as:

$$d(k) = \gamma^{*-1} \cdot c(N_u+1, k). \tag{30}$$

Then, with (28), one obtains $d(k) = \gamma^*(1+\gamma^*)^{N_u-k}$ leading to the final expression:

$$P_m = \gamma^* \sum_{k=1}^{N_u} \nu_k \cdot (1+\gamma^*)^{N_u-k}. \tag{31}$$

Now, numbering the nodes in the reverse order (noted $\bar{k}$ for clarity) i.e. from the farthest to the nearest one, ends the proof. □

This theorem shows that P_m is a linear combination of the equivalent noises $\nu_{\bar{k}}$ weighted with $d_{\bar{k}}$. Then, each coefficient relies only on γ^* and grows exponentially with $\bar{k}$. The optimal strategy, which minimizes P_m, should obviously allocate the users according to their channel quality.

We draw the reader's attention to the fact that the $\bar{k}$-th term of the sum in (26) should not be interpreted as the power used to transmit the $\bar{k}$-th message but as the additional power induced by the $\bar{k}$-th equivalent noise level in the sum-power. The power associated to each message is given by (25). Nevertheless, the linear relation (26) of Lemma 2 is more appropriate to demonstrate the optimality of the proposed scheduling policy.

Consider a set of L slots and a set of users $\mathcal{N}_U$ requesting a message, with $N_U = |\mathcal{N}_U|$. A scheduling policy associates $N_u(l)$ users for each slot s_l. Let us recall that according to our definition, the users $u_1(l) \dots u_{N_u(l)}(l)$ are ordered with respect to to their equivalent noise. Using the notation of Lemma 2, we refer to $\bar{k}$ as the coding level. A message encoded at level $\bar{k}$ means that the corresponding receiver needs to decode first the packets of lower level.

Definition 6 (Natural ordering policy). *Assume that $N_u(l) = N_u, \forall l$, and then $N_U = L \times N_u$. This comes without loss of generality, as shown at the end.*

Let now the nodes in $\mathcal{N}_U$ be ordered from the strongest to the weakest equivalent noise, from u_1 to u_{N_U}. The natural ordering policy proceeds by assigning the users u_1 to u_L to the first coding level over the L slots. Once the first coding level is filled out, the second level is filled and so on up to the last coding level. It follows that :

$$u_{\bar{k}}(l) = u_{l+(k-1)L}, \tag{32}$$

This scheduling policy is illustrated in Table 2.

		Slots				
		1	2	3	...	L
$\bar{k}$	N_u	u_{N_U-L+1}	u_{N_U-L+2}	u_{N_U-L+3}	...	u_{N_u}
	⋮	⋮	⋮	⋮	⋮	⋮
	2	u_{L+1}	u_{L+2}	u_{L+3}	...	u_{2L}
	1	u_1	u_2	u_3	...	u_L

Table 2. Optimal scheduling policy allocating the users through coding levels and slots.

Theorem 2 (Optimal scheduling). *For a given set of users indexed by $\{1, \dots, N_U\}$ and ordered from the strongest to the weakest equivalent noise, the natural ordering policy is optimal with respect to the average transmission power.*

Proof. From Lemma 2, it follows that the natural ordering must be used. The remaining question is about the repartition of the users through the different slots.

To prove that the natural ordering policy is optimal, let us consider another policy, for which one of the first L users noted u_i is not allocated to the first coding level, but to the level $\bar{k}_i$. Then, there exists a user u_j allocated to the first coding level, such that $j > L$. Then a simple permutation noted $\pi(u_i, u_j)$ is sufficient to reduce the sum power, since the equivalent noise of u_j is lower than that of u_i, and the power difference between the two policies is:

$$\Delta P = \left(d_{\bar{k}_j} - d_{\bar{k}_i}\right)\left(\nu_i - \nu_j\right), \tag{33}$$

which is strictly negative.

So starting from any policy, moving all the L first users with permutations to the first coding level reduces the sum-power. Then, proceeding the same way with the higher order coding levels, will also reduce the power. At the end of these permutations, each coding level $\bar{k}$ contains the same users as the natural ordering policy.

It should be also noted that any permutation between two users at the same coding level does not change the sum power. Therefore, the natural ordering policy, is one of the optimal ones. It is

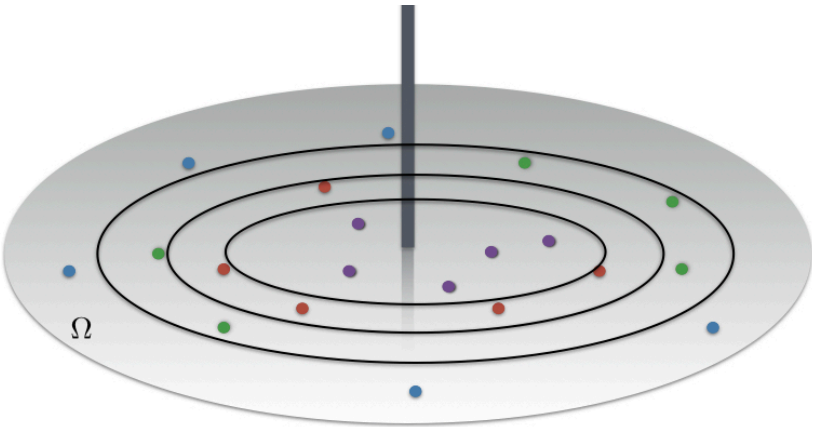


Figure 3. The optimal scheduling requires to divide the cell in equal rate subsets as a function of their channel quality.

worth noting that when doing such a permutation, the individual power allocated to each message may change, but with no impact on the sum-power. This complete the proof. □

Finally, if $N_U = L \times N_u$ is not verified, the same permutations can be used and one obtain a policy where the last coding level is partially filled in. In this case, the number of levels (or superposition codes) is given by $\lceil \eta_S / \eta_u \rceil$.

