Low-latency Ultra-Reliable 5G Communications: Finite-blocklength bounds and coding schemes

Low-latency Ultra-Reliable 5G Communications: Finite-Blocklength Bounds and Coding Schemes

Johan Östman[1], Giuseppe Durisi[1], Erik G. Ström[1], Jingya Li[2], Henrik Sahlin[2], and Gianluigi Liva[3]
[1]Chalmers University of Technology, Gothenburg, Sweden; [2]Ericsson Research, Gothenburg, Sweden;
[3]Deutsches Zentrum für Luft- und Raumfahrt (DLR), Wessling, Germany

***Abstract*—Future autonomous systems require wireless connectivity able to support extremely stringent requirements on both latency and reliability. In this paper, we leverage recent developments in the field of finite-blocklength information theory to illustrate how to optimally design wireless systems in the presence of such stringent constraints. Focusing on a multi-antenna Rayleigh block-fading channel, we obtain bounds on the maximum number of bits that can be transmitted within given bandwidth, latency, and reliability constraints, using an orthogonal frequency-division multiplexing system similar to LTE. These bounds unveil the fundamental interplay between latency, bandwidth, rate, and reliability. Furthermore, they suggest how to optimally use the available spatial and frequency diversity. Finally, we use our bounds to benchmark the performance of an actual coding scheme involving the transmission of short packets.**

I. Introduction

The next generation of wireless cellular systems (5G) is expected to be a key enabler of future autonomous systems, be them connected vehicles, smart meters, or automated factories [1], [2]. The characteristics of the wireless data traffic typically generated within these autonomous systems is, however, drastically different from the one encountered in traditional broadband wireless applications: short data packets (on the order of hundreds of bits) that need to be delivered with stringent requirements in terms of latency and reliability.

For example, machine-type communication (MTC) for factory automation may involve the transmission of packets containing 100 information bits within 100 μs and with packet error probability not exceeding 10^{-9} [3], [4]. In traffic safety applications, one may need the packet error probability not to exceed 10^{-5} [5]. These requirements are much more stringent than the ones that current wireless cellular systems, i.e., *long term evolution* (LTE), need to handle. Standardization activities are currently ongoing within the *3rd generation partnership project* (3GPP), with the aim of evolving LTE and achieving these new requirements.

One way to reduce latency is to assign to each user a resource block (RB) consisting of a smaller number of orthogonal frequency-division multiplexing (OFDM) symbols than currently done in LTE.[1] This yields a shorter transmission time interval (TTI). The impact of a reduced TTI on the performance of LTE has been recently analyzed in [6], [7].

In order to increase reliability, one can use the available transmit and receive antennas to provide spatial diversity rather than spatial multiplexing. This has been investigated in [3], [4] in a factory-automation scenario, under the assumption that perfect channel state information (CSI) is available at the receiver.

The problem of optimally designing a communication system operating under a stringent latency constraint can be addressed in a fundamental fashion using the finite-blocklength information theoretic bounds recently developed by Polyanskiy *et al.* [8]. Using these tools, Durisi *et al.* [9] developed bounds on the maximum coding rate over multi-antenna Rayleigh block-fading channels. Since these bounds do not assume the *a priori* availability of perfect CSI, they unveil the fundamental tradeoff between exploiting spatial and time-frequency diversity (to obtain high reliability) on the one hand, and reducing channel-estimation overhead on the other hand. The bounds in [9], however, require Monte-Carlo simulations and are difficult to compute for packet error probabilities below 10^{-6}. An alternative approach to obtaining achievability bounds on the maximum coding rate is through random-coding error exponent analyses [10]. The random coding error exponent of Rayleigh-fading channels for the case when no CSI is available at the receiver has been obtained in [11] for the single-input single-output case. However, no error exponent results are available for the no-CSI multiple-antenna case.

Contribution: In this paper, we analyze the problem of designing an OFDM based system (such as LTE) able to satisfy a given set of requirements on reliability, latency, and bandwidth occupancy. The specific contributions are as follows. We use the information theoretic bounds recently developed in [9] for the multiple-antenna Rayleigh block-fading channel to analyze the tradeoff between latency, bandwidth, and rate for the case when each transmit packet comprises a certain number of RBs that are assumed to be orthogonal in time and frequency, and subject to independent fading. Our analysis applies to both the uplink (UL), where we assume a fixed average power per use of the channel in time, and to the downlink (DL), where we assume instead a power spectral density (PSD) constraint. The analysis is performed for a target packet error probability of 10^{-5}. To understand how to optimally use spatial and frequency diversity when the requirement on packet error probability is 10^{-9} or lower (ultra-reliable communications), we extend the error-exponent analysis in [11] to the case of multiple-antenna systems and provide an upper bound on the error probability for the case when the input distribution is the so called *unitary space-time modulation* (USTM) [12]. Finally, we use our bounds

[1]In LTE release 13, an RB comprises 12 adjacent subcarriers over 7 consecutive OFDM symbol durations; in this paper, however, we allow an RB to span an arbitrary number of adjacent subcarriers and consecutive OFDM symbols.

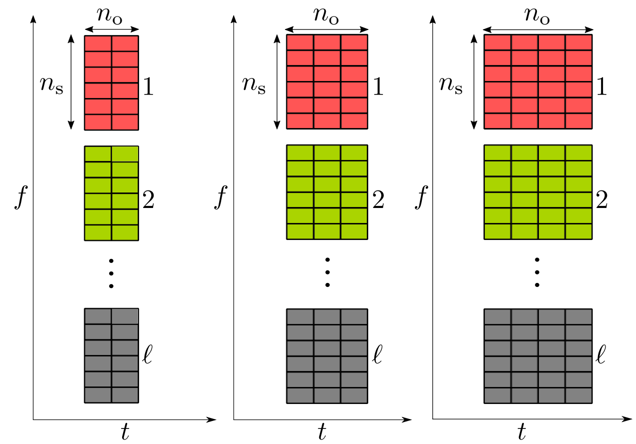

Fig. 1. An example of three different UE resource allocations. Here, $n_\mathrm{s} = 6$, $n_\mathrm{o} = \{2, 3, 4\}$. The fading process is assumed constant over an RB and the RBs are assumed to fade independently (RB spacing larger than the channel coherence bandwidth).

to benchmark the performance of a coding scheme based on pilot transmission and convolutional encoding of the information bits.

