Short-Packet Communications Over Multiple-Antenna Rayleigh-Fading Channels

Citation: Durisi, Giuseppe et al. "Short-Packet Communications Over Multiple-Antenna Rayleigh-Fading Channels." IEEE Transactions on Communications 64, 2 (February 2016): 618–629 © 2016 Institute of Electrical and Electronics Engineers (IEEE)

As Published: http://dx.doi.org/10.1109/tcomm.2015.2511087

Publisher: Institute of Electrical and Electronics Engineers (IEEE)

Persistent URL: http://hdl.handle.net/1721.1/111025

Version: Original manuscript: author's manuscript prior to formal peer review

Short-Packet Communications over Multiple-Antenna Rayleigh-Fading Channels

Giuseppe Durisi, *Senior Member, IEEE*, Tobias Koch, *Member, IEEE*, Johan Östman, Yury Polyanskiy, *Senior Member, IEEE*, Wei Yang, *Member, IEEE*

***Abstract*—Motivated by the current interest in ultra-reliable, low-latency, machine-type communication systems, we investigate the tradeoff between reliability, throughput, and latency in the transmission of information over multiple-antenna Rayleigh block-fading channels. Specifically, we obtain finite-blocklength, finite-SNR upper and lower bounds on the maximum coding rate achievable over such channels for a given constraint on the packet error probability. Numerical evidence suggests that our bounds delimit tightly the maximum coding rate already for short blocklengths (packets of about 100 symbols). Furthermore, our bounds reveal the existence of a tradeoff between the rate gain obtainable by spreading each codeword over all available time-frequency-spatial degrees of freedom, and the rate loss caused by the need of estimating the fading coefficients over these degrees of freedom. In particular, our bounds allow us to determine the optimal number of transmit antennas and the optimal number of time-frequency diversity branches that maximize the rate. Finally, we show that infinite-blocklength performance metrics such as the ergodic capacity and the outage capacity yield inaccurate throughput estimates.**

***Index Terms*—Ultra-reliable low-latency communications, mission-critical machine-type communications, multiple antennas, fading channels, transmit diversity, spatial multiplexing, finite-blocklength information theory.**

I. Introduction

Multi-antenna technology is a fundamental part of most modern wireless communication standards, due to its ability to provide tremendous gains in both spectral efficiency and reliability. The use of multiple antennas yields additional spatial degrees of freedom that can be used to lower the error probability for a given data rate, through the exploitation of *spatial diversity*, or increase the data rate for a given error probability, through the exploitation of *spatial multiplexing*. These two effects cannot be harvested concurrently and there exists a fundamental tradeoff between diversity and multiplexing. This tradeoff admits a particularly simple characterization in the high signal-to-noise ratio (SNR) regime [1].

Cellular systems offering mobile broadband services operate typically at maximum multiplexing [2] and do not make use of diversity-exploiting techniques such as space-time codes, whose purpose is to reduce the outage probability. Indeed, diversity-exploiting techniques are useless for low-mobility users, for which the fading coefficients can be learnt easily at the transmitter and outage events can be avoided altogether by rate adaptation. They are not advantageous for high-mobility users as well, because of the abundant time and frequency selectivity that is available, which is sufficient for modern cellular systems to operate at the target outage level.

These conclusions have been derived in [2] under the assumptions of long data packets (1000 channel uses or more) and moderately low packet-error rates (around 10^{-2}), which are relevant for current mobile broadband services.

In next-generation (5G) cellular systems, it is expected that enhanced mobile-broadband services (exploiting most likely the millimeter-wave part of the frequency spectrum and relying on advanced antenna solutions) will be complemented by new services centered on *machine-type communications* (MTC) [3]–[8]. An important emerging area among MTC systems is that of ultra-reliable, low-latency communications [9], [10], also known as *mission-critical* MTC [7]. This area targets MTC systems that require reliable real-time communications with stringent requirements on latency, reliability, and availability. Examples of mission-critical MTCs include smart grids for power distribution automation, industrial manufacturing and control, and intelligent transportation systems [7]. For example, in the case of industrial automation applications [9], [10], one is typically interested in transmitting short packets consisting of about 100 bits within $100\,\mu$s and with 10^{-9} packet error rate.

Motivated by mission-critical MTC systems, we investigate in this paper the fundamental tradeoff between throughput, reliability, and latency in short-packet wireless links. We also analyze how multiple antennas should be used in such links. Specifically, we address the following questions. Can the stringent reliability requirements of mission-critical MTC be met if the available transmit antennas are used to increase throughput (i.e., provide spatial multiplexing), or should these antennas be used to increase reliability (i.e., provide spatial diversity)? What is the cost of learning the fading coefficients, whose knowledge

This work was supported in part by the Swedish Research Council under grant 2012-4571, by the National Science Foundation CAREER award under grant agreement CCF-12-53205, by the European Community's Seventh Framework Programme FP7/2007-2013 under Grant 333680, by the Ministerio de Economía y Competitividad of Spain under Grants RYC-2014-16322, TEC2013-41718-R, and CSD2008-00010, and by the Comunidad de Madrid under Grant S2013/ICE-2845. The simulations were performed in part on resources at Chalmers Centre for Computational Science and Engineering (C3SE) provided by the Swedish National Infrastructure for Computing (SNIC).

The material of this paper was presented in part at the 2012 IEEE Information Theory Workshop, Lausanne, Switzerland, and in part at the 2014 IEEE International Symposium on Wireless Communication Systems, Barcelona, Spain.

G. Durisi and J. Östman are with the Department of Signals and Systems, Chalmers University of Technology, 41296, Gothenburg, Sweden (e-mail: {durisi,johanos}@chalmers.se).

T. Koch is with the Signal Theory and Communications Department, Universidad Carlos III de Madrid, 28911, Leganés, Spain and with the Gregorio Marañón Health Research Institute (e-mail: koch@tsc.uc3m.es).

Y. Polyanskiy is with the Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, 02139 USA (e-mail: yp@mit.edu).

W. Yang is with the Department of Electrical Engineering, Princeton University, NJ, 08544 USA (e-mail: weiy@princeton.edu)

is required to exploit the spatial degrees of freedom provided by multiple antennas, when the packet size is short? Does this cost overcome the benefits of using multiple antennas?

Contributions: The tension between reliability, throughput, and channel-estimation overhead in multiple-antenna communications have been investigated previously in the literature. However, as we shall review in Section III, most of the available results are asymptotic either in the packet length [11], [2], [12], or in the SNR [13], or in both [14]–[17]. Hence, their relevance in the context of mission-critical MTC is unclear.

In this paper, we address this issue by presenting a more refined *nonasymptotic* analysis of the tradeoff between reliability, throughput, latency, and channel-estimation overhead, which relies on the finite-blocklength bounds developed in [18]. Our main contributions are as follows:

- Focusing on the so-called *Rayleigh block-fading* model [19], [14], which is relevant for mission-critical MTC systems operating in a rich scattering environment [9], [10], we obtain nonasymptotic achievability and converse bounds on the maximum coding rate achievable for a given SNR, a given packet size, and a given packet reliability.
- We present numerical evidence that the newly derived achievability and converse bounds delimit tightly the maximum coding rate for packet lengths of interest for mission-critical MTC systems. Furthermore, our numerical examples show that the bounds allow one to identify accurately the throughput-maximizing number of transmit antennas as a function of the number of available time-frequency diversity branches. We also show that throughput estimates based on asymptotic performance metrics such as the ergodic capacity and the outage capacity are inaccurate, especially when the channel offers a significant amount of time-frequency diversity branches and the packet length is small.
- A comparison with nonasymptotic maximum coding rate bounds, obtained for specific space-time inner codes (such as the Alamouti scheme), allows us to identify when the available transmit antennas should be used to increase reliability, or throughput, or should be partly switched off to limit the channel-estimation overhead.

In previous works, researchers have drawn inspiration from the structure of the capacity achieving distribution of multiple-input multiple-output (MIMO) channels to design practical coded-modulation schemes (see e.g., [20]). In this paper, we go one step further and study how the choice of the input distribution affects the nonasymptotic achievability bounds and the corresponding converse bounds.

The results in this paper generalize to the multiple-antenna setting the analysis conducted in [21] for the single-input single-output case. A partial extension of the results in [21] to the MIMO case is provided in [22]. The analysis in [22], however, relies critically on the assumption that the codewords are orthogonal in space and that the transmit power is allocated uniformly both across antennas and across coherence intervals (see [22, Eq. (3)]). This assumption is dropped in the current paper. As we shall illustrate in Section VII, allocating the power uniformly across antennas is in fact suboptimal when the number of available time-frequency diversity branches is large. Bounds on the maximum coding rate for the case of *quasi-static* fading channels, i.e., channels for which the fading stays constant over the duration of each codeword are reported in [23]. Differently from [23], in this paper we allow each codeword to span multiple fading realizations in time and/or frequency.

Notation: Upper case letters such as X denote scalar random variables and their realizations are written in lower case, e.g., x. We use boldface upper case letters to denote random vectors, e.g., $\boldsymbol{X}$, and boldface lower case letters for their realizations, e.g., $\boldsymbol{x}$. Upper case letters of two special fonts are used to denote deterministic matrices (e.g., Y) and random matrices (e.g., $\mathbb{Y}$). The superscripts ^{H} and * stand for Hermitian transposition and complex conjugation, respectively, and we use $\mathrm{tr}\{\cdot\}$ and $\det\{\cdot\}$ to denote the trace and the determinant of a given matrix, respectively. The identity matrix of size $a \times a$ is written as I_a. The distribution of a zero-mean, circularly symmetric complex Gaussian random variable with variance σ^2 is denoted by $\mathcal{CN}(0, \sigma^2)$. For two functions $f(x)$ and $g(x)$, the notation $f(x) = \mathcal{O}(g(x))$, $x \to \infty$, means that $\limsup_{x\to\infty} |f(x)/g(x)| < \infty$, and $f(x) = o(g(x)), x \to \infty$, means that $\lim_{x\to\infty} |f(x)/g(x)| = 0$. Finally, $\ln(\cdot)$ indicates the natural logarithm, $[a]^+$ stands for $\max\{a, 0\}$, and $\Gamma(\cdot)$ denotes the Gamma function.

Following [24], we say that a scheme provides time, frequency, or spatial diversity if it allows the information symbols to pass through independently fading signal paths (diversity branches) in time, frequency, or space. We say that a scheme provides spatial multiplexing if it allows the transmission of multiple parallel data streams over the same channel. Throughout the paper, we shall rely on these broad notions of diversity and multiplexing. One exception is when we will review the diversity-multiplexing tradeoff (DMT) [13] in Section IV. To avoid any ambiguity, we shall refer to the quantities involved in the DMT, which are defined only in the high-SNR regime, as *diversity gain* and *multiplexing gain*.

II. System Model

We consider a Rayleigh block-fading channel with m_{t} transmit antennas and m_{r} receive antennas that stays constant for n_{c} channel uses. For a frequency-flat narrowband channel, n_{c} is the number of channel uses in time over which the channel stays constant (coherence time); for a frequency-selective channel and under the assumption that orthogonal frequency-division multiplexing (OFDM) is used, n_{c} is the number of subcarriers over which the channel stays constant (coherence bandwidth). More generally, n_{c} can be interpreted as the number of "time-frequency slots" over which the channel does not change.