4.3. Optimal scheduler when $T \to \infty$

According to the previous result the optimal scheduler relies on transmitting at each round to exactly N_u users with $N_u = \frac{\eta_s}{\eta_u}$. Let assumed this ratio be an integer. If it is not the case, alternate rounds may be used with $\lfloor \frac{\eta_s}{\eta_u} \rfloor$ and $\lfloor \frac{\eta_s}{\eta_u} \rfloor + 1$ users.

The optimal scheduling policy is enforced by partitioning the cells in N_u subsets of equivalent sum-rates:

$$\tilde{\mathcal{B}} := \{B_1, \dots B_{N_u}\}, \tag{34}$$

where $B_{\bar{k}} := \{x; \nu(x) \in [\nu_{\bar{k}}; \nu_{\bar{k}-1})\}$, as illustrated in Fig. 3 for a regular circular cell. The thresholds ν_k are defined with $\nu_0 = \nu_M$, $\nu_{N_u} = \nu_m$, and such that $|B_{\bar{k}}| = U_T / N_u; \forall k$ with U_T the total number of requests. Once this partition is done, at each slot, the BS picks up a user per subset and transmits to these users with SC.

This scheduler achieves the minimal average power when T (and L as well) tends to infinity. Indeed, the partition $\mathcal{B}^{(\infty)}$ converges to a partition where all $B_{\bar{k}}$ are of equal surface (due to the properties of the uniform PPP model) and the asymptotic average power is given by:

$$\bar{P}_m = \lim_{L \to \infty} \frac{1}{L} \sum_{l=1}^{L} \sum_{\bar{k}=1}^{N_u} d_{\bar{k}} \cdot \nu_{\bar{k},l}, \tag{35}$$

which can be expanded as:

$$\begin{aligned} \bar{P}_m &= \sum_{\bar{k}=1}^{N_u} d_{\bar{k}} \cdot \mathbb{E}_{x \in B_{\bar{k}}^{(\infty)}}[\nu(x)], \\ &= \sum_{\bar{k}=1}^{N_u} d_{\bar{k}} \cdot \mathbb{E}_{\nu \in [\nu_{\bar{k}}; \nu_{\bar{k}-1})}[\nu] = \sum_{\bar{k}=1}^{N_u} d_{\bar{k}} \cdot \bar{\nu}_{\bar{k}}, \end{aligned} \tag{36}$$

where $\bar{\nu}_k$ stands for the average equivalent noise over the $\bar{k}$-th subset.

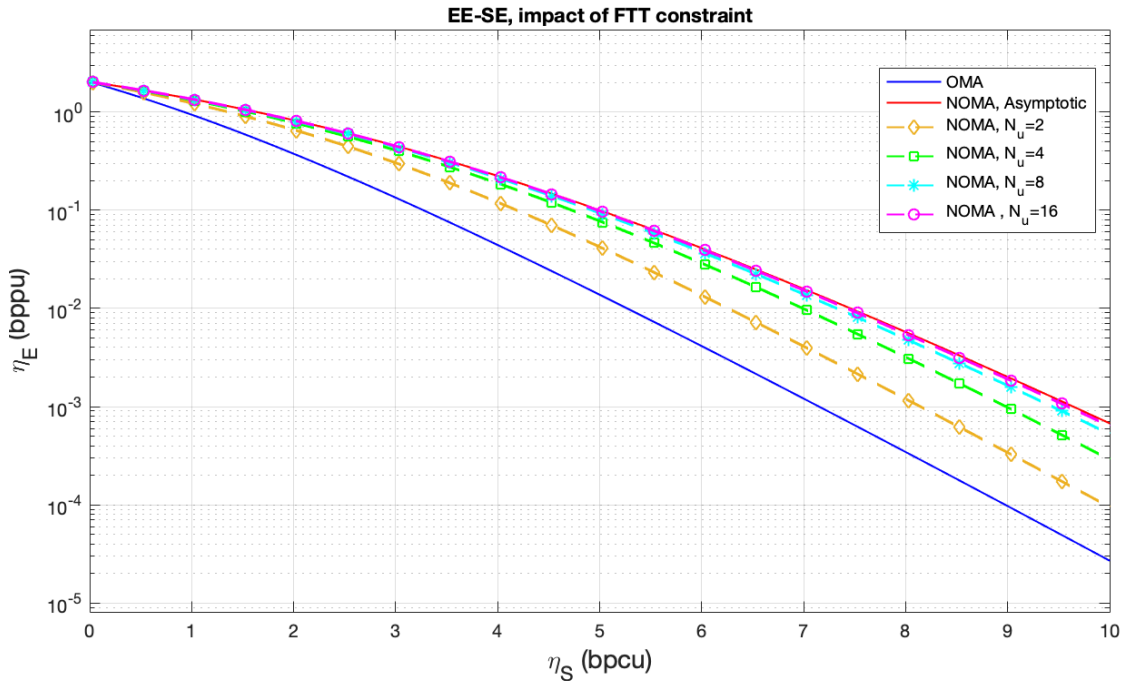


Figure 4. $\eta_E - \eta_S$ Trade-off for $N_u \in \{2, 4, 8, 16\}$.

Using the expression of $d_{\bar{k}}$ and expanding γ^*, one obtains

$$\bar{P}_m = (e^{a\eta_u} - 1) \cdot \sum_{\bar{k}=1}^{N_u} e^{a\eta_u(\bar{k}-1)} \bar{\nu}_{\bar{k}}, \tag{37}$$

$$\approx a\eta_s \cdot \sum_{\bar{k}=1}^{N_u} \frac{e^{a\eta_u(\bar{k}-1)} \bar{\nu}_{\bar{k}}}{N_u}. \tag{38}$$

The last approximation comes when the number of subsets is sufficiently large such that $a\eta_u << 1$.