Notation: Uppercase letters such as X denote scalar random variables and their realizations are written in lowercase, e.g., x. We use two different fonts to write deterministic matrices (e.g., $\mathbf{X}$) and random matrices (e.g., $\mathbb{X}$). The superscript ^{H} denotes Hermitian transposition and $\mathrm{tr}(\cdot)$ and $\det(\cdot)$ denote the trace and the determinant of a given matrix, respectively. The identity matrix of size $a \times a$ is written as $\mathbf{I}_a$. We denote by $\mathcal{V}(\cdot)$ the Vandermonde determinant [13, p. 22]. The distribution of a circularly symmetric complex Gaussian random variable with variance σ^2 is denoted by $\mathcal{CN}(0, \sigma^2)$. Finally, $\log(\cdot)$ indicates the natural logarithm, $[a]^+$ stands for $\max\{0, a\}$, $\lfloor \cdot \rfloor$ is the floor operator, $\Gamma(\cdot)$ denotes the Gamma function, and $\mathbb{E}[\cdot]$ denotes the expectation operator.

II. System Model

We consider a wireless multiple-antenna communication system employing OFDM, similar to what is used in LTE [14]. As shown in Fig. 1, a UE is assigned ℓ RBs that are orthogonal in frequency, and constitute a packet.[2] An RB consists of n_o OFDM symbols, each one spanning n_s consecutive subcarriers. Hence, an RB contains a total of $n_\mathrm{c} = n_\mathrm{o} n_\mathrm{s}$ time-frequency slots, also referred to as resource elements in LTE. Note that n_o is related to the packet duration (in LTE, this quantity is referred to as TTI), whereas $n_\mathrm{s}\ell$ is related to the bandwidth assigned to a given UE. In LTE, the duration of an OFDM symbol is approximately 71.4 μs and the subcarrier spacing is 15 kHz. Hence, an RB consisting of $n_\mathrm{o} = 7$ OFDM symbols and $n_\mathrm{s} = 12$ subcarriers occupies 180 kHz and lasts for 0.5 ms. Obviously, decreasing the number n_o of OFDM symbols within each RB results in shorter delays. This is currently under investigation within 3GPP [6].

We assume the channel fading to stay constant within each RB and to change independently from RB to RB (block-fading model [12]). This assumption is reasonable for propagation environments characterized by low delay and Doppler spreads. One such example is the so called LTE pedestrian model, where the coherence bandwidth and the delay spread are approximately 23 MHz and 200 ms, respectively [15]. The number of frequency diversity branches, i.e., the number of independent fading realizations within a given packet, is equal to the number of resource blocks ℓ. We shall focus on Rayleigh fading. We shall also assume that the channels between each transmit-receive antenna pair fade independently (no spatial correlation).

[2]In LTE UL, the product $n_\mathrm{s}\ell$ has to be a multiple of 2, 3 and 5 due to implementation constraints. This will not be taken into account in this paper.

The channel input-output relation within the kth RB, for the case when the number of transmit antennas is n_t and the number of receive antennas is n_r, can be expressed as:

$$\mathbb{Y}_k = \mathbf{X}_k \mathbb{H}_k + \mathbb{W}_k, \quad k = 1, \dots, \ell. \tag{1}$$

Here, $\mathbf{X}_k \in \mathbb{C}^{n_\mathrm{o} n_\mathrm{s} \times n_\mathrm{t}}$ and $\mathbb{Y}_k \in \mathbb{C}^{n_\mathrm{o} n_\mathrm{s} \times n_\mathrm{r}}$ are the transmitted and received matrices, respectively; $\mathbb{H}_k \in \mathbb{C}^{n_\mathrm{t} \times n_\mathrm{r}}$ is the fading matrix, whose entries are identical and independently distributed (i.i.d.) $\mathcal{CN}(0, 1)$ random variables. Finally, $\mathbb{W}_k \in \mathbb{C}^{n_\mathrm{o} n_\mathrm{s} \times n_\mathrm{r}}$, which denotes the thermal noise at the receiver, has independent $\mathcal{CN}(0, 1)$-distributed entries. The processes $\{\mathbb{H}_k\}$ and $\{\mathbb{W}_k\}$ are i.i.d. across k and are mutually independent.

Throughout the paper, we shall assume that the realizations of the random fading matrices $\{\mathbb{H}_k\}_{k=1}^{\ell}$ are unknown to the transmitter and the receiver. As discussed in, e.g., [9], [16], and [17], this allows us to take into account the potential rate loss caused by the transmission of training sequences for channel estimation at the receiver.

Next, we define a channel code for the channel (1) using standard information-theoretic terminology (see, e.g., [8], [9]).