Within the kth coherence interval, the channel input-output relation can be written as

$$\mathbb{Y}_k = \mathsf{X}_k \mathbb{H}_k + \mathbb{W}_k. \tag{1}$$

Here, $\mathsf{X}_k \in \mathbb{C}^{n_{\mathrm{c}} \times m_{\mathrm{t}}}$ and $\mathbb{Y}_k \in \mathbb{C}^{n_{\mathrm{c}} \times m_{\mathrm{r}}}$ are the transmitted and received matrices, respectively; the entries of the complex fading matrix $\mathbb{H}_k \in \mathbb{C}^{m_{\mathrm{t}} \times m_{\mathrm{r}}}$ are independent and identically distributed (i.i.d.) $\mathcal{CN}(0, 1)$; $\mathbb{W}_k \in \mathbb{C}^{n_{\mathrm{c}} \times m_{\mathrm{r}}}$ denotes the additive noise at the receiver and has i.i.d. $\mathcal{CN}(0, 1)$ entries. We assume $\{\mathbb{H}_k\}$ and $\{\mathbb{W}_k\}$ to take on independent realizations over successive coherence intervals. We further assume that $\mathbb{H}_k$ and $\mathbb{W}_k$ are independent and that their joint law does not depend on X_k.

Most throughput analyses available in the literature rely on the assumption that the receiver has perfect channel state information (CSI), i.e., that a "genie" informs the receiver about the realizations of the fading process $\{\mathbb{H}_k\}$. As discussed in [25], [16], [26], this assumption relies on the fact that CSI can be acquired by transmitting some known training symbols that are used by the receiver to learn $\{\mathbb{H}_k\}$. Unfortunately, throughput estimates based on the assumption of perfect CSI at the receiver are overly optimistic for two reasons: i) CSI will always be imperfect, no matter how long the training sequences are; ii) transmitting training sequences yields a rate loss (channel-estimation overhead), which—as we shall see—can be significant for short-packet transmission. Analyses relying on the perfect-CSI assumption simply ignore this overhead.

To obtain more realistic throughput estimates, in this paper we drop the assumption of perfect CSI at the receiver. Instead, we assume that the receiver has knowledge only of the statistics of the Rayleigh-fading process (i.e., its mean and its autocovariance function) but no *a priori* knowledge of the realizations of $\{\mathbb{H}_k\}$. Note that this does not prevent the receiver from performing channel estimation. We merely view the transmission of training sequences to learn the channel at the receiver as a specific form of channel coding. This implies that in our setup the overhead associated with the transmission of such sequences is automatically accounted for.

Throughout the paper, we also assume no *a priori* CSI at the transmitter. The transmitter has only knowledge of the statistics of the fading process. This assumption is reasonable in a high-mobility scenario, where fast channel variations make channel tracking at the transmitter unfeasible. It is also appropriate for mission-critical applications where it may be desirable to avoid the creation of the feedback link required to provide CSI at the transmitter.

III. Maximum Coding Rate

We next introduce the notion of a channel code for the channel (1). For simplicity, we shall restrict ourselves to codes whose blocklength n is an integer multiple of the coherence interval n_{c}, i.e., $n = ln_{\mathrm{c}}$ for some $l \in \mathbb{N}$.

Definition 1: An $(l, n_{\mathrm{c}}, M, \epsilon, \rho)$ code for the channel (1) consists of

- An encoder $f : \{1, \ldots, M\} \to \mathbb{C}^{n_{\mathrm{c}} \times m_{\mathrm{t}} l}$ that maps the message $J \in \{1, \ldots, M\}$ to a codeword in the set $\{\mathsf{C}_1, \ldots, \mathsf{C}_M\}$. Since each codeword C_m, $m = 1, \ldots, M$, spans l coherence intervals, it is convenient to express it as the concatenation of l subcodewords

$$\mathsf{C}_m = \big[\mathsf{C}_{m,1}, \cdots, \mathsf{C}_{m,l}\big]. \tag{2}$$

  We require that each subcodeword $\mathsf{C}_{m,k} \in \mathbb{C}^{n_{\mathrm{c}} \times m_{\mathrm{t}}}$ satisfies the power constraint

$$\mathrm{tr}\big\{\mathsf{C}_{m,k}^H \mathsf{C}_{m,k}\big\} = n_{\mathrm{c}}\rho, \;\; m = 1, \ldots, M, \;\; k = 1, \ldots, l. \tag{3}$$

  Evidently, (3) implies the per-codeword power constraint[1]

$$\mathrm{tr}\big\{\mathsf{C}_m^H \mathsf{C}_m\big\} = ln_{\mathrm{c}}\rho \tag{4}$$

$$= n\rho. \tag{5}$$

  Since the noise has unit variance, ρ in (4) can be thought of as the SNR.

- A decoder $g : \mathbb{C}^{n_{\mathrm{c}} \times m_{\mathrm{t}} l} \to \{1, \ldots, M\}$ satisfying a maximum error probability constraint

$$\max_{1 \le j \le M} \Pr\big[g\big(\mathbb{Y}^l\big) \neq J \,|\, J = j\big] \le \epsilon \tag{6}$$

  where

$$\mathbb{Y}^l = \big[\mathbb{Y}_1, \cdots, \mathbb{Y}_l\big] \tag{7}$$

  is the channel output induced by the transmitted codeword

$$\mathsf{X}^l = \big[\mathsf{X}_1, \cdots, \mathsf{X}_l\big] = f(j) \tag{8}$$

  according to (1).

The *maximal channel coding rate* $R^*(l, n_{\mathrm{c}}, \epsilon, \rho)$ is defined as the largest rate $(\ln M)/(ln_{\mathrm{c}})$ for which there exists an $(l, n_{\mathrm{c}}, M, \epsilon, \rho)$ code. Formally,

$$R^*(l, n_{\mathrm{c}}, \epsilon, \rho) \triangleq \sup\left\{\frac{\ln M}{ln_{\mathrm{c}}} \,:\, \exists (l, n_{\mathrm{c}}, M, \epsilon, \rho) \text{ code}\right\}. \tag{9}$$

Recall that neither the encoder nor the decoder are assumed to have access to side information about the fading channel. For the case when CSI is available at the receiver, $R^*(l, n_{\mathrm{c}}, \epsilon, \rho)$ has been characterized up to second order for specific scenarios in [27]–[29].

The maximal channel coding rate $R^*(l, n_{\mathrm{c}}, \epsilon, \rho)$ captures the fundamental tension between the error probability ϵ and the transmission rate R^* for a given blocklength $n = ln_{\mathrm{c}}$ and SNR ρ. Furthermore, its dependency on the coherence interval n_c, on the number of diversity branches l, and on the number of transmit and receive antennas[2] allows one to study how this tension depends on the characteristics of the fading channel.

IV. Relation to Previous Results

Most of the results available in the literature can be interpreted as asymptotic characterizations of $R^*(l, n_{\mathrm{c}}, \epsilon, \rho)$ for $l \to \infty$, or $n_{\mathrm{c}} \to \infty$, or $\rho \to \infty$, or a combination of these limits.

Ergodic capacity: For the case when $l \to \infty$ for fixed n_{c}, fixed ρ, and fixed $0 < \epsilon < 1$, the maximum coding rate $R^*(l, n_{\mathrm{c}}, \epsilon, \rho)$ converges to the ergodic capacity $C_{\mathrm{erg}}(\rho)$

$$\lim_{l \to \infty} R^*(l, n_{\mathrm{c}}, \epsilon, \rho) = C_{\mathrm{erg}}(\rho) = \frac{1}{n_{\mathrm{c}}} \sup I(\mathbb{X}; \mathbb{Y}) \tag{10}$$

where $\mathbb{X} \in \mathbb{C}^{n_{\mathrm{c}} \times m_{\mathrm{t}}}$ denotes the channel input, $\mathbb{Y} \in \mathbb{C}^{n_{\mathrm{c}} \times m_{\mathrm{r}}}$ is the corresponding channel output, obtained through (1), and the supremum in (10) is over all probability distributions on $\mathbb{X}$ satisfying $\mathrm{tr}\{\mathbb{X}^H \mathbb{X}\} = n_{\mathrm{c}}\rho$ almost surely. Note that, by the strong converse [30], the ergodic capacity $C_{\mathrm{erg}}(\rho)$ does not

[1] It is more common in information-theoretic analyses to impose a power constraint per codeword and not per coherence interval. The benefit of the per-codeword power constraint is that it leads to simple closed-form expressions for capacity. However, practical systems typically operate under constraint (3).

[2] This dependency is not made explicit in the notation used in (9), in order to keep the notation compact.

depend on ϵ. Although $C_{\mathrm{erg}}(\rho)$ is not known in closed form when CSI is not available *a priori* at the receiver, its high-SNR behavior is well understood [19], [31], [32], [14], [17]. Specifically, Zheng and Tse [14] showed that, under the assumption $n_{\mathrm{c}} > 1$,

$$C_{\mathrm{erg}}(\rho) = m^* \left(1 - \frac{m^*}{n_{\mathrm{c}}}\right) \ln \rho + \mathcal{O}(1), \quad \rho \to \infty \quad (11)$$

where

$$m^* = \min\{m_{\mathrm{t}}, m_{\mathrm{r}}, \lfloor n_{\mathrm{c}}/2 \rfloor\}. \quad (12)$$

We remark that (11) holds also when the maximization in (10) is performed under the less stringent constraint that $\mathbb{E}\big[\mathrm{tr}\{\mathbb{X}^H\mathbb{X}\}\big] \leq n_{\mathrm{c}}\rho$. Since $C_{\mathrm{erg}}(\rho) = \min\{m_{\mathrm{t}}, m_{\mathrm{r}}\} \ln \rho + \mathcal{O}(1)$ for the case when the receiver has perfect CSI [11], we see from (11) that the *prelog* penalty due to lack of *a priori* CSI is equal to $(m^*)^2/n_{\mathrm{c}}$ (provided that $n_{\mathrm{c}} \geq m_{\mathrm{t}} + m_{\mathrm{r}}$). This is roughly m^* times the number of pilots per time-frequency slot needed to learn the channel at the receiver when m^* transmit antennas are used. The prelog penalty vanishes as n_{c} becomes large.

By tightening the high-SNR expansion (11) [14], [17], one obtains an accurate finite-SNR approximation of capacity [21], [33]. The input distribution that achieves the first two terms in the resulting high-SNR expansion of $C_{\mathrm{erg}}(\rho)$ depends on the relationship between n_{c}, m_{t} and m_{r}. When $n_{\mathrm{c}} \geq m_{\mathrm{t}} + m_{\mathrm{r}}$, it is optimal at high SNR to choose $\mathbb{X}$ to be a scaled isotropically distributed matrix that has orthonormal columns [14]. This input distribution is sometimes referred to as USTM. When $n_{\mathrm{c}} < m_{\mathrm{t}} + m_{\mathrm{r}}$, Beta-variate space-time modulation (BSTM) should be used instead [17]. In BSTM, the USTM unitary matrix is multiplied by a diagonal matrix whose nonzero entries are distributed as the square-root of the eigenvalues of a Beta-distributed random matrix. Throughout this paper, we shall focus on the case $n_{\mathrm{c}} \geq m_{\mathrm{t}} + m_{\mathrm{r}}$.

Although the ergodic capacity captures the rate penalty due to the channel-estimation overhead, and although its high-SNR expansion (11) describes compactly how this penalty depends on the channel coherence interval, the infinite-blocklength nature of (10) and its independence on the packet reliability ϵ limit its usefulness for the short-packet scenario considered in this paper.