Interestingly, this result can be compared to the fundamental limit established in Theorem 1. These expressions are similar, except that the continuous integral has been replaced by a discrete sum, and the equivalent noise ν by $\bar{\nu}_{\bar{k}}$. The term $a\eta_s G_\nu(t)$ in the exponential is replaced by its discrete version $a\eta_u k$ and $f_\nu(t)$ by $1/N_u$. It is then straightforward to show that Theorem 1 is obtained as the limit of (38) when N_u tends to infinity, i.e. when the constraint $n^* \to \infty$, which proves the doubly asymptotic optimality of this scheduler.

4.4. Application example

The former analytical results are applied to the reference scenario of section 2.2 and represented in Fig. 4 with the cross curves for different numbers of SC layers (indicated by N_u). Moreover, the path loss exponent is $\alpha = 3.65$, the reference path loss and the noise power are normalized, i.e. $g_0 = \sigma^2 = 1$. The two asymptotic curves are given in blue for OMA, i.e. with Lemma 1 and red for NOMA, i.e. with Theorem 1. The orange curve with diamonds corresponds to the 2-user NOMA. The green, cyan and magenta curves with square, stars or circles are obtained with 4, 8 and 16−user NOMA respectively. These curves are obtained by applying (37) where η_u being simply the total spectral efficiency η_S divided by the number of users, since all users receive the same amount of information. These curves highlight the performance loss when the number of superposed codes is equal or lower than 4. This model also shows that the fundamental limit established for NOMA in the asymptotic regime is almost

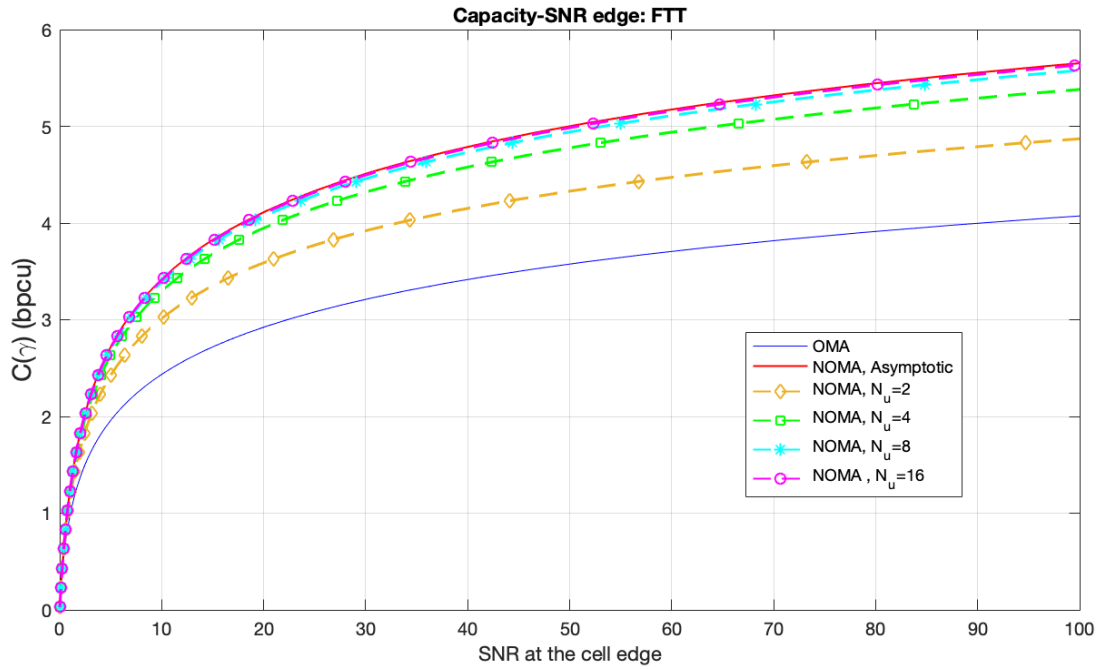


Figure 5. Cell capacity $C(\gamma)$ under symmetric-information for a given SNR at cell edge (red), achievable symmetric-rate with OMA (blue) or under optimal partitioning in N_u equal sum-information regions, with $N_u \in \{2, 4, 8, 16\}$.

achievable with a reasonable number of coding levels. 90% of the gain is achieved with only 4 coding levels, and even 30% of the gain is achieved with 2 coding levels.

In addition, the capacity of the cell is represented in Fig. 5 in the same conditions. In both figures, one can remark the sub-optimality of OMA in the doubly asymptotic regime. Moreover, one can also observe the quick convergence to the optimal performance, i.e. the EE-SE Pareto front in Fig. 4 and the asymptotic cell capacity in Fig. 5, with the number of partitions of the cell.

Both figures highlight the interest of our fundamental limit given by theorem 1, being almost achievable with a NOMA strategy.

5. Finite block-length (FBL) constrained Model

The last step addressed in this section to compare practical NOMA schemes to the fundamental limit, is to relax the error-free assumption to cope with the FBL regime, more appropriate for small packets.

We herein develop an approximation of the achievability bound with a NOMA scheme (SC) by exploiting the normal approximation derived in [8] for a point-to-point transmission and reviewed in (2).