Definition 1: An $(\ell, n_\mathrm{s}, n_\mathrm{o}, M, \epsilon, \rho)$–code for the channel (1) consists of

- An encoder $f : \{1, \dots, M\} \to \mathbb{C}^{n_\mathrm{o} n_\mathrm{s} \times n_\mathrm{t} \ell}$ that maps a message $J \in \{1, \dots, M\}$ to a codeword $\mathbf{C}(J) \in \{\mathbf{C}_1, \dots, \mathbf{C}_M\}$. Each codeword can be expressed as a concatenation of ℓ subcodewords, each spanning an RB. Specifically, $\mathbf{C}_m = [\mathbf{C}_{m,1}, \dots, \mathbf{C}_{m,\ell}]$, $m \in \{1, \dots, M\}$, where $\mathbf{C}_{m,k} \in \mathbb{C}^{n_\mathrm{o} n_\mathrm{s} \times n_\mathrm{t}}$ for $k = 1, \dots, \ell$. Each subcodeword satisfies the power constraint
$$\mathrm{tr}\big(\mathbf{C}_{m,k}^H \mathbf{C}_{m,k}\big) = \rho. \tag{2}$$
- A decoder $g : \mathbb{C}^{n_\mathrm{o} n_\mathrm{s} \times n_\mathrm{r} \ell} \to \{1, \dots, M\}$ that satisfies the maximum error probability constraint
$$\max_{1 \le j \le M} \Pr\{g(\mathbb{Y}^\ell) \ne J \,|\, J = j\} \le \epsilon \tag{3}$$
where $\mathbb{Y}^\ell = [\mathbb{Y}_1, \dots, \mathbb{Y}_\ell]$ is the channel output induced by codeword $\mathbf{X}^\ell = [\mathbf{X}_1, \dots, \mathbf{X}_\ell] = f(j)$ through (1).

For the UL, we shall set ρ in (2) as follows:

$$\rho = n_\mathrm{o} \rho_\mathrm{u} / \ell. \tag{4}$$

Here, ρ_u can be thought as the average SNR per use of the channel in time (recall that the noise is assumed to have unit variance). In the DL, we shall instead assume a constraint on the PSD, i.e., on the average SNR per time-frequency slot. Specifically,

$$\rho = n_\mathrm{o} n_\mathrm{s} \rho_\mathrm{d}. \tag{5}$$

The subcodeword power constraints (4) and (5) imply the per-codeword power constraints $\mathrm{tr}\left(\mathbf{C}_m^H \mathbf{C}_m\right) = n_\mathrm{o}\rho_\mathrm{u}$ and $\mathrm{tr}\left(\mathbf{C}_m^H \mathbf{C}_m\right) = n_\mathrm{o} n_\mathrm{s} \ell \rho_\mathrm{d}, m = 1, \dots, M$, for the UL and the DL, respectively. Constraint (4) is motivated by the limited battery power at the UE, whereas constraint (5) captures that cellular base-stations need to fulfill spectral transmission masks.

The *maximum coding rate* R^* denotes the largest number of bits per time-frequency slot that can be transmitted with probability of error no larger than ϵ, for given ρ, n_s, ℓ and n_o:

$$R^* \triangleq \sup\left\{\frac{\log_2(M)}{n_\mathrm{s} n_\mathrm{o} \ell} : \exists\,(\ell, n_\mathrm{s}, n_\mathrm{o}, M, \epsilon, \rho)\text{–code}\right\}. \quad (6)$$

For a given subcarrier spacing and a given OFDM symbol duration, R^* is related to the largest number of bits $\lfloor n_\mathrm{o} n_\mathrm{s} \ell R^* \rfloor$ that can be transmitted with reliability $(1-\epsilon)$ through the channel (1) for given latency and bandwidth constraints.

III. Finite-Blocklength Bounds

Finite-blocklength bounds for the multiple-antenna Rayleigh block-fading channel were recently proposed in [9]. Here, we will review these bound and adapt them to our setting (differently from [9], we allow coding over frequency, which requires a different power normalization). The following definition will turn out useful.

Definition 2: Assume that $n_\mathrm{c} = n_\mathrm{o} n_\mathrm{s}$ is larger than the total number of antennas, $n_\mathrm{t} + n_\mathrm{r}$. Let $\mathbf{\Sigma}_k$ be an $n_\mathrm{c} \times n_\mathrm{c}$ diagonal matrix with positive diagonal entries. Let ξ be a positive real constant, $q = \min\{n_\mathrm{t}, n_\mathrm{r}\}$ and $p = \max\{n_\mathrm{t}, n_\mathrm{r}\}$. For $k = 1, \dots, \ell$ we define the random variable

$$S_k(\mathbf{\Sigma}_k, \xi) = c(\mathbf{\Sigma}_k) - \mathrm{tr}\left(\mathbb{Z}_k^H \mathbb{Z}_k\right) - \log(\psi(\mathbb{\Lambda}, \xi)) \quad (7)$$

where $\{\mathbb{Z}_k\}_{k=1}^{\ell}$ are independent complex Gaussian $n_\mathrm{o} n_\mathrm{s} \times n_\mathrm{r}$ matrices with i.i.d. $\mathcal{CN}(0,1)$ entries and $\mathbb{\Lambda} = \mathrm{diag}(\Lambda_1, \dots, \Lambda_{n_\mathrm{r}})$ is a diagonal matrix whose diagonal entries are the ordered eigenvalues of $\mathbb{Z}_k^H \mathbf{\Sigma}_k \mathbb{Z}_k$. The function, $c(\mathbf{\Sigma}_k)$ is given as follows:

$$\begin{aligned} c(\mathbf{\Sigma}_k) &= n_\mathrm{t}(n_\mathrm{c} - n_\mathrm{t}) \log\left(\frac{\rho}{n_\mathrm{t}}\right) - n_\mathrm{r} \log(\det(\mathbf{\Sigma}_k)) \\ &\quad - n_\mathrm{t}(n_\mathrm{c} - n_\mathrm{t} - n_\mathrm{r}) \log\left(1 + \frac{\rho}{n_\mathrm{t}}\right) \\ &\quad + \sum_{u=1}^{n_\mathrm{t}} \log(\Gamma(u)) - \sum_{u=n_\mathrm{c}-q+1}^{n_\mathrm{c}} \log(\Gamma(u)). \end{aligned} \quad (8)$$