Outage capacity: For the case when $n_{\mathrm{c}} \to \infty$ for fixed l, ϵ, and ρ, the maximum coding rate $R^*(l, n_{\mathrm{c}}, \epsilon, \rho)$ converges to the outage capacity $C_{\mathrm{out}}(\rho, \epsilon)$, defined as [34]

$$\lim_{n_{\mathrm{c}}\to\infty} R^*(l, n_{\mathrm{c}}, \epsilon, \rho) = C_{\mathrm{out}}(\rho, \epsilon)$$
$$= \sup\left\{R : \inf_{\{\mathsf{Q}_k\}_{k=1}^l} P_{\mathrm{out}}\big(\{\mathsf{Q}_k\}_{k=1}^l, R\big) \leq \epsilon\right\}. \quad (13)$$

Here, $P_{\mathrm{out}}(\cdot,\cdot)$ is the outage probability

$$P_{\mathrm{out}}\big(\{\mathsf{Q}_k\}_{k=1}^l, R\big)$$
$$= \Pr\left\{\frac{1}{l}\sum_{k=1}^{l} \ln\det(\mathsf{I}_{m_{\mathrm{r}}} + \mathbb{H}_k^H \mathsf{Q}_k \mathbb{H}_k) \leq R\right\} \quad (14)$$

where, for the Rayleigh-fading case considered in this paper, $\{\mathsf{Q}_k\}$, $k = 1, \ldots, l$, are $m_{\mathrm{t}} \times m_{\mathrm{t}}$ diagonal matrices with nonnegative entries that satisfy $\mathrm{tr}\{\mathsf{Q}_k\} = \rho$, and where the infimum in (13) is over all $\{\mathsf{Q}_k\}$. For the case $l = 1$, Telatar [11] conjectured that the optimal diagonal matrix Q_1 is of the form

$$\mathsf{Q}_1 = \frac{\rho}{m}\,\mathrm{diag}\{\underbrace{1, \ldots, 1}_{m}, \underbrace{0, \ldots, 0}_{m_{\mathrm{t}}-m}\} \quad (15)$$

for some $m \in \{1, \ldots, m_{\mathrm{t}}\}$. This conjecture was proved in [35] for the multiple-input single-output case.

The outage capacity in (13) characterizes in an implicit way the tension between the reliability ϵ and the throughput R. Note that (13) holds irrespectively of whether CSI is available at the receiver or not. Indeed, as the coherence interval n_{c} gets large, the cost of learning the channel at the receiver vanishes [36, p. 2632], [23]. Consequently, analyses based on outage capacity do not capture the overhead due to channel estimation, which may be significant for short-packet communications.

Diversity-multiplexing tradeoff: Consider the scenario where l and n_{c} are fixed, CSI is available at the receiver, and the packet error rate ϵ vanishes as a function of ρ according to

$$\epsilon(\rho) = \rho^{-d\,l} \quad (16)$$

where $d \in \{0, 1, \ldots, m_{\mathrm{t}} m_{\mathrm{r}}\}$ is the so-called *spatial diversity gain*. For the case when $n_{\mathrm{c}} \geq m_{\mathrm{t}} + m_{\mathrm{r}} - 1$, Zheng and Tse proved that [1]

$$\lim_{\rho\to\infty} \frac{R^*(n_{\mathrm{c}}, l, \epsilon(\rho), \rho)}{\ln \rho} = r(d) \quad (17)$$

where the *multiplexing gain* $r(d)$ is the piece-wise linear function connecting the points

$$r\big((m_{\mathrm{t}} - k)(m_{\mathrm{r}} - k)\big) = k, \quad k = 0, \ldots, \min\{m_{\mathrm{t}}, m_{\mathrm{r}}\}. \quad (18)$$

The condition $n_{\mathrm{c}} \geq m_{\mathrm{t}} + m_{\mathrm{r}} - 1$ has been relaxed to $n_{\mathrm{c}} \geq m_{\mathrm{t}}$ in [37], where an explicit code construction that achieves (17) is provided.

For the case when CSI is not available at the receiver and $n_{\mathrm{c}} \geq 2m^* + m_{\mathrm{r}} + 1$ (where m^* is given in (12)), the diversity-multiplexing tradeoff becomes [13], [38]

$$\lim_{\rho\to\infty} \frac{R^*(n_{\mathrm{c}}, l, \epsilon(\rho), \rho)}{\ln \rho} = \left(1 - \frac{m^*}{n_{\mathrm{c}}}\right) r(d). \quad (19)$$

The expressions in (18) and in (19) describe elegantly and succinctly the tradeoff between diversity gain and multiplexing gain. The price to be paid for such a characterization is its high-SNR nature, which may limit its significance for the scenarios analyzed in this paper.

Finite-SNR versions of the DMT have been proposed in [39], [40]. However, these extensions rely on the outage probability and are, in contrast to the original formulation in [13], only meaningful asymptotically as the blocklength tends to infinity.

To summarize, the performance metrics developed so far for the analysis of wireless systems, i.e., the ergodic capacity, the outage capacity, and the DMT have shortcomings when applied to short-packet wireless communications. We address these shortcomings in the next section by developing nonasymptotic bounds on $R^*(l, n_{\mathrm{c}}, \epsilon, \rho)$.

V. Bounds on the Maximal Coding Rate

A. Output Distribution Induced by USTM Inputs

Let $\mathbb{A}$ be an $n \times m$ ($n > m$) random matrix. We say that $\mathbb{A}$ is isotropically distributed if, for every deterministic $n \times n$ unitary matrix V, the matrix $\mathsf{V}\mathbb{A}$ has the same probability distribution as $\mathbb{A}$. A key ingredient of the nonasymptotic bounds on $R^*(l, n_{\mathrm{c}}, \epsilon, \rho)$ described in this section is the following closed-form expression for the probability density function (pdf) induced on the channel output $\mathbb{Y}_k$ in (1) when $\mathbb{X}_k$ is a scaled isotropically distributed matrix with orthonormal columns. Such an input distribution is commonly referred to as USTM. It will turn out convenient to consider a minor modification of the USTM distribution, in which only $\widetilde{m}_{\mathrm{t}}$ out of the available m_{t} transmit antennas are used.

Lemma 1: Assume that $n_{\mathrm{c}} \geq m_{\mathrm{t}} + m_{\mathrm{r}}$. Let $q = \min\{\widetilde{m}_{\mathrm{t}}, m_{\mathrm{r}}\}$ and $p = \max\{\widetilde{m}_{\mathrm{t}}, m_{\mathrm{r}}\}$. Let also $\mathbb{X} = \sqrt{\rho n_{\mathrm{c}}/\widetilde{m}_{\mathrm{t}}}\mathbb{U}$ where $\mathbb{U} \in \mathbb{C}^{n_{\mathrm{c}} \times \widetilde{m}_{\mathrm{t}}}$ ($1 \leq \widetilde{m}_{\mathrm{t}} \leq m_{\mathrm{t}}$) satisfies $\mathbb{U}^H\mathbb{U} = \mathsf{I}_{\widetilde{m}_{\mathrm{t}}}$ and is isotropically distributed. Further, let $\mathbb{Y} = \mathbb{X}\mathbb{H} + \mathbb{W}$ where $\mathbb{H} \in \mathbb{C}^{\widetilde{m}_{\mathrm{t}} \times m_{\mathrm{r}}}$ and $\mathbb{W} \in \mathbb{C}^{n_{\mathrm{c}} \times m_{\mathrm{r}}}$ are defined as in (1). The pdf of $\mathbb{Y}$ is given by

$$f_{\mathbb{Y}}(\mathsf{Y}) = \frac{\prod_{u=n_{\mathrm{c}}-q+1}^{n_{\mathrm{c}}} \Gamma(u)}{\pi^{m_{\mathrm{r}} n_{\mathrm{c}}} \prod_{u=1}^{\widetilde{m}_{\mathrm{t}}} \Gamma(u)} \frac{(1+\mu)^{\widetilde{m}_{\mathrm{t}}(n_{\mathrm{c}}-\widetilde{m}_{\mathrm{t}}-m_{\mathrm{r}})}}{\mu^{\widetilde{m}_{\mathrm{t}}(n_{\mathrm{c}}-\widetilde{m}_{\mathrm{t}})}} \cdot \psi_{\widetilde{m}_{\mathrm{t}}}(\sigma_1^2, \ldots, \sigma_{m_{\mathrm{r}}}^2). \tag{20}$$

Here, $\sigma_1 > \cdots > \sigma_{m_{\mathrm{r}}}$ denote the m_{r} nonzero singular values of Y, which are positive and distinct almost surely [41], $\mu = \rho n_{\mathrm{c}}/\widetilde{m}_{\mathrm{t}}$, and

$$\psi_{\widetilde{m}_{\mathrm{t}}}(\sigma_1^2, \ldots, \sigma_{m_{\mathrm{r}}}^2) = \frac{\det\{\mathsf{M}\}}{\prod_{i<j}^{m_{\mathrm{r}}} (\sigma_i^2 - \sigma_j^2)} \prod_{k=1}^{m_{\mathrm{r}}} \frac{e^{-\sigma_k^2/(1+\mu)}}{\sigma_k^{2(n_{\mathrm{c}}-m_{\mathrm{r}})}}. \tag{21}$$

The entries of the $p \times p$ real matrix M are given by

$$[\mathsf{M}]_{ij} = \begin{cases} b_i^{\widetilde{m}_{\mathrm{t}}-j}\widetilde{\gamma}\big(n_{\mathrm{c}}+j-p-\widetilde{m}_{\mathrm{t}}, b_i\mu/(1+\mu)\big), \\ \qquad 1 \leq i \leq m_{\mathrm{r}}, \quad 1 \leq j \leq \widetilde{m}_{\mathrm{t}} \\ e^{-b_i\mu/(1+\mu)} \left[\dfrac{\partial^{\widetilde{m}_{\mathrm{t}}-j}}{\partial\delta^{\widetilde{m}_{\mathrm{t}}-j}}\delta^{n_{\mathrm{c}}-i}\Big|_{\delta=\frac{\mu}{1+\mu}}\right], \\ \qquad m_{\mathrm{r}} < i \leq p, \quad 1 \leq j \leq \widetilde{m}_{\mathrm{t}} \\ b_i^{n_{\mathrm{c}}-j}e^{-b_i\mu/(1+\mu)} \\ \qquad 1 \leq i \leq m_{\mathrm{r}}, \quad \widetilde{m}_{\mathrm{t}} < j \leq p \end{cases} \tag{22}$$

where $b_i = \sigma_i^2$, $i = 1, \ldots, m_{\mathrm{r}}$, and

$$\widetilde{\gamma}(n, x) \triangleq \frac{1}{\Gamma(n)} \int_0^x t^{n-1}e^{-t}dt \tag{23}$$

denotes the regularized incomplete Gamma function.

Proof: The proof, which relies on the Itzykson-Zuber integral [42, Eq. (3.2)] and on repeated use of [43, Lem. 5], can be found, e.g., in [17, App. A] and, more recently, in [44]. ∎

Remark 1: A different expression for $f_{\mathbb{Y}}(\mathsf{Y})$ can be found in [32]. The expression in Lemma 1 appears to be easier to compute and more stable numerically.

B. USTM Dependence-Testing (DT) Lower Bound

We first present a lower bound on $R^*(l, n_{\mathrm{c}}, \epsilon, \rho)$ that is based on the dependence-testing (DT) bound [18, Th. 22] (maximal error probability) and makes use of the USTM-induced output distribution given in Lemma 1.