For a fixed number of bits I_0 to be transmitted, (2) leads to:

$$I_0 \approx nC(\gamma) - Q^{-1}(\epsilon) \cdot \sqrt{n \cdot V(\gamma)}, \tag{39}$$

which provides a relationship between P_m (through γ), n and ϵ. Consider a 2-users BC before the generalization to N_u-users BC. In the following, we denote $\epsilon_{i,j}$ the decoding error probability of message j by user i.

5.0.1. Achievable minimal power for the 2-user Gaussian BC

Let us review that in our setup, the BS aims at transmitting two independent packets of I_0 bits each, to two users in Gaussian channels in at most n channel uses and with an average individual error probability lower than ϵ^* for each user, i.e. $\epsilon_i \leq \epsilon^*, \forall i \in \{1, \cdots, N_u\}$.

Considering the targeted rate is $R^* = I_0/n$ and assuming the interference caused by the other users to be Gaussian, we can write for the weakest user:

$$C(\gamma_2^*) - Q^{-1}(\epsilon^*)\sqrt{\frac{V(\gamma_2^*)}{n}} \geq \frac{I_0}{n}. \tag{40}$$

Contrarily to the asymptotic situation described in Section 4, the target SNR value needs to be adapted for each user as a consequence of the SC technique.

Considering user 2, γ_2^* is obtained as the unique solution of (40) for some tuple (n, ϵ^*, I_0). The solution is unique because this equation is monotonically increasing with respect to γ. This imposes the following relation between P_1 and P_2 :

$$P_2 \geq \gamma_2^*(P_1 + \nu_2). \tag{41}$$

Now considering user 1, it first decodes message 2 with a lower error noted $\epsilon_{1,2} < \epsilon^*$, because the SNR is stronger.

By the union bound, the decoding error probability of user 1 is bounded by the sum of the decoding errors associated to the two messages $\epsilon_1 \leq \epsilon_{1,1} + \epsilon_{1,2}$. Then, to keep a global error probability lower than ϵ^*, the error probability on its intended message $\epsilon_{1,1}$, should satisfy:

$$\epsilon_{1,1} \leq \epsilon^* - \epsilon_{1,2}. \tag{42}$$

Therefore, the minimum required SNR for the strongest user, γ_1^*, is the solution of:

$$C(\gamma_1^*) - Q^{-1}(\epsilon^* - \epsilon_{1,2})\sqrt{\frac{V(\gamma_1^*)}{n}} = \frac{I_0}{n}, \tag{43}$$

which is bigger than γ_2^*, because the error constraint is stronger.

Solving these equations provide the minimal transmission powers P_1 and P_2 as:

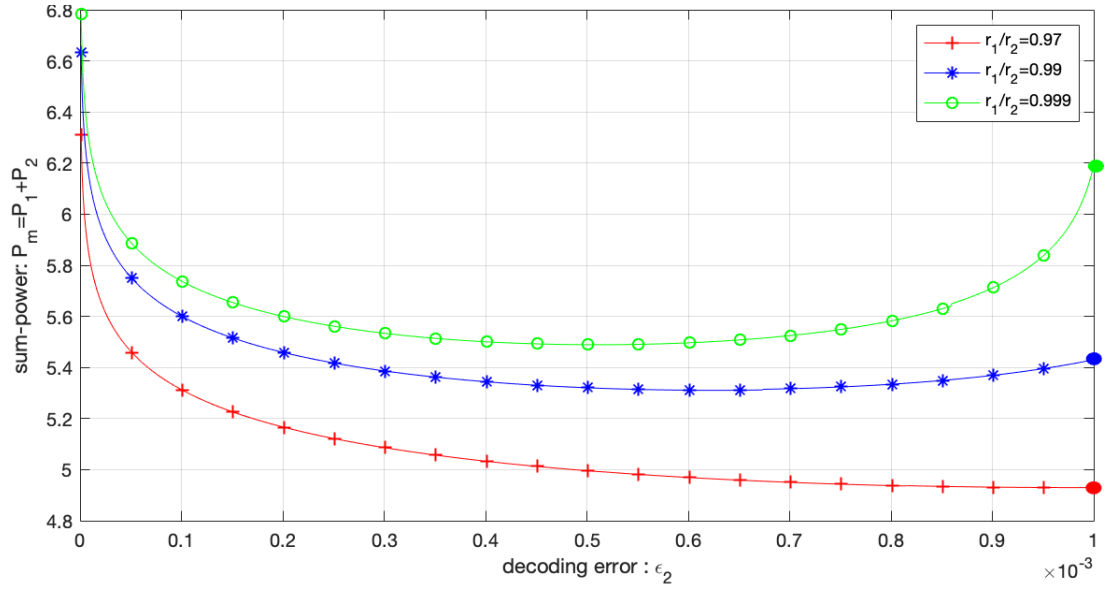
$$\begin{aligned} &P_1 = \gamma_1^* \cdot \nu_1, \text{ and} \\ &P_2 = \gamma_2^* \cdot (P_1 + \nu_2) = \gamma_2^* \cdot (\gamma_1^* \cdot \nu_1 + \nu_2). \end{aligned} \tag{44}$$

Although an analytic expression cannot be written, numerical computation is straightforward.