Furthermore,

$$\psi(\mathbb{\Lambda}, \xi) = \frac{\det(\mathbf{M}(\mathbb{\Lambda}, \xi))}{\mathcal{V}(\mathbb{\Lambda})} \prod_{i=1}^{n_\mathrm{r}} \frac{\exp(-\Lambda_i/(1+\rho/n_\mathrm{t}))}{\Lambda_i^{n_\mathrm{c}-n_\mathrm{r}}} \quad (9)$$

where

$$[\mathbf{M}(\mathbb{\Lambda}, \xi)]_{i,j} = \begin{cases} \Lambda_i^{n_\mathrm{t}-j} \tilde{\gamma}\left([n_\mathrm{c} + j - p - n_\mathrm{t}]^+, \Lambda_i \xi\right), \\ \quad 1 \le i \le n_\mathrm{r}, \quad 1 \le j \le n_\mathrm{t}; \\ \exp(-\Lambda_i \xi) \left[\frac{\partial^{n_\mathrm{t}-j}}{\partial \delta^{n_\mathrm{t}-j}} \delta^{n_\mathrm{c}-i}\Big|_{\delta=\xi}\right], \\ \quad n_\mathrm{r} < i \le p, \quad 1 < j \le n_\mathrm{t}; \\ \Lambda_i^{n_\mathrm{c}-j} \exp(-\Lambda_i \xi), \\ \quad 1 \le i \le n_\mathrm{r}, \quad n_\mathrm{t} < j \le p \end{cases} \quad (10)$$

with

$$\tilde{\gamma}(n, x) \triangleq \frac{1}{\Gamma(n)} \int_0^x t^{n-1} \exp(-t)\, dt \quad (11)$$

denoting the regularized incomplete Gamma function.

With the help of Definition 2, we shall provide in the next two theorems an achievability (lower) and a converse (upper) bound on the maximum coding rate R^* defined in (6).

Theorem 1: The max. coding rate R^* is lower-bounded as

$$R^* \ge \max\left\{\frac{\log_2(M)}{n_\mathrm{s} n_\mathrm{o} \ell} : \epsilon_\mathrm{ub}(M) \le \epsilon\right\} \quad (12)$$

where

$$\epsilon_\mathrm{ub}(M) = \mathbb{E}\left[\exp\left(-\left[\sum_{k=1}^{\ell} S_k(\mathbf{\Sigma}_k, \xi) - \log(M-1)\right]^+\right)\right]. \quad (13)$$

Here, $S_k(\mathbf{\Sigma}_k, \xi)$ is defined in (7), $\xi = \rho/(n_\mathrm{t} + \rho)$, and $\mathbf{\Sigma}_k = \mathrm{diag}(\underbrace{\rho/n_\mathrm{t}+1, \dots, \rho/n_\mathrm{t}+1}_{n_\mathrm{t}}, \underbrace{1, \dots, 1}_{n_\mathrm{c}-n_\mathrm{t}})$.

Proof: The bound is obtained by applying the dependence testing bound [8, Thm. 22] to the channel (1) with input distribution chosen as USTM. For details, see [9, Thm. 1]. ■

Theorem 2: The max. coding rate R^* is upper-bounded as

$$\begin{aligned} R^* \le \sup_{\{\mathbf{\Sigma}_k\}_{k=1}^{\ell}} \inf_{\gamma>0} \frac{1}{n_\mathrm{s} n_\mathrm{o} \ell \log(2)} \\ \times \left\{\gamma - \log\left(\left[\Pr\left\{\sum_{k=1}^{\ell} S_k\left(\tilde{\mathbf{\Sigma}}_k, \xi\right) \le \gamma\right\} - \epsilon\right]^+\right)\right\}. \end{aligned} \quad (14)$$

Here, $S_k(\tilde{\mathbf{\Sigma}}_k, \xi)$ is defined in (7), $\xi = \rho/(n_\mathrm{t} + \rho)$ and the matrices $\{\tilde{\mathbf{\Sigma}}_k\}_{k=1}^{\ell}$ are given as follows

$$\tilde{\mathbf{\Sigma}}_k = \begin{bmatrix} \mathbf{\Sigma}_k + \mathbf{I}_{n_\mathrm{t}} & 0 \\ 0 & \mathbf{I}_{n_\mathrm{c}-n_\mathrm{t}} \end{bmatrix} \quad (15)$$

with $\{\mathbf{\Sigma}_k\}_{k=1}^{\ell}$ being $n_\mathrm{t} \times n_\mathrm{t}$ diagonal matrices with nonnegative elements satisfying the power constraint $\mathrm{tr}(\mathbf{\Sigma}_k) = \rho$.

Proof: The proof relies on the metaconverse theorem [8, Thm. 28]. The auxiliary distribution is chosen as the output distribution induced by an USTM input through the channel (1). For details, see [9, Thm. 2 and Remark 2]. ■

In the next section, the bounds in Theorem 1 and Theorem 2 will be used to characterize R^* for given latency and bandwidth occupancy constraints. Our implementation of the numerical routines needed for the evaluation of these bounds (available as part of `spectre`–*short-packet communications toolbox* [18]) requires Monte-Carlo analysis, rendering these bounds difficult to compute for packet error probabilities below 10^{-5}. To address this problem, we shall complement these bounds with an achievability bound on R^* based on Gallager's random coding error exponent that can be easily computed for low error probabilities.