Theorem 1: Let $\Lambda_{k,\widetilde{m}_{\mathrm{t}},1} > \cdots > \Lambda_{k,\widetilde{m}_{\mathrm{t}},m_{\mathrm{r}}}$ be the ordered eigenvalues of $\mathbb{Z}_k^H\mathsf{D}_{\widetilde{m}_{\mathrm{t}}}\mathbb{Z}_k$ where $\{\mathbb{Z}_k\}_{k=1}^l$ are independent complex Gaussian $n_{\mathrm{c}} \times m_{\mathrm{r}}$ matrices with i.i.d. $\mathcal{CN}(0,1)$ entries, and

$$\mathsf{D}_{\widetilde{m}_{\mathrm{t}}} = \mathrm{diag}\Big\{\underbrace{1+\rho n_{\mathrm{c}}/\widetilde{m}_{\mathrm{t}}, \ldots, 1+\rho n_{\mathrm{c}}/\widetilde{m}_{\mathrm{t}}}_{\widetilde{m}_{\mathrm{t}}}, \underbrace{1, \ldots, 1}_{n_{\mathrm{c}}-\widetilde{m}_{\mathrm{t}}}\Big\} \tag{24}$$

for $\widetilde{m}_{\mathrm{t}} \in \{1, \ldots, m_{\mathrm{t}}\}$. It can be shown that the eigenvalues are positive and distinct almost surely. Let

$$S_{k,\widetilde{m}_{\mathrm{t}}} = \widetilde{m}_{\mathrm{t}}(n_{\mathrm{c}} - \widetilde{m}_{\mathrm{t}}) \ln \frac{\rho n_{\mathrm{c}}}{\widetilde{m}_{\mathrm{t}} + \rho n_{\mathrm{c}}} - \sum_{u=n_{\mathrm{c}}-q+1}^{n_{\mathrm{c}}} \ln \Gamma(u) + \sum_{u=1}^{\widetilde{m}_{\mathrm{t}}} \ln \Gamma(u) - \mathrm{tr}\big\{\mathbb{Z}_k^H\mathbb{Z}_k\big\} - \ln \psi_{\widetilde{m}_{\mathrm{t}}}(\Lambda_{k,\widetilde{m}_{\mathrm{t}},1}, \ldots, \Lambda_{k,\widetilde{m}_{\mathrm{t}},m_{\mathrm{r}}}) \tag{25}$$

where $q = \min\{\widetilde{m}_{\mathrm{t}}, m_{\mathrm{r}}\}$ and the function $\psi_{\widetilde{m}_{\mathrm{t}}} : \mathbb{R}_+^{m_{\mathrm{r}}} \to \mathbb{R}$ was defined in (21). Finally, let

$$\epsilon_{\mathrm{ub}}(M) = \min_{1 \leq \widetilde{m}_{\mathrm{t}} \leq m_{\mathrm{t}}} \mathbb{E}\Big[e^{-\left[\sum_{k=1}^l S_{k,\widetilde{m}_{\mathrm{t}}} - \ln(M-1)\right]^+}\Big]. \tag{26}$$

Then

$$R^*(l, n_{\mathrm{c}}, \epsilon, \rho) \geq \max\left\{\frac{\ln M}{n_{\mathrm{c}} l} \,:\, \epsilon_{\mathrm{ub}}(M) \leq \epsilon\right\}. \tag{27}$$

Proof: See Appendix A. ∎

C. Meta-converse (MC) Upper Bound

We next give an upper bound on $R^*(l, n_{\mathrm{c}}, \epsilon, \rho)$ that is based on the meta-converse (MC) theorem for maximal error probability of error [18, Th. 31] and uses the output distribution induced by the USTM input distribution (see (20)) as auxiliary output distribution.

Theorem 2: For a fixed $\widetilde{m}_{\mathrm{t}} \in [1, \ldots, m_{\mathrm{t}}]$, let the random variables $\{\bar{\mathbb{Y}}_k\}_{k=1}^l$ be i.i.d. $f_{\mathbb{Y}}$-distributed, with $f_{\mathbb{Y}}$, defined in (20), being the output distribution corresponding to an USTM input distribution over $\widetilde{m}_{\mathrm{t}}$ antennas. Let $\Delta_{k,\widetilde{m}_{\mathrm{t}},1} > \cdots > \Delta_{k,\widetilde{m}_{\mathrm{t}},m_{\mathrm{r}}}$ be the ordered eigenvalues of $\bar{\mathbb{Y}}_k^H\bar{\mathbb{Y}}_k$, $k = 1, \ldots, l$, and let

$$\mathbb{\Delta}_{k,\widetilde{m}_{\mathrm{t}}} = \mathrm{diag}\{\Delta_{k,\widetilde{m}_{\mathrm{t}},1}, \ldots, \Delta_{k,\widetilde{m}_{\mathrm{t}},m_{\mathrm{r}}}\}. \tag{28}$$

It can be shown that the eigenvalues are positive and distinct almost surely. Let $\{\mathsf{\Sigma}_k\}_{k=1}^l$ be $m_{\mathrm{t}} \times m_{\mathrm{t}}$ *diagonal* matrices with nonnegative diagonal entries, satisfying $\mathrm{tr}\{\mathsf{\Sigma}_k\} = n_{\mathrm{c}}\rho$, $k = 1, \ldots, l$. Let

$$\widetilde{\mathsf{\Sigma}}_k = \begin{bmatrix} \mathsf{I}_{m_{\mathrm{t}}} + \mathsf{\Sigma}_k & \mathbf{0} \\ \mathbf{0} & \mathsf{I}_{n_{\mathrm{c}}-m_{\mathrm{t}}} \end{bmatrix}. \tag{29}$$

Further let $\{\mathbb{U}_k\}_{k=1}^l$ be i.i.d. isotropically distributed (truncated) $n_{\mathrm{c}} \times m_{\mathrm{r}}$ unitary matrices, and let $\{\bar{\mathbb{Z}}_k\}_{k=1}^l$ be independent

complex Gaussian $n_{\mathrm{c}} \times m_{\mathrm{r}}$ matrices with i.i.d. $\mathcal{CN}(0,1)$ entries. Finally, let

$$\begin{aligned}\bar{c}_{\widetilde{m}_{\mathrm{t}}}(\Sigma_k) &= \widetilde{m}_{\mathrm{t}}(n_{\mathrm{c}} - \widetilde{m}_{\mathrm{t}}) \ln \frac{\rho n_{\mathrm{c}}}{\widetilde{m}_{\mathrm{t}}} \\ &\quad - \widetilde{m}_{\mathrm{t}}(n_{\mathrm{c}} - \widetilde{m}_{\mathrm{t}} - m_{\mathrm{r}}) \ln\left(1 + \frac{\rho n_{\mathrm{c}}}{\widetilde{m}_{\mathrm{t}}}\right) \\ &\quad - m_{\mathrm{r}} \ln \det \widetilde{\Sigma}_k - \sum_{u=n_{\mathrm{c}}-p+1}^{n_{\mathrm{c}}} \ln \Gamma(u) + \sum_{u=1}^{\widetilde{m}_{\mathrm{t}}} \ln \Gamma(u) \end{aligned} \quad (30)$$

$$\begin{aligned}T_{k,\widetilde{m}_{\mathrm{t}}}(\Sigma_k) &= \bar{c}_{\widetilde{m}_{\mathrm{t}}}(\Sigma_k) - \mathrm{tr}\{\mathbb{U}_k \mathbb{\Lambda}_{k,\widetilde{m}_{\mathrm{t}}} \mathbb{U}_k^H \widetilde{\Sigma}_k^{-1})\} \\ &\quad - \ln \psi_{\widetilde{m}_{\mathrm{t}}}(\Delta_{k,\widetilde{m}_{\mathrm{t}},1}, \ldots, \Delta_{k,\widetilde{m}_{\mathrm{t}},m_{\mathrm{r}}}) \end{aligned} \quad (31)$$

and

$$\begin{aligned}\bar{S}_{k,\widetilde{m}_{\mathrm{t}}}(\Sigma_k) &= \bar{c}_{\widetilde{m}_{\mathrm{t}}}(\Sigma_k) - \mathrm{tr}\{\bar{\mathbb{Z}}_k^H \bar{\mathbb{Z}}_k\} \\ &\quad - \ln \psi_{\widetilde{m}_{\mathrm{t}}}(\bar{\Lambda}_{k,\widetilde{m}_{\mathrm{t}},1}, \ldots, \bar{\Lambda}_{k,\widetilde{m}_{\mathrm{t}},m_{\mathrm{r}}}). \end{aligned} \quad (32)$$

Here, $\bar{\Lambda}_{k,\widetilde{m}_{\mathrm{t}},1} > \cdots > \bar{\Lambda}_{k,\widetilde{m}_{\mathrm{t}},m_{\mathrm{r}}}$ are the ordered eigenvalues of $\mathbb{Z}_k^H \widetilde{\Sigma}_k \mathbb{Z}_k$ (which are positive and distinct almost surely), $p = \max\{\widetilde{m}_{\mathrm{t}}, m_{\mathrm{r}}\}$, and $\psi_{\widetilde{m}_{\mathrm{t}}}$ is defined in (21). Then, for every n and for every $0 < \epsilon < 1$, the maximal channel coding rate $R^*(l, n_{\mathrm{c}}, \epsilon, \rho)$ is upper bounded by

$$\begin{aligned}&R^*(l, n_{\mathrm{c}}, \epsilon, \rho) \leq \\ &\quad \min_{1 \leq \widetilde{m}_{\mathrm{t}} \leq m_{\mathrm{t}}} \sup_{\{\Sigma_k\}_{k=1}^l} \frac{1}{n} \ln \frac{1}{\Pr\left\{\sum_{k=1}^l T_{k,\widetilde{m}_{\mathrm{t}}}(\Sigma_k) \geq \gamma\right\}} \end{aligned} \quad (33)$$

where $\gamma = \gamma\big(\{\Sigma_k\}_{k=1}^l\big)$ is the solution of

$$\Pr\left\{\sum_{k=1}^l \bar{S}_{k,\widetilde{m}_{\mathrm{t}}}(\Sigma_k) \leq \gamma\right\} = \epsilon. \quad (34)$$

Proof: See Appendix B. ∎

Remark 2: To facilitate its numerical evaluation, the MC upper bound (33) can be relaxed by using [18, Eq. (102)], which yields

$$\begin{aligned}&R^*(l, n_{\mathrm{c}}, \epsilon, \rho) \leq \min_{1 \leq \widetilde{m}_{\mathrm{t}} \leq m_{\mathrm{t}}} \sup_{\{\Sigma_k\}_{k=1}^l} \inf_{\lambda > 0} \\ &\quad \frac{1}{n}\left[\lambda - \ln\left(\left[\Pr\left\{\sum_{k=1}^l \bar{S}_{k,\widetilde{m}_{\mathrm{t}}}(\Sigma_k) \leq \lambda\right\} - \epsilon\right]^+\right)\right]. \end{aligned} \quad (35)$$

We will use this upper bound in the numerical evaluations reported in Section VII.

Remark 3: A converse bound that holds when the per-coherence-interval power constraint (4) is replaced by the less stringent (and perhaps more common) per-codeword power constraint

$$\mathrm{tr}\big\{\mathsf{C}_m^H \mathsf{C}_m\big\} \leq l n_{\mathrm{c}} \rho \quad (36)$$

can be obtained by evaluating the supremum in (33) and (35) over all $\{\Sigma_k\}_{k=1}^l$ that satisfy

$$\sum_{k=1}^l \mathrm{tr}\{\Sigma_k\} \leq l n_{\mathrm{c}} \rho. \quad (37)$$

VI. Bounds on the Coding Rate for Orthogonal Space-Time Codes

In the previous section, we provided bounds on the maximum coding rate without imposing any constraint on how the multiple antennas available at the transmitter should be used. In this section, we focus on orthogonal space-time codes that use the available transmit antennas to provide transmit diversity and, hence, improve reliability. Specifically, we consider a setup where an outer code, defined along the same lines as in Definition 1, is combined with a specific orthogonal space-time inner code. By treating this inner code as part of the channel, one can obtain achievability and converse bounds similar to the ones reported in Theorems 1 and 2. For simplicity, we shall focus on the 2×2 and 4×4 MIMO configurations.