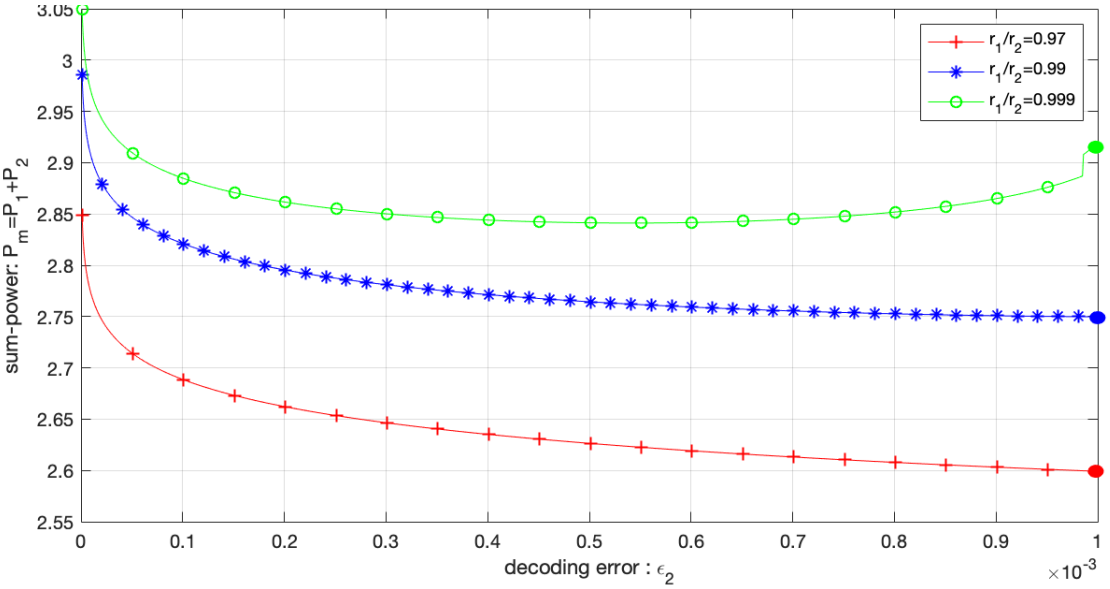
5.0.2. Impact of the power sharing between P_1 and P_2

In (40), we determined the minimal power allowing to achieve the error target on user 2. But the use of a larger power P_2 could be justified from a theoretical point of view, since it would allow to reduce $\epsilon_{1,2}$, then allowing to reduce P_1 as the solution to (43). The influence of reducing ϵ_2 on the sum-power is illustrated in Fig. 6 for the simulation parameters described in Section 2.2 and for a target individual error probability $\epsilon^* = 10^{-3}$. The sum-power is plotted for different information size and block-lengths ($(I_0 = 40, n = 100)$, and $(I_0 = 400, n = 1000)$). Each curve is obtained when the users are positioned at distance r_1 and r_2 from the BS.

The reference solution obtained with $\epsilon_2 = \epsilon^*$ is on the right of each plot (indicated with a plain circle). A sum-power reduction by increasing P_2 and reducing P_1 exists but is significant only when r_1/r_2 approaches 1. Clearly, when SC is used for users with significantly different positions, the reference solution is nearly optimal. This is justified because when the SNRs of the two users are sufficiently different, then $\epsilon_{1,2} << \epsilon_2 = \epsilon^*$, and therefore the impact of $\epsilon_{1,2}$ in (43) is negligible.

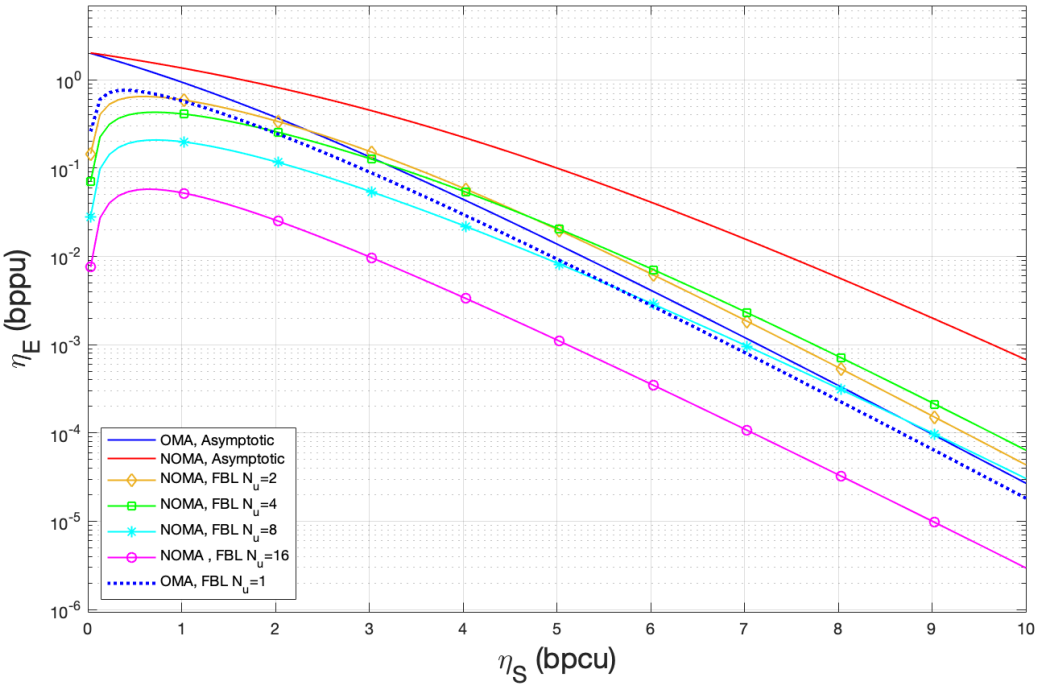


(a) Sum-power with $I_o = 40$, $n = 100$, $\epsilon^* = 0.001$.