Theorem 3: Let $n_\mathrm{c} = n_\mathrm{o} n_\mathrm{s}$ be larger than the total number of antennas $n_\mathrm{t} + n_\mathrm{r}$. Fix a rate R and let $q = \min\{n_\mathrm{t}, n_\mathrm{r}\}$. Let also $\mathbb{Y} = \mathbb{X}\mathbb{H} + \mathbb{W}$ where $\mathbb{H}$ and $\mathbb{W}$ are defined as in (1) and

$\mathbb{X} = (\rho/n_{\rm t})\Phi$, where $\Phi \in \mathbb{C}^{n_{\rm o}n_{\rm s}\times n_{\rm t}}$ is unitary and isotropically distributed. Finally, let $\mathbb{\Lambda} = \mathrm{diag}(\Lambda_1, \dots, \Lambda_{n_{\rm r}})$ denote the ordered eigenvalues of $\mathbb{Y}^H\mathbb{Y}$. The average error probability $\bar{\epsilon}$ is upper-bounded by

$$\bar{\epsilon} \leq \min_{0\leq\mu\leq 1} \exp(-\ell(\mathcal{E}(\mu) - \mu R)) \tag{16}$$

where

$$\begin{aligned}\mathcal{E}(\mu) &= c(\mu)\\ &- \log \mathbb{E}_{\mathbb{\Lambda}}\left[\left(\frac{\prod_{i=1}^{n_{\rm r}} e^{\xi\Lambda_i}\Lambda_i^{n_{\rm r}-n_{\rm c}}}{\mathcal{V}(\mathbb{\Lambda})}\det(\mathbf{M}(\mathbb{\Lambda},\xi))\right)^{(1+\mu)}\right]\end{aligned} \tag{17}$$

with $\xi = \rho/((1+\rho)(1+\mu))$ and

$$c(\mu) = (1+\mu)\log\left(\frac{\left(1+\frac{\rho}{n_{\rm t}}\right)^{\frac{n_{\rm r}n_{\rm t}}{1+\mu}} \xi^{n_{\rm t}(n_{\rm c}-n_{\rm t})}\prod_{i=1}^{n_{\rm t}}\Gamma(i)}{\prod_{i=n_{\rm c}-q+1}^{n_{\rm c}}\Gamma(i)}\right). \tag{18}$$

The matrix $\mathbf{M}(\mathbb{\Lambda},\xi)$ in (17) is defined in (10). Furthermore, the probability distribution function of the ordered eigenvalues $(\Lambda_1, \dots, \Lambda_{n_{\rm r}})$ is given by

$$f_{\mathbb{\Lambda}}(\mathbf{\Lambda}) = \frac{\exp(-\sum_{i=1}^{n_{\rm r}}\lambda_i)\left(\prod_{i=1}^{n_{\rm r}}\lambda_i\right)\mathcal{V}(\mathbf{\Lambda})^2}{\prod_{i=1}^{n_{\rm r}}\Gamma(n_{\rm c}-i+1)\,\Gamma(n_{\rm r}-i+1)}. \tag{19}$$

Proof: This result follows essentially from [11] by choosing USTM as input distribution. The details are omitted due to space constraints. ■

Remark 1: The average error probability $\bar{\epsilon}$ in (16) can be converted into maximum error probability (see (3)) by following a standard procedure (see, e.g., [19, p. 204]).

Unfortunately, the expectation in Theorem 3 seems formidable to solve in closed form. However, for small $n_{\rm r}$, say $n_{\rm r} \leq 3$, it can be efficiently evaluated numerically.

IV. Numerical Results

In this section, we shall use the bounds (12), (14), and (16) to derive guidelines on the optimal design of the OFDM system described in Section II as a function of the latency, bandwidth, and reliability constraints.

The numerical evaluation of the upper bound (14) is challenging because it requires one to maximize over the diagonal matrices $\{\mathbf{\Sigma}_k\}_{k=1}^{\ell}$. Throughout this section we simplify the numerical evaluations by assuming $\mathbf{\Sigma}_k = (\rho/n_{\rm t})\mathbf{I}_{n_{\rm t}}$. The accuracy of this approximation was validated numerically in [9].

A. Dependency of R^* on ℓ and $n_{\rm o}$

In this section, we shall use the bounds in Theorem 1 and 2 to investigate how R^* depends on the number of resource blocks ℓ and the number of OFDM symbols $n_{\rm o}$. We shall consider both a 1×2 and a 2×2 multiple-input multiple-output (MIMO) system in the UL and both a 2×1 and a 2×2 MIMO system in the DL. The target packet error probability is 10^{-5}. Furthermore, we shall assume throughout this subsection that the number of subcarriers per RB, $n_{\rm s}$, is 12 and consider both the case $n_{\rm o} = 2$ and $n_{\rm o} = 4$ OFDM symbols. For an OFDM symbol duration of 71.4 μs (including cyclic prefix) as in LTE, these values of $n_{\rm o}$ yield a packet duration of 142.8 μs and 285.6 μs, respectively. We shall also assume a subcarrier spacing of 15 kHz (again as in LTE) so that we can relate the product $n_{\rm s}\ell$ to the bandwidth assigned to a given UE.

Our results for the UL are reported in Fig. 2 (1×2 MIMO) and Fig. 3 (2×2 MIMO). As expected, achievability and converse bounds are loose for small values of ℓ and become progressively tighter as ℓ (and, hence, the packet size $n_{\rm o}n_{\rm s}\ell$) increases. As expected, R^* is larger for the case $n_{\rm o} = 4$ because the packet size is larger, which allows for more resilience against the additive noise. The crossing between the converse curve for the case $n_{\rm o} = 2$ and the one for the case $n_{\rm o} = 4$ for values of ℓ smaller than 3 is merely a consequence of the looseness of our converse bound for very small values of ℓ (note that this crossing does not occur for the achievability bound). As far as the dependency of R^* on ℓ is concerned, we observe that our bounds are not monotonic in ℓ, but that there exists an optimal ℓ (roughly about $\ell = 5$ for the $n_{\rm o} = 4$ case) beyond which the maximum coding rate decreases. To the left of this optimal value, the main bottleneck is the limited time-frequency diversity, whereas to the right of this optimal value the main bottleneck is the low power per resource block available (the power scales inversely with ℓ, see (4)).