These bounds, which pertain to the case when the transmit antennas are used to provide full spatial diversity, are then compared in Section VII to the general bounds in Theorems 1 and 2. This will allow us to characterize the rate penalty incurred by employing diversity-exploiting transmission strategies.

A. 2×2 *Case: Alamouti*

For the 2×2 case, we consider an Alamouti space-time inner code [45]. In order to analyze the finite-blocklength performance of such a scheme when CSI is not *a priori* available at the receiver, we proceed as in Section V: we first obtain a closed-form expression for the output distribution induced by the Alamouti scheme, and then use this output distribution to obtain a DT lower bound and a MC upper bound on the maximum coding rate obtainable with such a scheme.

We assume that the coherence interval n_{c} is even, and we let the $n_{\mathrm{c}} \times 2$ input matrix X_k in (1) be given by

$$\mathsf{X}_k = [\boldsymbol{a}_k \quad e(\boldsymbol{a}_k)] \quad (38)$$

where $\boldsymbol{a}_k$ is an n_{c}-dimensional vector satisfying $\|\boldsymbol{a}_k\|^2 = \rho n_{\mathrm{c}}/2$, and where the function $e : \mathbb{C}^{n_{\mathrm{c}}} \to \mathbb{C}^{n_{\mathrm{c}}}$ maps an input vector $\boldsymbol{a}$ into the output vector $\boldsymbol{b}$ according to the Alamouti rule [45]:

$$[\boldsymbol{b}]_{2l-1} = [e(\boldsymbol{a})]_{2l-1} = [\boldsymbol{a}]_{2l}^*, \quad l = 1, 2, \ldots, n_{\mathrm{c}}/2 \quad (39a)$$

$$[\boldsymbol{b}]_{2l} = [e(\boldsymbol{a})]_{2l} = -[\boldsymbol{a}]_{2l-1}^*, \quad l = 1, 2, \ldots, n_{\mathrm{c}}/2. \quad (39b)$$

In Lemma 2 below, we provide the pdf of the channel output $\mathbb{Y}$ induced by an input matrix $\mathbb{X}$ constructed as in (38) and whose first column $\boldsymbol{A}$ is uniformly distributed over the hypersphere of radius $\sqrt{\rho n_{\mathrm{c}}/2}$ (this corresponds to USTM for the case of a single transmit antenna).

Lemma 2: Assume that $m_{\mathrm{t}} = m_{\mathrm{r}} = 2$ and that n_{c} is even and larger or equal to 4. Let

$$\mathbb{X} = [\boldsymbol{A} \quad e(\boldsymbol{A})] \quad (40)$$

where $e(\cdot)$ is defined in (39) and where $\boldsymbol{A} = \sqrt{\rho n_{\mathrm{c}}/2}\, \boldsymbol{U}$, with $\boldsymbol{U}$ being an isotropically distributed unit-norm n_{c}-dimensional complex random vector. Let $\mathbb{Y} = \big[\boldsymbol{Y}_1 \ \boldsymbol{Y}_2\big] = \mathbb{X}\mathbb{H} + \mathbb{W}$, where $\mathbb{H} \in \mathbb{C}^{2\times 2}$ and $\mathbb{W} \in \mathbb{C}^{n_{\mathrm{c}} \times 2}$ are defined similarly as in (1). Furthermore, let

$$\hat{\mathbb{Y}} = [\boldsymbol{Y}_1 \quad e(\boldsymbol{Y}_1) \quad \boldsymbol{Y}_2 \quad e(\boldsymbol{Y}_2)] \quad (41)$$

and let Σ_1 and Σ_3 (with realizations σ_1 and σ_3, respectively) be the first and the third largest eigenvalue of the 4×4 matrix $\rho n_{\mathrm{c}}/(2+\rho n_{\mathrm{c}})\hat{\mathbb{Y}}^H\hat{\mathbb{Y}}$.[3] Then, the pdf of $\mathbb{Y}$ is given by

$$f_{\mathbb{Y}}(\mathsf{Y}) = \frac{\exp\bigl(\mathrm{tr}\{\mathsf{Y}^H\mathsf{Y}\}\bigr)}{\pi^{2n_{\mathrm{c}}}(1+\rho n_{\mathrm{c}}/2)^{2n_{\mathrm{c}}}} \frac{\Gamma(n_{\mathrm{c}})}{(\sigma_1-\sigma_3)^4} \det\{\mathsf{M}(\sigma_1,\sigma_3)\} \quad (42)$$

where the 4×4 matrix M is given by

$$\begin{bmatrix} e^{\sigma_1}\widetilde{\gamma}(n_{\mathrm{c}}-5,\sigma_1) & (n_{\mathrm{c}}-2)\sigma_1^{n_{\mathrm{c}}-3} & (n_{\mathrm{c}}-3)\sigma_1^{n_{\mathrm{c}}-4} & (n_{\mathrm{c}}-4)\sigma_1^{n_{\mathrm{c}}-5} \\ e^{\sigma_1}\widetilde{\gamma}(n_{\mathrm{c}}-4,\sigma_1) & \sigma_1^{n_{\mathrm{c}}-2} & \sigma_1^{n_{\mathrm{c}}-3} & \sigma_1^{n_{\mathrm{c}}-4} \\ e^{\sigma_3}\widetilde{\gamma}(n_{\mathrm{c}}-5,\sigma_3) & (n_{\mathrm{c}}-2)\sigma_3^{n_{\mathrm{c}}-3} & (n_{\mathrm{c}}-3)\sigma_3^{n_{\mathrm{c}}-4} & (n_{\mathrm{c}}-4)\sigma_3^{n_{\mathrm{c}}-5} \\ e^{\sigma_3}\widetilde{\gamma}(n_{\mathrm{c}}-4,\sigma_3) & \sigma_3^{n_{\mathrm{c}}-2} & \sigma_3^{n_{\mathrm{c}}-3} & \sigma_3^{n_{\mathrm{c}}-4} \end{bmatrix}$$

if $n_{\mathrm{c}} > 4$, and by

$$\begin{bmatrix} e^{\sigma_1} & 2\sigma_1 & 1 & 0 \\ e^{\sigma_1} & \sigma_1^2 & \sigma_1 & 1 \\ e^{\sigma_3} & 2\sigma_3 & 1 & 0 \\ e^{\sigma_3} & \sigma_3^2 & \sigma_3 & 1 \end{bmatrix} \quad (43)$$

if $n_{\mathrm{c}} = 4$.

Proof: The proof follows along the same lines as the proof of Lemma 1. ∎

Remark 4: Note that although $\boldsymbol{A}$ in (40) is isotropically distributed, the matrix $\mathbb{X}$ is not. Hence, $\mathbb{X}$ does not follow a USTM distribution.

Treating the Alamouti space-time inner code as part of the channel, we next report lower and upper bounds on the maximum coding rate $R^*_{\mathrm{ala}}(l, n_{\mathrm{c}}, \epsilon, \rho)$ achievable when an Alamouti space-time inner code is used. These bounds rely on the closed-form expression for $f_{\mathbb{Y}}(\cdot)$ given in (42).

1) DT lower bound: We provide first an achievability bound, which is based on the DT bound [18, Th. 22].

Theorem 3: Let $\{\mathbb{Z}_k\}_{k=1}^l$ be independent complex Gaussian $n_{\mathrm{c}} \times 2$ matrices with i.i.d. $\mathcal{CN}(0,1)$ entries. Let

$$\mathsf{D} = \mathrm{diag}\left\{1+\frac{\rho n_{\mathrm{c}}}{2}, 1+\frac{\rho n_{\mathrm{c}}}{2}, \underbrace{1, \ldots, 1}_{n_{\mathrm{c}}-2}\right\} \quad (44)$$

$\mathbb{V}_k = \bigl[\boldsymbol{V}_{k,1}\ \boldsymbol{V}_{k,2}\bigr] = \mathsf{D}^{1/2}\mathbb{Z}_k$, and

$$\hat{\mathbb{V}}_k = \bigl[\boldsymbol{V}_{k,1}\, e(\boldsymbol{V}_{k,1})\, \boldsymbol{V}_{k,2}\, e(\boldsymbol{V}_{k,2})\bigr] \quad (45)$$

where the function $e(\cdot)$ was defined in (39). Furthermore, let $\Sigma_{k,1}$ and $\Sigma_{k,3}$ be the first and third largest eigenvalue of $(\rho n_{\mathrm{c}}/(2+\rho n_{\mathrm{c}}))\hat{\mathbb{V}}_k^H\hat{\mathbb{V}}_k$ (which are positive and distinct almost surely), let

$$\begin{aligned} S_k = {} & \mathrm{tr}\{\mathbb{Z}_k^H\mathsf{D}\mathbb{Z}_k\} - \mathrm{tr}\{\mathbb{Z}_k^H\mathbb{Z}\} - \ln\Gamma(n_{\mathrm{c}}) \\ & + \ln\det\{\mathsf{M}(\Sigma_{k,1},\Sigma_{k,3})\} - 4\ln(\Sigma_{k,1}-\Sigma_{k,3}) \end{aligned} \quad (46)$$

and let

$$\epsilon_{\mathrm{ala}}(M) = \mathbb{E}\left[\exp\left\{-\left[\sum_{k=1}^{l} S_k - \ln(M-1)\right]^+\right\}\right]. \quad (47)$$

Then

$$R^*_{\mathrm{ala}}(l, n_{\mathrm{c}}, \epsilon, \rho) \geq \max\left\{\frac{\ln M}{n_{\mathrm{c}} l} \,:\, \epsilon_{\mathrm{ala}}(M) \leq \epsilon\right\}. \quad (48)$$

[3]The matrix $\hat{\mathbb{Y}}^H\hat{\mathbb{Y}}$ has two distinct positive eigenvalues with multiplicity two almost surely.

Proof: The proof follows along the same lines as the proof of Theorem 1. ∎

2) MC upper bound: Using [18, Th. 22] and [18, Eq. (102)] we obtain the following converse bound on $R^*_{\mathrm{ala}}(l, n_{\mathrm{c}}, \epsilon, \rho)$.

Theorem 4: Let S_k be defined as in (46). Then

$$\begin{aligned} & R^*_{\mathrm{ala}}(l, n_{\mathrm{c}}, \epsilon, \rho) \\ & \qquad \leq \inf_{\lambda>0} \frac{1}{n}\left[\lambda - \ln\left(\Pr\left\{\sum_{k=1}^{l} S_k \leq \lambda\right\} - \epsilon\right)\right]. \end{aligned} \quad (49)$$

Proof: The proof follows along the same lines as the proof of Theorem 2. ∎

B. 4×4 Case: Frequency-Switched Transmit Diversity

Since no generalization of the Alamouti space-time inner code exists beyond the 2×2 configuration [46], we consider instead for the 4×4 case the combination of Alamouti and frequency-switched transmit diversity (FSTD) used in LTE [47, Sec. 11.2.2.1]. According to this scheme, in the odd time-frequency slots only transmit antennas 1 and 2 are used, and in the even time-frequency slots only transmit antennas 3 and 4 are used. In each time-frequency slot, an Alamouti space-time inner code is used for transmission. For example, for the case $n_{\mathrm{c}} = 4$, this scheme results in the following 4×4 input matrix

$$\begin{bmatrix} a_1 & a_2 & 0 & 0 \\ 0 & 0 & b_1 & b_2 \\ -a_2^* & a_1^* & 0 & 0 \\ 0 & 0 & -b_2^* & b_1^* \end{bmatrix} \quad (50)$$

where $|a_1|^2 + |a_2|^2 = |b_1|^2 + |b_2|^2 = \rho$. The combination of Alamouti and FSTD transforms a 4×4 MIMO channel with coherence interval n_{c} into two parallel 2×4 MIMO channels with coherence interval $n_{\mathrm{c}}/2$. Upper and lower bounds on the maximum coding rate achievable with this scheme can be obtained using a similar approach as for the 2×2 case.