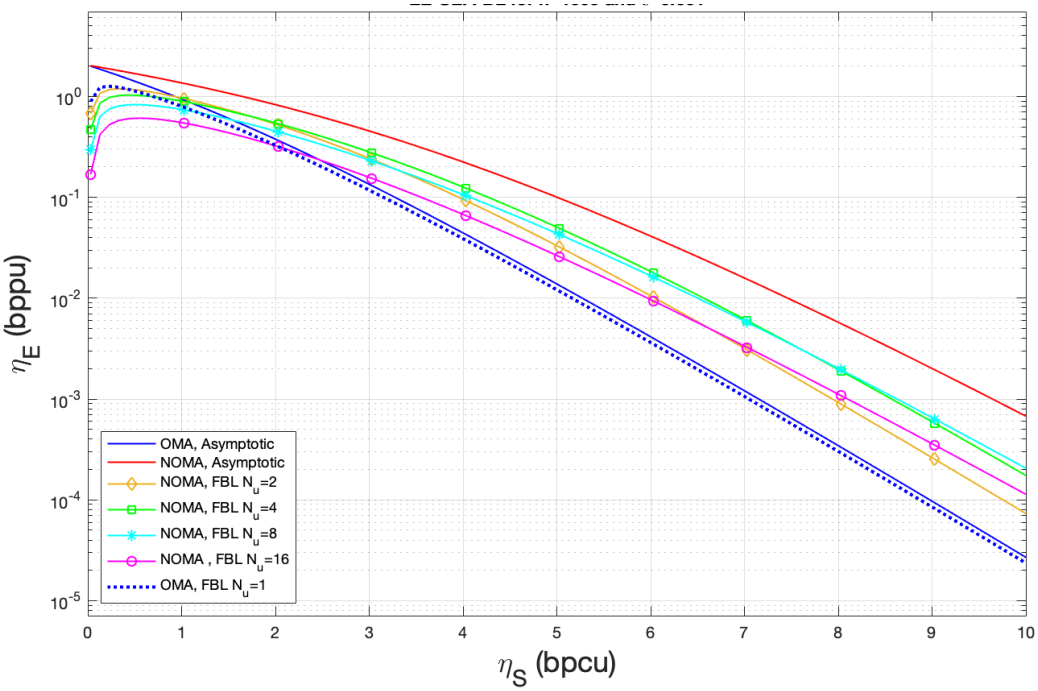


(b) Sum-power with $I_o = 400$, $n = 1000$, $\epsilon^* = 0.001$.

Figure 6. Minimal sum-power $P_m = P_1 + P_2$ for the 2-user Gaussian BC with SC, for different ratios between BS-user distances r_1/r_2 and with respect to the decoding error probability of user 2 ϵ_2. The plain circles indicate the reference solution when $\epsilon_2 = \epsilon^*$. Reducing ϵ_2 induces an increase of P_2 but permits a decrease of P_1. When the reduction in P_1 is larger than the increase in P_2, a power gain can be obtained. This gain is more significant for small packets. Note that the vertical scale has been zoomed to highlight the sum-power variations. Indeed the variations are relatively small compared to the effective sum-power values.



(a) EE-SE for $n = 100$ and $\epsilon = 10^{-3}$



(b) EE-SE for $n = 1000$ and $\epsilon = 10^{-3}$

Figure 7. EE-SE trade-off for different number of channel uses n and for a per-user probability of error $\epsilon = 10^{-3}$.

5.0.3. Achievable power for the N-user BC

Extending the former result to the N-user Gaussian BC is straightforward with SC, when the power of each user is optimized according to (43). At each level, an additional penalty on the error is introduced. $\gamma^*_{\bar{k}}$ is thus solution to:

$$C(\gamma^*_{\bar{k}}) - Q^{-1}(\epsilon^* - \sum_{j=1}^{\bar{k}-1} \epsilon_{\bar{k},j}) \sqrt{\frac{V(\gamma^*_{\bar{k}})}{n}} = \frac{I_0}{n}. \tag{45}$$

The sum introduced in the Q^{-1} function shows how the error probabilities accumulate, which is the key issue of a SC approach in FBL regime.

The consequence for the sum-power follows from the iterative relations:

$$P_{\bar{k}} \geq \gamma^*_{\bar{k}} \cdot \left(\nu_{\bar{k}} + \sum_{j=\bar{k}+1}^{N_u} P_j \right), \tag{46}$$

where $\gamma^*_{\bar{k}}$ is the solution of (45).

The NOMA achievable EE-SE trade-off for two different block-lengths ($n = 100$ or $n = 1000$) and an individual error probability threshold $\epsilon^* = 10^{-3}$ is represented in Fig. 7 with the iterative power allocation described above for a number of coding levels in $N_u \in \{1, 2, 4, 8, 16\}$. The EE-SE trade-off of OMA in asymptotic and FBL regimes is also plotted for reference.

Clearly, for small block-length ($n = 100$, Fig. 7(a)), the achievable region shrinks the most with the 16-user SC due to the impact of error accumulation. The best FBL SC configuration is the 4-user SC (green curve), in moderate to high spectral efficiency regime. The 2-user SC is almost optimal in these regimes. In the low spectral efficiency regime (below 1), OMA (dotted blue line) outperforms NOMA. When the block-length is larger ($n = 1000$, Fig. 7(b)), OMA remains optimal in the low spectral efficiency regime. But in the moderate to high spectral efficiency regime, the degradation of SC reduces significantly, and all NOMA schemes outperform OMA. In this situation, the 4-user or 8-user SC are the best ones.

An important conclusion is that SC is inappropriate for very small packets at low SNR. This is in line with [20] that pointed out the better performance of OMA when the density of users $\mu = K/n << 1$, with K the number of users, compared to a full decoder. Note that [20] considered a MAC scenario while we are considering BC in this paper. However, thanks to the MAC-BC duality the conclusions could be easily transposed to the MAC scenario because in both cases, successive decoding is used as a baseline.

6. Conclusions

In this paper, we proposed an analytic model to evaluate the performance of NOMA with many users when the transmission time is constrained and when small packets are transmitted. For that, we merged the spatial continuum model introduced in [25] with the finite block-length second-order rate expansion limit introduced by Polyanskiy *et al.*.