In Fig. 3, we consider the 2×2 MIMO case. We see that adding a second transmit antenna is beneficial for small values of ℓ. Indeed, for the case $n_{\rm o} = 4$, the achievable bound peaks at about 0.94 bits per time-frequency slots compared to 0.54 bits per time-frequency slots in the 1×2 case. Furthermore, this peak occurs at smaller values of ℓ (3 instead of 5), which implies that the additional spatial diversity provided by the second antenna reduces the need for frequency diversity. This results in bandwidth savings. We note also that, as ℓ increases, the rate gains resulting from the use of a second antenna diminish. This is accordance to the well-known result that in the low-SNR regime, using a single antenna is optimal when the channel is not known to the receiver [20].

The DL scenario is analyzed in Fig. 4 for the 2×1 and 2×2 cases. Since now the available power increases with ℓ because of the PSD constraint (5), all curves become monotonic in ℓ. As shown in the figures, our bounds allow one to estimate accurately the bandwidth and latency required to operate at a given rate. We see for example that for the 2×1 MIMO case, one can operate at a rate of approximately 1.4 bits per time-frequency slot using a packet of duration 285.6 μs ($n_{\rm o} = 4$) and a bandwidth of 1.26 MHz, or alternatively a packet of duration 142.8 μs ($n_{\rm o} = 2$) and a bandwidth of 1.44 MHz.

B. Optimal use of spatial and frequency diversity

To investigate how to optimally use spatial and frequency diversity, we analyze in this section a DL system with a variable number of transmit antennas $n_{\rm t}$ and a single receive antenna ($n_{\rm r} = 1$). We consider a scenario in which 130 information bits are transmitted over $n_{\rm o}n_{\rm s}\ell = 168$ time-frequency slots ($R \approx 0.77$). We shall assume $n_{\rm o} = 2$ and investigate how the error probability behaves as a function of $\rho_{\rm d}$ for different values of ℓ and $n_{\rm t}$. Note that since the total packet length is fixed to 168, larger values of ℓ imply smaller RBs. Specifically, each OFDM symbol is assumed to span $n_{\rm s} = 84/\ell$ subcarriers.

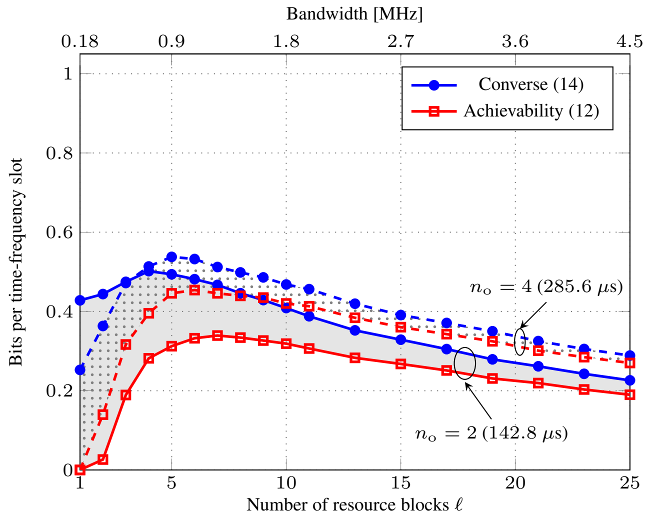

Fig. 2. Achievability bound (12) and converse bound (14) on the maximum coding rate in a UL 1×2 system for different number of resource blocks ℓ. Here, $\rho_u = 20$ dB, $\epsilon = 10^{-5}$, and $n_s = 12$.

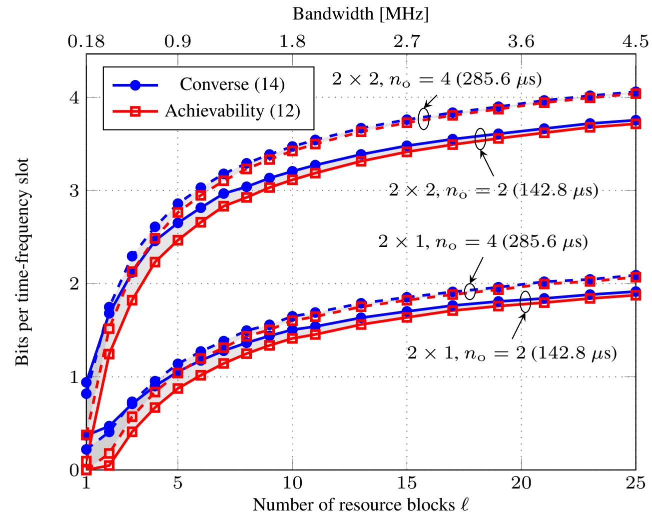

Fig. 4. Achievability bound (12) and converse bound (14) on the maximum coding rate in a DL 2×1 and a DL 2×2 system for different number of resource blocks ℓ. Here, $\rho_d = 10$ dB, $\epsilon = 10^{-5}$, and $n_s = 12$.

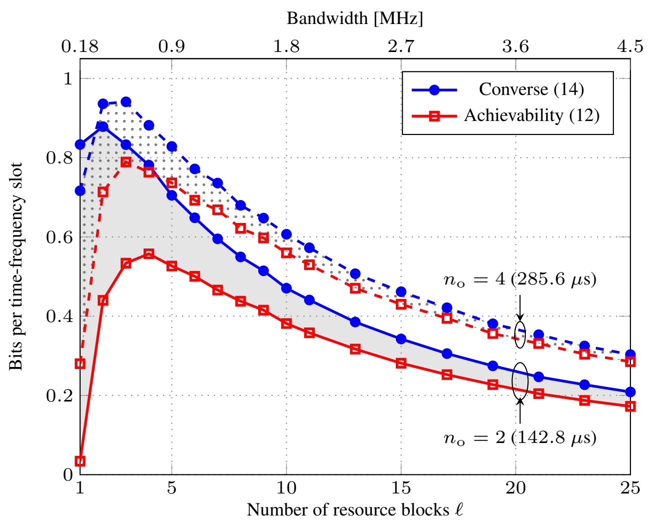

Fig. 3. Achievability bound (12) and converse bound (14) on the maximum coding rate in a UL 2×2 system for different number of resource blocks ℓ. Here, $\rho_u = 20$ dB, $\epsilon = 10^{-5}$, and $n_s = 12$.