VII. Numerical Results

Since a 5G standard for mission-critical MTC is not available yet, we base our numerical simulations on the setup analyzed in [2]. Specifically, we assume that packets of $n = 168$ symbols are used to transmit within one millisecond 14 OFDM symbols, each one consisting of 12 tones. The available bandwidth is of 10 MHz at a central frequency of 2 GHz. In a typical urban environment, one can obtain 12 frequency-diversity branches by spreading the tones uniformly over the available bandwidth [2]. Throughout, we set $\rho = 6$ dB. We consider both the case where the packet error rate is $\epsilon = 10^{-3}$, which may be appropriate for the exchange of short packets carrying control signaling, and the case $\epsilon = 10^{-5}$, which may be relevant for the transmission of critical information, e.g., in traffic-safety applications [3], [5]. For both cases, we shall compute the DT lower bound (27) and the MC upper bound (35). The numerical evaluation of the MC upper bound (35) is challenging because it involves a maximization over the diagonal matrices $\{\Sigma_k\}$. As for the

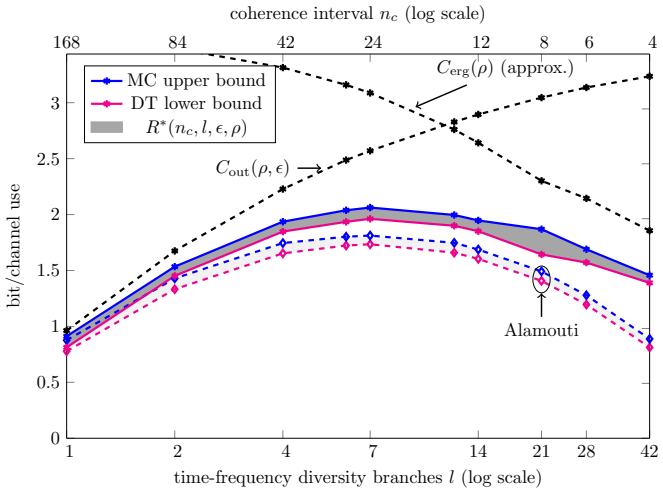

Fig. 1. $m_{\mathrm{t}} = m_{\mathrm{r}} = 2$, $n = 168$, $\epsilon = 10^{-3}$, $\rho = 6\,\mathrm{dB}$. Because of computational complexity, in the MC upper bound (35) the supremum over $\{\mathbf{\Sigma}_k\}_{k=1}^{l}$ is restricted to $\{\mathbf{\Sigma}_k\}_{k=1}^{l}$ of the form given in (51) when $l > 7$.

outage capacity in (13), the symmetry in (35) suggests that the supremum over $\{\mathbf{\Sigma}_k\}_{k=1}^{l}$ is achieved when

$$\mathbf{\Sigma}_k = \frac{\rho}{m_k}\,\mathrm{diag}\{\underbrace{1,\ldots,1}_{m_k},\underbrace{0,\ldots,0}_{m_{\mathrm{t}}-m_k}\} \tag{51}$$

for some $m_k \in \{1,\ldots,m_{\mathrm{t}}\}$, $k \in \{1,\ldots,l\}$. We can think of (51) as a finite-blocklength equivalent of Telatar conjecture [11]. Although far from conclusive, the numerical results reported in this section support the validity of this conjecture.

Control signaling: In Fig. 1 we plot[4] the DT lower bound (27) and the MC upper bound (35) for the 2×2 case. Here, $\epsilon = 10^{-3}$. These bounds delimit $R^*(l, n_{\mathrm{c}}, \epsilon, \rho)$ tightly and demonstrate that $R^*(l, n_{\mathrm{c}}, \epsilon, \rho)$ is not monotonic in the coherence interval n_{c}, but that there exists an optimal value n_{c}^*, or, equivalently, an optimal number $l^* = n/n_{\mathrm{c}}^*$ of time-frequency diversity branches, that maximizes $R^*(l, n_{\mathrm{c}}, \epsilon, \rho)$. A similar observation was reported in [21] for the single-antenna case. For $n_{\mathrm{c}} < n_{\mathrm{c}}^*$, the cost of estimating the channel dominates. For $n_{\mathrm{c}} > n_{\mathrm{c}}^*$, the bottleneck is the limited number of time-frequency diversity branches offered by the channel. For the parameters considered in Fig. 1, the optimal coherence interval length is $n_{\mathrm{c}}^* \approx 24$, which corresponds to about 7 time-frequency diversity branches.

In the figure, we also plot the outage capacity $C_{\mathrm{out}}(\rho, \epsilon)$ in (13) as a function of the number of time-frequency diversity branches $l = n/n_{\mathrm{c}}$ (with $n = 168$), and a lower bound on the ergodic capacity $C_{\mathrm{erg}}(\rho)$ as a function of the coherence interval n_{c}. This lower bound on $C_{\mathrm{erg}}(\rho)$, which is obtained by computing the mutual information on the RHS of (10) for the case when $\mathbb{X}$ is USTM-distributed and by optimizing over the number of active transmit antennas, approximates $C_{\mathrm{erg}}(\rho)$ accurately already at moderate SNR values [33].

As shown in the figure, $C_{\mathrm{out}}(\epsilon, \rho)$ provides a good approximation for $R^*(l, n_{\mathrm{c}}, \epsilon, \rho)$ only when l is small ($n_{\mathrm{c}} \approx n$), i.e., when the fading channel is essentially constant over the duration of the packet (quasi-static scenario). Furthermore, $C_{\mathrm{out}}(\epsilon, \rho)$ fails to capture the loss in throughput due to the channel estimation overhead, which is relevant for small n_{c}. For example, for $n_{\mathrm{c}} = 4$, the outage capacity overestimates $R^*(l, n_{\mathrm{c}}, \epsilon, \rho)$ by a factor two.

The lower bound on $C_{\mathrm{erg}}(\rho)$ plotted in the figure approximates $R^*(l, n_{\mathrm{c}}, \epsilon, \rho)$ poorly when n_{c} is large. For example, it overestimates $R^*(l, n_{\mathrm{c}}, \epsilon, \rho)$ by a factor four when $n_{\mathrm{c}} = 168$. As expected, the approximation gets better as n_{c} becomes smaller.

The number of active transmit antennas $\widetilde{m}_{\mathrm{t}}$ that maximizes the DT achievability bound is $\widetilde{m}_{\mathrm{t}} = 2$ (both antennas active) for $1 \le l \le 21$, and it is $\widetilde{m}_{\mathrm{t}} = 1$ (only one antenna active) for $l > 21$. The lower bound on $C_{\mathrm{erg}}(\rho)$, which also involves a maximization over the number of active antennas, exhibits the same behavior. We also note that the intersection between $C_{\mathrm{out}}(\epsilon, \rho)$ and $C_{\mathrm{erg}}(\rho)$ predicts coarsely the optimal number l^* of time-frequency diversity branches.

The optimal $\widetilde{m}_{\mathrm{t}}$ value for the MC upper bound (35) is again $\widetilde{m}_{\mathrm{t}} = 2$ for $1 \le l \le 21$ and $\widetilde{m}_{\mathrm{t}} = 1$ for $l > 21$. Furthermore, the optimal[5] $\{\mathbf{\Sigma}_k\}_{k=1}^{l}$ take all the same value and are equal to a 2×2 scaled identity matrix for $1 \le l \le 14$ and for $l = 28$, and to a 2×2 diagonal matrix with diagonal entries equal to ρ and to 0, respectively, for $l = 21$ and $l = 42$.

In the same figure, we plot the achievability and the converse bounds for the case when an Alamouti code is used as inner code. One can see that for small values of l, the Alamouti scheme is almost optimal, but the gap between the DT lower bound and the Alamouti converse increases as l grows. This is in agreement with the findings based on an outage-capacity analysis reported in [2]. However, in contrast to what has been observed for the outage capacity, our bounds on $R^*(l, n_{\mathrm{c}}, \epsilon, \rho)$ reveal that it is better to switch off the second transmit antenna when l is large. In this regime, the cost of estimating the channel resulting from the use of a second antenna overcomes the advantage of having additional spatial degrees of freedom.

We would like to emphasize that, in contrast to our approach, outage-capacity-based analyses are inherently insensitive to the cost of estimating the fading parameters and are therefore not suitable to capture the channel-estimation overhead. Although the high-SNR ergodic capacity approximation (11) and the DMT for the case of no *a priori* CSI (19) do predict that transmit antennas must be progressively switched off as l grows large, their predictions are coarse. Indeed, our numerical results suggest that the second transmit antenna should be switched off when $l > 21$, or equivalently $n_{\mathrm{c}} < 8$, whereas both (11) and (19) suggest that the second antenna should be switched off only when $n_{\mathrm{c}} \le 3$. Using both antennas when $3 < n_{\mathrm{c}} < 8$ results in a rate loss that can be as large as 30%.

The gap between the DT lower bound and the MC upper bound in Fig. 1 is largest around the value of n_{c} (or equivalently $l = n/n_{\mathrm{c}}$) for which the second transmit antenna must be switched off. One could tighten both the DT and the MC bound by considering a larger class of input distributions (and the induced class of output distributions for the MC bound). For example, one could drop the assumption that the input

[4]The numerical routines used to obtain these results are available at https://github.com/yp-mit/spectre

[5]Because of computational complexity, for $l > 7$ only $\{\mathbf{\Sigma}_k\}_{k=1}^{l}$ of the form given in (51) are considered.

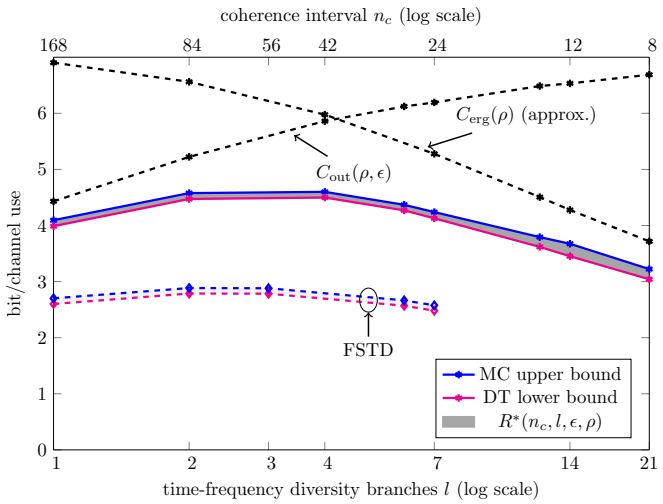

Fig. 2. $m_{\text{t}} = m_{\text{r}} = 4$, $n = 168$, $\epsilon = 10^{-3}$, $\rho = 6\,\text{dB}$. Because of computational complexity, in the MC upper bound (35) the supremum over $\{\mathbf{\Sigma}_k\}_{k=1}^{l}$ is restricted to $\{\mathbf{\Sigma}_k\}_{k=1}^{l}$ of the form given in (51).