We first show that the fundamental limit obtained with the spatial continuum model is relevant as this fundamental limit can be reached with a reasonable number of superposition coding layers when the messages are transmitted over large block-lengths. This result justifies the use of the proposed fundamental limit (Theorem 1) to optimize the design of cellular networks for NOMA IoT cells.

By exploiting a SC scheme in FBL, we further show the performance degradation when n reduces below 1000. However, it is worth mentioning that our FBL analysis relies on assumptions which prevent us from claiming that NOMA is necessarily worst than OMA in FBL. Indeed, i) we used the normal approximation, ii) we impose a SC strategy and iii) we used a sub-optimal reference power allocation in SC.

Even if the normal approximation has been observed to be tight by simulation, the classical Berry-Esseen bounds are not sufficient to prove this tightness [8,16,17]. The recent paper [34] explores the tightness of saddle-point approximations for the P2P channel and could be used in the future to determine tighter bounds for the N-user BC. Nevertheless, additional simulations not presented here for the sake of consistency show that the degradation of SC in the FBL is not due to this approximation.

Concerning the SC strategy, clearly responsible of the performance degradation due to the successive decoding algorithm, an open question is to determine if a dirty paper coding technique in FBL could outperform the SC technique. Answering this question may rely on [17,35].

Author Contributions: The initial model of the spatial continuum has been proposed by Jean-Marie Gorce as a visiting scholar at Princeton University in collaboration with Prof. H. Vincent Poor. The extension developed in this paper has been proposed by Jean-Marie Gorce and Philippe Mary in the framework of the ANR Arburst project and through fruitful discussions with Jean-Marc Kélif who contributed to the development of section IV. Dadja Anade contributed to the derivation of the results of section V and produced the simulation results used throughout the manuscript.

Funding: This work has been supported by the French National Agency for Research (ANR) under grant ANR-16-CE25-0001- ARBURST and by Orange Labs under External Research Contract F05151.

Acknowledgments: The authors thank Prof H. Vincent Poor for his insights in the development of the initial model. They also thank all members of the ANR project Arburst, especially Laurent Clavier, who contributed to the development of this paper throught insightful comments. The authors also thank Dr Samir Perlaza for his suggestions and proof reading.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AWGN	Additive White Gaussian Noise
BC	Broadcast Channel
BS	Base Station
ccdf	complementary cumulative distribution function
EE	Energy Efficiency
FBL	Finite block-length
FTT	Finite Time Transmission (constraint)
GSCBC	Gaussian Spatial Continuum Broadcast Channel
IoT	Internet of Things
MAC	Multiple Access Channel
NOMA	Non Orthogonal Multiple Access
OMA	Orthogonal Multiple Access
pdf	probability density function
PPP	Poisson Point Process
SC	Superposition Coding
SCBC	Spatial Continuum Broadcast Channel
SCMAC	Spatial Continuum Multiple Access Channel
SE	Spectral Efficiency
SIC	Successive Interference Cancellation
SINR	Signal To Interference Plus Noise Ratio
SNR	Signal To Noise Ratio
ULLC	Ultra Low Latency Communications

Appendix A. Proof of Lemma 1

Given (19), the power needed at distance r is $P(r) = (e^{a\eta_s} - 1)\frac{\sigma^2 r^\alpha}{g_0}$. In the case of spatial continuum, the equi-repartition of the power among the users is the total power integrated over the surface divided over the area of the cell. In a circular cell, the power distributed over a ring located at

distance r from the BS and whose the thickness is dr is $P(r) \cdot 2\pi r dr$. The power used in OMA is hence the total power spread over the cell divided by the total area. For a circular cell of radius R_c, we have

$$P_{\text{OMA}} = \frac{\int_{\Omega} P(r) 2\pi r dr}{\pi R_c^2} \tag{A1}$$

After straightforward computation, we end up with

$$P_{\text{OMA}} = \frac{2\sigma^2 R_c^{\alpha}}{g_0 \left(\alpha + 2\right)} \left(e^{a\eta_s} - 1\right) \tag{A2}$$

EE is simply SE normalized by P_{OMA}/ν_c, and the proof is complete.