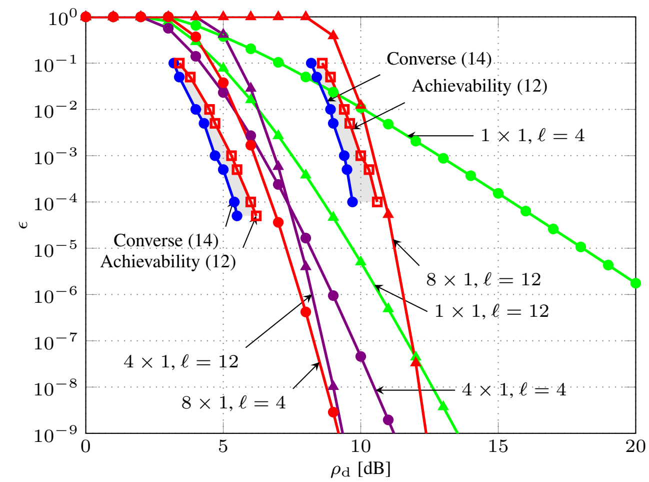

Fig. 5. Achievability bound (16) for a DL $n_t \times 1$ system for different n_t, ℓ, and ρ_d. It is assumed that $n_o = 2$, $n_s\ell = 84$, and $R = 0.77$ bits per time-frequency slot. The achievability bound (12) and converse bound (14) are included for comparison for the cases $n_t = 8$ and $\ell = \{4, 12\}$.

In Fig. 5, we plot the achievability bound (16) after converting it to maximum error probability [19, p. 204] for the case $n_t \in \{1, 2, 4\}$ and $\ell = \{4, 12\}$ (which yield $n_s = 21$ and $n_s = 7$, respectively). For the 8×1 case and both $\ell = 4$ and $\ell = 12$, we compare the achievability bound (16) to the achievability and converse bounds (12) and (14) up to the values of ϵ for which these two bounds can be computed. As expected, (16) is less accurate than (12) for moderate error probabilities. For example, when $n_t = 8$ and $\ell = 4$, the gap between these two achievability bounds is 0.26 dB at $\epsilon = 10^{-4}$. The gap turns out to be larger for smaller n_t values. For example, when $n_t = 1$, the gap between the two bounds at $\epsilon = 10^{-4}$ (not shown in the figure) is 3.8 dB.

We see from the figure that at $\epsilon = 10^{-9}$ the minimum value of SNR ρ_d predicted by our bound (16) is achieved by selecting $n_t = 8$ and $\ell = 4$, which yields 32 independent fading branches. A similar SNR value is needed when $n_t = 4$ and $\ell = 12$, yielding 48 fading branches. The figure also illustrates that further increasing the number of fading branches, as in the $n_t = 8$, $\ell = 12$ case, is not effective because one is limited by the channel estimation overhead. Reducing the number of diversity branches as in the $n_t = 1$, $\ell = 12$ case is also not effective, because of the lack of diversity (which is reflected by the more gentle slope of the curves).

C. *Practical coding schemes*

We finally benchmark the performance of an actual coding scheme against the bounds provided in Theorems 1 and 2. We consider a 1×2 MIMO system in UL and assume $n_o = 2$, $n_s = 12$, and $\ell = 8$, which results in a packet length of 192 time-frequency slots consisting of 8 RBs. We also assume that 92 information bits are transmitted, which results in a rate of $92/192 \approx 0.48$ bits per time-frequency slot. Within each RB, we reserve n_{p} time-frequency slots for pilot transmission. In the remaining $(n_{\mathrm{o}}n_{\mathrm{s}} - n_{\mathrm{p}})\ell$ slots we transmit coded bits mapped into QPSK symbols. As coding scheme, we consider a tail-biting $(368, 92)$ convolutional code with a memory-15 nonsystematic encoder, which is designed for the case $n_{\mathrm{p}} = 1$ (indeed, 368 coded bits yield 184 QPSK symbols, which together with the 8 pilot symbols, yield the desired blocklength of 192). For values of n_{p} larger than 1, the encoder output is punctured. Specifically, two coded bits are punctured for each additional pilot symbol.

At the receiver side, the pilot symbols are used to estimate the channel coefficients by means of a maximum-likelihood (ML) estimator. Thereafter, maximum ratio combining is performed and the bit-wise log-likelihood ratios are derived and given as input to the decoder. A sub-optimum list decoding algorithm based on ordered statistics has been used for the simulations. Specifically, ordered statistics decoding with test patterns of maximum weight equal to 3 has been adopted. This was shown to provide a negligible loss with respect to ML decoding for codes of length up to a few hundred bits [21].

The packet error probability-SNR tradeoff of this coding scheme is depicted in Fig. 6 for $n_{\mathrm{p}} = \{1, 2, 4, 6, 8\}$. As a benchmark, we also depict the performance predicted by the finite-blocklength bounds in Theorem 1 and 2. We note that, for the chosen coding scheme, the optimal number of pilots turns out to be $n_{\mathrm{p}} = 6$. For this value of n_{p}, the SNR gap between the coding scheme and the achievability bound (12) is about 2.68 dB at $\epsilon = 10^{-2}$. Furthermore, our numerical results illustrate that the performance of our coding scheme is extremely sensitive to the chosen number of pilot symbols.

V. Conclusion

We considered the problem of designing an OFDM-based system, similar to LTE, operating under stringent constraints on latency and reliability. Information-theoretic finite-blocklength bounds turned out to provide valuable insight on how to choose the system bandwidth as a function of the desired reliability and latency constraints, and how to exploit the available spatial and frequency diversity. We also used our bounds to benchmark the performance of an actual coding scheme, which relies on convolutional encoding and the transmission of pilot symbols.