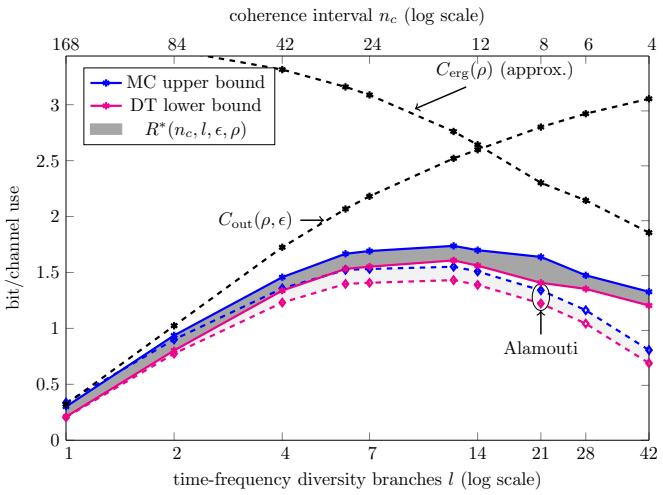

Fig. 3. $m_{\text{t}} = m_{\text{r}} = 2$, $n = 168$, $\epsilon = 10^{-5}$, $\rho = 6\,\text{dB}$. Because of computational complexity, in the MC upper bound (35) the supremum over $\{\mathbf{\Sigma}_k\}_{k=1}^{l}$ is restricted to $\{\mathbf{\Sigma}_k\}_{k=1}^{l}$ of the form given in (51) when $l > 7$.

distribution is identical across coherence intervals. Indeed, using a different number of transmit antennas in different coherence intervals could be beneficial since it would essentially allow one to extend the optimization in both (57) and (35) over fractional values of $\widetilde{m}_{\text{t}}$.

In Fig. 2, we present a similar comparison for the case of a 4×4 system. As shown in the figure, the gap between the MC upper bound and the DT lower bound is small, allowing for an accurate characterization of $R^*(l, n_{\text{c}}, \epsilon, \rho)$. In contrast, the gap between the DT lower bound and the FSTD upper bound is large, which suggests that using all 4 transmit antennas to provide spatial diversity is suboptimal even when the number of time-frequency diversity branches is limited (i.e., l is small). As in the 2×2 case, the transmit antennas should progressively be switched off as l increases, in order to mitigate the channel-estimation overhead. Specifically, the DT achievability bound is maximized by using 4 transmit antennas ($\widetilde{m}_{\text{t}} = 4$) when $1 \le l < 12$, by using 3 antennas when $l = 12$, and by using only two antennas when $12 < l \le 21$. Also in this case, the lower bound on $C_{\text{erg}}(\rho)$ and the MC upper bound exhibit a similar behavior.

Ultra-reliable communication: In Figs. 3 and 4, we consider the case $\epsilon = 10^{-5}$. We observe a similar behavior as for the case $\epsilon = 10^{-3}$, with the difference that the gap between the optimal schemes and the orthogonal space-time schemes (Alamouti for the 2×2 configuration, and FSTD for the 4×4 case) becomes smaller. This comes as no surprise, since the higher reliability requirement makes the exploitation of transmit diversity advantageous.

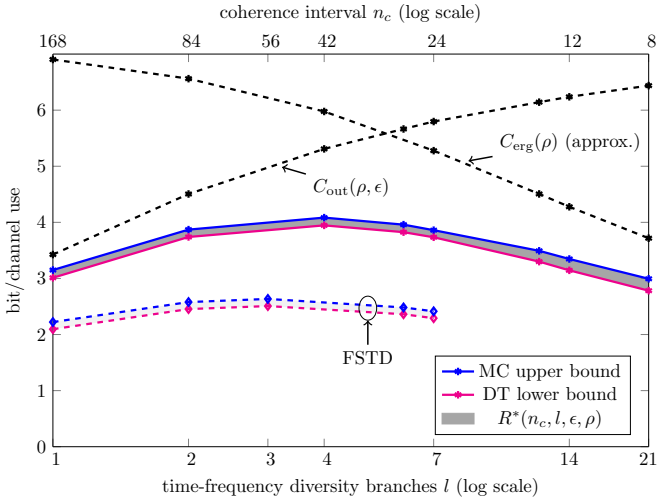

Fig. 4. $m_{\text{t}} = m_{\text{r}} = 4$, $n = 168$, $\epsilon = 10^{-5}$, $\rho = 6\,\text{dB}$. Because of computational complexity, in the MC upper bound (35) the supremum over $\{\mathbf{\Sigma}_k\}_{k=1}^{l}$ is performed only over $\{\mathbf{\Sigma}_k\}_{k=1}^{l}$ values of the form given in (51).

VIII. Conclusions

We presented finite-blocklength bounds on the maximum coding rate achievable over a MIMO Rayleigh block-fading channel, under the assumption that neither the transmitter nor the receiver have *a priori* CSI. Our bounds are explicit in the packet error rate ϵ, the coherence interval n_{c}, and the number of time-frequency diversity branches l. Furthermore, they allow one to determine, for a fixed packet size $n = n_{\text{c}}l$, the number of time-frequency diversity branches and the number of transmit antennas that maximize the rate. The optimal choice balances the rate gain resulting from exploiting of the available time-frequency-spatial resources, against the cost of estimating the channel coefficients over these resources. The bounds provide also an indication of whether the available transmit antennas should be used to provide transmit diversity or spatial multiplexing.

Our numerical results demonstrate that traditional infinite-blocklength performance metrics, such as the outage and the ergodic capacity, provide inaccurate estimates on the maximum coding rate when the packet size is short. They further fail to capture the fundamental tradeoff between reliability, throughput, latency and channel-estimation overhead. This suggests that the optimal design of the novel low-latency, ultra-reliable MTC that will be provided by next-generation wireless systems must rely on a more refined analysis of the interplay between packet-error probability, communication rate, and packet size, than the one

offered by traditional infinite-blocklength performance metrics.

Appendix A
Proof of Theorem 1

The transmitter uses only $\widetilde{m}_{\mathrm{t}}$ out of the available m_{t} antennas. This yields an $\widetilde{m}_{\mathrm{t}} \times m_{\mathrm{r}}$ MIMO Rayleigh block-fading channel. Let $\mathbb{X}_k = \sqrt{\rho n_{\mathrm{c}}/\widetilde{m}_{\mathrm{t}}}\mathbb{U}_k$, $k = 1, \ldots, l$, where $\{\mathbb{U}_k\}_{k=1}^{l}$ are independent, isotropically distributed $n_{\mathrm{c}} \times \widetilde{m}_{\mathrm{t}}$ random matrices with orthonormal columns. The induced channel outputs $\mathbb{Y}_k = \sqrt{\rho n_{\mathrm{c}}/\widetilde{m}_{\mathrm{t}}}\mathbb{U}_k\mathbb{H}_k + \mathbb{W}_k$, $k = 1, \ldots, l$, are i.i.d. $f_{\mathbb{Y}}$-distributed, where $f_{\mathbb{Y}}$ is given in (20). Let $\mathsf{U}^l = [\mathsf{U}_1, \ldots, \mathsf{U}_l]$. Since the channel is block-memoryless, the information density [18, Eq. (4)] can be decomposed as

$$\imath\big(\mathsf{U}^l; \mathsf{Y}^l\big) = \sum_{k=1}^{l} \imath(\mathsf{U}_k; \mathsf{Y}_k) = \sum_{k=1}^{l} \ln \frac{f_{\mathbb{Y}\,|\,\mathbb{U}}(\mathsf{Y}_k \,|\, \mathsf{U}_k)}{f_{\mathbb{Y}}(\mathsf{Y}_k)} \tag{52}$$

where

$$f_{\mathbb{Y}\,|\,\mathbb{U}}(\mathsf{Y}_k \,|\, \mathsf{U}_k) = \frac{e^{-\operatorname{tr}\left\{\mathsf{Y}_k^H\left(\mathsf{I}_{n_{\mathrm{c}}} + (\rho n_{\mathrm{c}}/\widetilde{m}_{\mathrm{t}})\mathsf{U}_k\mathsf{U}_k^H\right)^{-1}\mathsf{Y}_k\right\}}}{\pi^{m_{\mathrm{r}} n_{\mathrm{c}}}(1+\rho n_{\mathrm{c}}/\widetilde{m}_{\mathrm{t}})^{\widetilde{m}_{\mathrm{t}} m_{\mathrm{r}}}}. \tag{53}$$

We next note that, for every $n_{\mathrm{c}} \times n_{\mathrm{c}}$ unitary matrix V,

$$f_{\mathbb{Y}\,|\,\mathbb{U}}(\mathsf{Y} \,|\, \mathsf{V}^H\mathsf{U}) = f_{\mathbb{Y}\,|\,\mathbb{U}}(\mathsf{V}\mathsf{Y} \,|\, \mathsf{U}) \tag{54}$$

and

$$f_{\mathbb{Y}}(\mathsf{V}\mathsf{Y}) = f_{\mathbb{Y}}(\mathsf{Y}). \tag{55}$$

Consequently, the probability law of the information density $\imath(\mathsf{U}_k; \mathbb{Y}_k)$ in (52) (where $\mathbb{Y}_k \sim f_{\mathbb{Y}}$) does not depend on U_k. Without loss of generality, we shall then set $\mathsf{U}_k = \bar{\mathsf{U}}$, $k = 1, \ldots, l$, with

$$\bar{\mathsf{U}} = \begin{bmatrix} \mathsf{I}_{\widetilde{m}_{\mathrm{t}}} \\ \mathbf{0}_{n_{\mathrm{c}}-\widetilde{m}_{\mathrm{t}} \times \widetilde{m}_{\mathrm{t}}} \end{bmatrix}. \tag{56}$$

Using [18, Th. 22], we conclude that there exists an $(l, n_{\mathrm{c}}, M, \epsilon, \rho)$ code satisfying

$$\epsilon \le \mathbb{E}\left[\exp\left\{-\left[\sum_{k=1}^{l} \imath\big(\bar{\mathsf{U}}; \mathbb{Y}_k\big) - \ln(M-1)\right]^{+}\right\}\right] \tag{57}$$

where the expectation is with respect to $\mathbb{Y}_k \sim f_{\mathbb{Y}\,|\,\mathbb{U}}(\cdot \,|\, \bar{\mathsf{U}})$. Through algebraic manipulations, one can show that $\imath\big(\bar{\mathsf{U}}; \mathbb{Y}_k\big)$ has the same distribution as the random variable $S_{k,\widetilde{m}_{\mathrm{t}}}$ in (25). Minimizing (57) over the number of effectively used transmit antennas $\widetilde{m}_{\mathrm{t}}$, and solving the resulting inequality for the rate $(\ln M)/(n_{\mathrm{c}} l)$ yields (27).

Appendix B
Proof of Theorem 2

Fix $1 \le \widetilde{m}_{\mathrm{t}} \le m_{\mathrm{t}}$. To upper-bound $R^*(l, n_{\mathrm{c}}, \epsilon, \rho)$, we use the meta-converse theorem for maximal error probability [18, Th. 31] with auxiliary pdf

$$q_{\mathbb{Y}^l}(\mathsf{Y}^l) = \prod_{k=1}^{l} f_{\mathbb{Y}}(\mathsf{Y}_k) \tag{58}$$

where $f_{\mathbb{Y}}$ is the USTM-induced output pdf defined in (20). This yields

$$R^*(l, n_{\mathrm{c}}, \epsilon, \rho) \le \sup_{\mathsf{X}^l} \frac{1}{n} \ln \frac{1}{\beta_{1-\epsilon}\big(\mathsf{X}^l, q_{\mathbb{Y}^l}\big)} \tag{59}$$

where the supremum is over all codewords $\mathsf{X}^l \in \mathbb{C}^{n_{\mathrm{c}} \times m_{\mathrm{t}} l}$ satisfying the power constraint (3), and where $\beta_{1-\epsilon}(\cdot, \cdot)$ is defined as in [18, Eq. (105)].[6] By the Neyman-Pearson lemma, we have that

$$\beta_{1-\epsilon}(\mathsf{X}^l, q_{\mathbb{Y}^l}) = \Pr\big\{\imath\big(\mathsf{X}^l; \mathbb{Y}^l\big) \ge \gamma\big\}, \quad \mathbb{Y}^l \sim q_{\mathbb{Y}^l} \tag{60}$$

where γ is the solution of

$$\Pr\big\{\imath\big(\mathsf{X}^l; \mathbb{Y}^l\big) \le \gamma\big\} = \epsilon, \quad \mathbb{Y}^l \sim f_{\mathbb{Y}^l\,|\,\mathbb{X}^l}(\cdot \,|\, \mathsf{X}^l) \tag{61}$$

and where $\imath(\cdot; \cdot)$ is defined as in (52).