1. Shannon, C.E. A Mathematical Theory of Communication. *Bell Sys. Tech. J.* **1948**, *27*, 379–423.
2. Vaezi, M.; Poor, H.V. NOMA: An information-theoretic perspective. In *Multiple Access Techniques for 5G Wireless Networks and Beyond*; Springer, 2019; pp. 167–193.
3. Wu, Y.; Gao, X.; Zhou, S.; Yang, W.; Polyanskiy, Y.; Caire, G. Massive Access for Future Wireless Communication Systems. *IEEE Wireless Communications* **2020**, *27*, 148–156. doi:10.1109/MWC.001.1900494.
4. Verdú, S. Spectral efficiency in the wideband regime. *IEEE Transactions on Information Theory* **2002**, *48*, 1319–1343.
5. Feinstein, A. A new basic theorem of information theory. *IRE Trans. Inform. Theory* **1954**, *4*, 2–22.
6. Shannon, C.E. Certain results in coding theory for noisy channels. *Information and Control* **1957**, *1*, 6–25.
7. Gallager, R.G. *Information Theory and Reliable Communication*; Wiley, 1968.
8. Polyanskiy, Y.; Poor, H.; Verdú, S. Channel Coding Rate in the Finite Blocklength Regime. *IEEE Transactions on Information Theory* **2010**, *56*, 2307–2359.
9. Yang, W.; Durisi, G.; Koch, T.; Polyanskiy, Y. Quasi-Static Multiple-Antenna Fading Channels at Finite Blocklength. *IEEE Transactions on Information Theory* **2014**, *60*, 4232–4265.
10. Hoydis, J.; Couillet, R.; Piantanida, P. The Second-Order Coding Rate of the MIMO Quasi-Static Rayleigh Fading Channel. *IEEE Transactions on Information Theory* **2015**, *61*, 6591–6622.
11. Durisi, G.; Koch, T.; Östman, J.; Polyanskiy, Y.; Yang, W. Short-Packet Communications Over Multiple-Antenna Rayleigh-Fading Channels. *IEEE Transactions on Communications* **2016**, *64*, 618–629.
12. Mary, P.; Gorce, J.M.; Unsal, A.; Poor, H.V. Finite blocklength information theory: What is the practical impact on wireless communications? 2016 IEEE Globecom Workshops (GC Wkshps). IEEE, 2016, pp. 1–6.
13. Jindal, N.; Vishwanath, S.; Goldsmith, A. On the duality of Gaussian multiple-access and broadcast channels. *IEEE Trans. Inf. Theory* **2004**, *50*, 768–783.
14. Verdu, S. Non-Asymptotic Achievability Bounds in Multi-User Information Theory. Proc. 50th Allerton Conference on Communication, Control and Computing, 2012.
15. Huang, Y.W.; Moulin, P. Finite Blocklength Coding for Multiple Access Channels. IEEE International Symposium on Information Theory (ISIT), 2012.
16. MolavianJazi, E.; Laneman, J. A Second-Order Achievable Rate Region for Gaussian Multi-Access Channels via a Central Limit Theorem for Functions. *IEEE Transactions on Information Theory* **2015**, *61*, 6719–6733.
17. Tan, V.Y.F.; Kosut, O. On the Dispersion of Three Network Information Theory Problems. *IEEE Transactions on Information Theory* **2014**, *60*, 881–903.
18. Unsal, A.; Gorce, J.M. The Dispersion of Superposition Coding for Gaussian Broadcast Channels. IEEE Information Theory Workshop 2017; , 2017.
19. Shahi, S.; Tuninetti, D.; Devroye, N. On the capacity of strong asynchronous multiple access channels with a large number of users. 2016 IEEE International Symposium on Information Theory (ISIT), 2016, pp. 1486–1490.
20. Kowshik, S.S.; Polyanskiy, Y. Quasi-static fading MAC with many users and finite payload. 2019 IEEE International Symposium on Information Theory (ISIT), 2019, pp. 440–444.
21. Chen, X.; Chen, T.Y.; Yu, W.; Guo, D. Capacity of Gaussian many-access channels. *IEEE Transactions on Information Theory* **2017**, *63*, 3516–3539.

22. Cao, W.; Dytso, A.; Shkel, Y.; Feng, G.; Poor, H.V. Sum-Capacity of the MIMO Many-Access Gaussian Noise Channel. *IEEE Transactions on Communications* **2019**, *67*, 5419–5433.
23. Shirvanimoghaddam, M.; Condoluci, M.; Dohler, M.; Johnson, S.J. On the Fundamental Limits of Random Non-Orthogonal Multiple Access in Cellular Massive IoT. *IEEE Journal on Selected Areas in Communications* **2017**, *35*, 2238–2252. doi:10.1109/JSAC.2017.2724442.
24. Maatouk, A.; Assaad, M.; Ephremides, A. Minimizing the age of information: NOMA or OMA ? IEEE INFOCOM 2019-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS). IEEE, 2019, pp. 102–108.
25. Gorce, J.M.; Poor, H.V.; Kelif, J.M. Spatial continuum model: Toward the fundamental limits of dense wireless networks. IEEE Global Communications Conference (GLOBECOM), 2016.
26. Gorce, J.M.; Fadlallah, Y.; Kelif, J.M.; Poor, H.V.; Gati, A. Fundamental limits of a dense IoT cell in the uplink. 15th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), 2017.
27. Gorce, J.M.; Mary, P.; Kélif, J.M. Towards Fundamental Limits of Bursty Multi-user Communications in Wireless Network. Proc. of the 10th workshop on information theory methods in science and engineering, Paris. University of Helsinki, 2017.
28. Gorce, J.M.; Tsilimantos, D.; Ferrand, P.; Poor, H.V. Energy-Capacity Trade-off Bounds in a downlink typical cell. Proc. IEEE Int. Symp. on Personal, Indoor and Mobile Radio Commun.(PIMRC), 2014.
29. Kelif, J.M.; Gorce, J.M.; Gati, A. Performance and Energy in Green Superposition Coding Wireless Networks: An Analytical Model. Proc. IEEE Globecom, 2017.
30. Chetot, L.; Gorce, J.M.; Kélif, J.M. Fundamental Limits in Cellular Networks with Point Process Partial Area Statistics. 17th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), 2019.
31. Yu, W. On the fundamental limits of massive connectivity. 2017 Information Theory and Applications Workshop (ITA), 2017, pp. 1–6. doi:10.1109/ITA.2017.8023482.
32. Cover, T.M.; Thomas, J.A. *Elements of information theory*; John Wiley & Sons, 2012.
33. Gradshteyn, I.S.; Ryzhik, I. *Table of integrals, series, and products. Translated from the Russian.*; Ac. Press, San Diego, CA, 2000.
34. Anade, D.; Gorce, J.M.; Mary, P.; Perlaza, S.M. An Upper Bound on the Error Induced by Saddlepoint Approximations—Applications to Information Theory. *Entropy* **2020**, *22*, 690.
35. Scarlett, J. On the dispersion of dirty paper coding. 2014 IEEE International Symposium on Information Theory. IEEE, 2014, pp. 2282–2286.