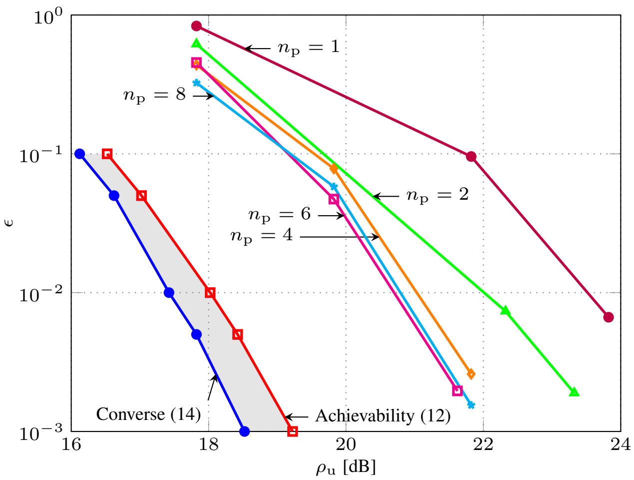

Fig. 6. Comparison of the achievability (12) and converse bound (14) with the performance of a coding scheme of rate $R \approx 0.48$ bits per time-frequency slot for $n_{\mathrm{p}} = \{1, 2, 4, 6, 8\}$. Here, $n_{\mathrm{t}} = 1$, $n_{\mathrm{r}} = 2$, $\ell = 8$, $n_{\mathrm{o}} = 2$ and $n_{\mathrm{s}} = 12$.

References

[1] A. Osseiran, F. Boccardi, V. Braun, K. Kusume, P. Marsch, M. Maternia, O. Queseth, M. Schellmann, H. Schotten, H. Taoka, H. Tullberg, M. A. Uusitalo, B. Timus, and M. Fallgren, "Scenarios for 5G mobile and wireless communications: the vision of the METIS project," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 26–35, May 2014.

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

[3] N. A. Johansson, Y.-P. E. Wang, E. Eriksson, and M. Hessler, "Radio access for ultra-reliable and low-latency 5G communications," in *Proc. IEEE Int. Conf. Commun. (ICC)*, London, U.K., Jun. 2015.

[4] O. N. C. Yilmaz, Y.-P. E. Wang, N. A. Johansson, N. Barhmi, S. A. Ashraf, and J. Sachs, "Analysis of ultra-reliable and low-latency 5G communication for a factory automation use case," in *Proc. IEEE Int. Conf. Commun. (ICC)*, London, U.K., Jun. 2015.

[5] "Scenarios, requirements and KPIs for 5G mobile and wireless system," METIS deliverable D1.1, document ICT-317669-METIS/D1.1, Apr. 2013, available: https://www.metis2020.com/documents/deliverables/.

[6] 3GPP, "Link evaluation for PUSCH for short TTI," 3GPP TSG RAN1#84-BIS, Tech. Rep. R1-163411, Apr. 2016. [Online]. Available: http://www.3gpp.org/DynaReport/TDocExMtg--R1-84b--31661.htm

[7] ——, "Study on latency reduction techniques for LTE," 3GPP TSG RAN#72, Tech. Rep. RP-161024, Jun. 2016. [Online]. Available: http://www.3gpp.org/DynaReport/TDocExMtg--RP-72--31638.htm

[8] 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. 2307–2359, May 2010.

[9] G. Durisi, T. Koch, J. Östman, Y. Polyanskiy, and W. Yang, "Short-packet communications over multiple-antenna Rayleigh-fading channels," *IEEE Trans. Commun.*, vol. 64, no. 2, pp. 618–629, Feb. 2016.

[10] R. G. Gallager, *Information Theory and Reliable Communication*. New York, NY, USA: John Wiley & Sons Inc., 1968.

[11] I. Abou-Faycal and B. M. Hochwald, "Coding requirements for multiple-antenna channels with unknown Rayleigh fading," Bell Labs Technical Memorandum, Tech. Rep., Mar. 1999.

[12] T. L. Marzetta and B. M. Hochwald, "Capacity of a mobile multiple-antenna communication link in Rayleigh flat fading," *IEEE Trans. Inf. Theory*, vol. 45, no. 1, pp. 139–157, Jan. 1999.

[13] R. Couillet and M. Debbah, *Random Matrix Methods for Wireless Communications*. Cambridge, UK: Cambridge Univ. Press, 2011.

[14] E. Dahlman, S. Parkvall, and J. Sköld, *4G LTE/LTE-Advanced for Mobile Broadband*, 1st ed. Waltham, U.S.: Academic Press, 2011.

[15] 3GPP, "LTE; evolved universal terrestrial radio access (E-UTRA); base station (BS) radio transmission and reception," 3GPP Specification series, Tech. Rep. TS 36.104, Nov. 2008. [Online]. Available: http://www.3gpp.org/dynareport/36-series.htm

[16] W. Yang, G. Durisi, and E. Riegler, "On the capacity of large-MIMO block-fading channels," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 117–132, Feb. 2013.

[17] A. Lapidoth, "On the asymptotic capacity of stationary Gaussian fading channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 2, pp. 437–446, Feb. 2005.

[18] A. Collins, G. Durisi, T. Erseghe, V. Kostina, J. Östman, Y. Polyanskiy, I. Tal, and W. Yang, *SPECTRE: short-packet communication toolbox*, v2.0, Sep. 2016. [Online]. Available: https://github.com/yp-mit/spectre

[19] T. M. Cover and J. A. Thomas, *Elements of information theory*, 2nd ed. New York, U.S.: Wiley, 2006.

[20] S. Verdú, "Spectral efficiency in the wideband regime," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1319–1343, Jun. 2002.

[21] M. P. C. Fossorier and S. Lin, "Soft-decision decoding of linear block codes based on ordered statistics," *IEEE Trans. Inf. Theory*, vol. 41, no. 5, pp. 1379–1396, Sep 1995.