For a given codeword $\mathsf{X}^l = [\mathsf{X}_1, \ldots, \mathsf{X}_l]$, let

$$\mathsf{X}_k\mathsf{X}_k^H = \mathsf{V}_k\mathsf{\Sigma}_k\mathsf{V}_k^H, \quad k = 1, \ldots, l. \tag{62}$$

Here, $\mathsf{V}_k \in \mathbb{C}^{n_{\mathrm{c}} \times m_{\mathrm{t}}}$ contains the eigenvectors of $\mathsf{X}_k\mathsf{X}_k^H$, and $\mathsf{\Sigma}_k \in \mathbb{C}^{m_{\mathrm{t}} \times m_{\mathrm{t}}}$ is a diagonal matrix with nonnegative entries containing the m_{t} eigenvalues of $\mathsf{X}_k^H\mathsf{X}_k$. It follows from (54) and (55) that $\beta_{1-\epsilon}(\mathsf{X}^l, q_{\mathbb{Y}^l})$ depends on X^l only through the diagonal matrices $\{\mathsf{\Sigma}_k\}_{k=1}^{l}$. Hence, we can replace the infimum over X^l in (59) by an infimum over $\{\mathsf{\Sigma}_k\}_{k=1}^{l}$.

We continue the proof by noting that, when $\mathbb{Y}^l \sim f_{\mathbb{Y}^l\,|\,\mathbb{X}^l}(\cdot \,|\, \mathsf{X}^l)$, the information density $\imath\big(\mathsf{X}^l; \mathbb{Y}^l\big)$ is distributed as $\sum_{k=1}^{l} \bar{S}_{k,\widetilde{m}_{\mathrm{t}}}$, with $\bar{S}_{k,\widetilde{m}_{\mathrm{t}}}$ defined in (32); and when $\mathbb{Y}^l \sim q_{\mathbb{Y}^l}$ the information density is distributed as $\sum_{k=1}^{l} T_{k,\widetilde{m}_{\mathrm{t}}}$, with $T_{k,\widetilde{m}_{\mathrm{t}}}$ defined in (31). Finally, (33) follows by minimizing over $\widetilde{m}_{\mathrm{t}} \in \{1, \ldots, m_{\mathrm{t}}\}$.

References

[1] L. Zheng and D. Tse, "Diversity and multiplexing: A fundamental tradeoff in multiple-antenna channels," *IEEE Trans. Inf. Theory*, vol. 49, no. 5, pp. 1073–1096, May 2003.

[2] A. Lozano and N. Jindal, "Transmit diversity vs. spatial multiplexing in modern MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 9, no. 1, pp. 186–197, Sep. 2010.

[3] METIS project, Deliverable D1.1, "Scenarios, requirements and KPIs for 5G mobile and wireless system," Tech. Rep., Apr. 2013.

[4] F. Boccardi, R. Heath, A. Lozano, T. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 74–80, Feb. 2014.

[5] P. Popovski, "Ultra-reliable communication in 5G wireless systems," in *IEEE Int. Conf. 5G for Ubiquitous Connectivity*, Levi, Finland, Nov. 2014.

[6] G. Fettweis and S. Alamouti, "5G: Personal mobile internet beyond what cellular did to telephony," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 140–145, Feb. 2014.

[7] E. Dahlman, G. Mildh, S. Parkvall, J. Peisa, J. Sachs, Y. Selén, and J. Sköld, "5G wireless access: Requirements and realization," *IEEE Commun. Mag.*, vol. 52, no. 12, pp. 42–47, Dec. 2014.

[8] G. Durisi, T. Koch, and P. Popovski, "Towards massive, ultra-reliable, and low-latency wireless: The art of sending short packets," Apr. 2015. [Online]. Available: http://arxiv.org/abs/1504.06526

[9] 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.

[10] 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.

[6] To be precise, the second argument of $\beta_{1-\epsilon}(\cdot, \cdot)$ in [18, Eq. (105)] is an arbitrary probability measure. In our case, since the chosen probability measure is absolutely continuous, it is convenient to let the second argument of $\beta_{1-\epsilon}(\cdot, \cdot)$ be a pdf.

[11] İ. E. Telatar, "Capacity of multi-antenna Gaussian channels," *Eur. Trans. Telecommun.*, vol. 10, pp. 585–595, Nov. 1999.

[12] B. Hassibi and B. M. Hochwald, "How much training is needed in multiple-antenna wireless links?" *IEEE Trans. Inf. Theory*, vol. 49, no. 4, pp. 951–963, Apr. 2003.

[13] L. Zheng and D. N. C. Tse, "The diversity-multiplexing tradeoff for non-coherent multiple antenna channels," in *Proc. Allerton Conf. Commun., Contr., Comput.*, Monticello, IL, U.S.A., Oct. 2002, pp. 1011–1020.

[14] ——, "Communication on the Grassmann manifold: A geometric approach to the noncoherent multiple-antenna channel," *IEEE Trans. Inf. Theory*, vol. 48, no. 2, pp. 359–383, Feb. 2002.

[15] A. Lapidoth and S. Shamai (Shitz), "Fading channels: How perfect need 'perfect side information' be?" *IEEE Trans. Inf. Theory*, vol. 48, no. 5, pp. 1118–1134, May 2002.

[16] S. M. Moser, "The fading number of multiple-input multiple-output fading channels with memory," *IEEE Trans. Inf. Theory*, vol. 55, no. 6, pp. 2716–2755, Jun. 2009.

[17] 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.

[18] 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.

[19] 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.

[20] B. M. Hochwald, T. L. Marzetta, T. J. Richardson, W. Sweldens, and R. Urbanke, "Systematic design of unitary space-time constellations," *IEEE Trans. Inf. Theory*, vol. 46, no. 6, pp. 1962–1973, Sep. 2000.

[21] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Diversity versus channel knowledge at finite block-length," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Lausanne, Switzerland, Sep. 2012, pp. 572–576.

[22] J. Östman, W. Yang, G. Durisi, and T. Koch, "Diversity versus multiplexing at finite blocklength," in *Proc. IEEE Int. Symp. Wirel. Comm. Syst. (ISWCS)*, Barcelona, Spain, Aug. 2014, pp. 702–706.

[23] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Quasi-static multiple-antenna fading channels at finite blocklength," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 4232–4265, Jul. 2014.

[24] D. N. C. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

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

[26] G. Durisi and H. Bölcskei, "High-SNR capacity of wireless communication channels in the noncoherent setting: A primer," *Int. J. Electron. Commun. (AEÜ)*, vol. 65, no. 8, pp. 707–712, Aug. 2011, invited paper.

[27] Y. Polyanskiy and S. Verdú, "Scalar coherent fading channel: Dispersion analysis," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Aug. 2011, pp. 2959–2963.

[28] S. Vituri and M. Feder, "Dispersion of infinite constellations in MIMO fading channels," in *Proc. Conv. of Electrical & Electronics Eng. Israel (IEEEI)*, Eilat, Israel, Nov. 2012.

[29] A. Collins and Y. Polyanskiy, "Orthogonal designs optimize achievable dispersion for coherent MISO channels," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Honolulu, HI, Jul. 2014.

[30] J. Wolfowitz, "The coding of messages subject to chance errors," *Illinois J. Math*, vol. 1, pp. 591–606, Dec. 1957.

[31] B. M. Hochwald and T. L. Marzetta, "Unitary space–time modulation for multiple-antenna communications in Rayleigh flat fading," *IEEE Trans. Inf. Theory*, vol. 46, no. 2, pp. 543–564, Mar. 2000.

[32] B. Hassibi and T. L. Marzetta, "Multiple-antennas and isotropically random unitary inputs: The received signal density in closed form," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1473 –1484, Jun. 2002.

[33] R. Devassy, G. Durisi, J. Östman, W. Yang, T. Eftimov, and Z. Utkovski, "Finite-SNR bounds on the sum-rate capacity of Rayleigh block-fading multiple-access channels with no a priori CSI," *IEEE Trans. Commun.*, vol. 63, no. 10, pp. 3621–3632, Oct. 2015.

[34] L. H. Ozarow, S. Shamai (Shitz), and A. D. Wyner, "Information theoretic considerations for cellular mobile radio," *IEEE Trans. Veh. Technol.*, vol. 43, no. 2, pp. 359–378, May 1994.

[35] E. Abbe, E. Telatar, and S. Huang, "Proof of the outage probability conjecture for MISO channels," *IEEE Trans. Inf. Theory*, vol. 59, no. 5, pp. 2596–2602, May 2013.

[36] E. Biglieri, J. G. Proakis, and S. Shamai (Shitz), "Fading channels: Information-theoretic and communications aspects," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2619–2692, Oct. 1998.

[37] P. Elia, K. R. Kumar, S. A. Pawar, P. V. Kumar, and H.-F. Lu, "Explicit space-time codes achieving the diversity-multiplexing gain tradeoff," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 3869–3884, Sep. 2006.

[38] L. Zheng, "Diversity-multiplexing tradeoff: A comprehensive view of multiple antenna systems," Ph.D. dissertation, University of California at Berkeley, Berkeley, CA, Nov. 2002.

[39] K. Azarian and H. El-Gamal, "The throughput–reliability tradeoff in block-fading MIMO channels," *IEEE Trans. Inf. Theory*, vol. 53, no. 2, pp. 488–501, Feb. 2007.

[40] S. Loyka and G. Levin, "Finite-SNR diversity-multiplexing tradeoff via asymptotic analysis of large MIMO systems," *IEEE Trans. Inf. Theory*, vol. 56, no. 10, pp. 4781–4792, Oct. 2010.

[41] A. M. Tulino and S. Verdú, "Random matrix theory and wireless communications," in *Foundations and Trends in Communications and Information Theory*. Delft, The Netherlands: now Publishers, 2004, vol. 1, no. 1, pp. 1–182.

[42] C. Itzykson and J. B. Zuber, "The planar approximation. II," *J. Math. Phys.*, vol. 21, pp. 411–421, 1980.

[43] A. Ghaderipoor, C. Tellambura, and A. Paulraj, "On the application of character expansions for MIMO capacity analysis," *IEEE Trans. Inf. Theory*, vol. 58, no. 5, pp. 2950–2962, May 2012.

[44] G. Alfano, C.-F. Chiasserini, A. Nordio, and S. Zhou, "Closed-form output statistics of MIMO block-fading channels," *IEEE Trans. Inf. Theory*, vol. 60, no. 12, pp. 7782–7797, Dec. 2014.

[45] S. Alamouti, "A simple transmit diversity technique for wireless communications," *IEEE J. Sel. Areas Commun.*, vol. 16, no. 8, pp. 1451–1458, Oct. 1998.

[46] V. Tarokh, H. Jafarkhani, and A. Calderbank, "Space-time block codes from orthogonal designs," *IEEE Trans. Inf. Theory*, vol. 45, no. 5, pp. 1456–1467, Jul. 1999.

[47] S. Sesia, I. Toufik, and M. Baker, Eds., *LTE–The UMTS long term evolution: From Theory to Practice*, 2nd ed. UK: Wiley, 2011.