Short-Packet Communications with Multiple Antennas: Transmit Diversity, Spatial Multiplexing, and Channel Estimation Overhead

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

Abstract

We present finite-blocklength upper and lower bounds on the maximum coding rate achievable over a multiple-antenna Rayleigh block-fading channel under the assumption that neither the transmitter nor the receiver have *a priori* knowledge of the channel realizations. Numerical evidence suggests that the bounds delimit tightly the maximum coding rate already for short blocklengths (packets of 168 channel uses). The bounds allow us to estimate the number of transmit antennas and the degree of time-frequency diversity that trade off optimally the rate gains resulting from an increase in the number of independent time-frequency-spatial diversity branches against the corresponding increase in channel estimation overhead.

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 Communitys Sevenths Framework Programme FP7/2007-2013 under Grant 333680, by the Ministerio de Economa y Competitividad of Spain under Grant TEC2013-41718-R, by Consolider-Ingenio 2010 CSD2008-00010 COMONSENS, 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 W. Yang are with the Department of Signals and Systems, Chalmers University of Technology, 41296, Gothenburg, Sweden (e-mail: {durisi,ywei}@chalmers.se).

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

J. Östman is with the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Hong Kong, China (email: eejohan@ust.hk)

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

Furthermore, a comparison with finite-blocklength maximum coding rate bounds obtained for specific space-time codes (such as the Alamouti code) allows us to determine when the available transmit antennas should be used to provide spatial diversity, and when to provide spatial multiplexing. Finally, infinite-blocklength performance metrics, such as the ergodic and the outage capacity, are shown to predict inaccurately the maximum coding rate for the short packet sizes considered in this paper. This demonstrates the necessity of accurate finite-blocklength analyses when studying wireless communications with short packet size.

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].

Current cellular systems operate typically at maximum multiplexing [2]. Indeed, diversity-exploiting techniques such as space-time codes turn out to be detrimental 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. Diversity-exploiting techniques are not advantageous for high-mobility user as well, because of the abundant time and frequency selectivity that is available. 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 cellular systems.

However, as we move towards next generation cellular systems (5G), the two assumptions of long data packets and moderately low packet error rate may cease to be valid. Indeed, 5G is expected to support a much wider range of use-case scenarios; emerging applications, such as metering, traffic safety, and telecontrol of industrial plants, may require the exchange of much shorter packets, sometimes under stringent latency and reliability constraints [3]–[5]. The question addressed in this paper is how multiple antennas should be used in this scenario. Is diversity more beneficial than multiplexing in the regime of short packet length (say 100 channel uses, roughly equal to a LTE resource block) and ultra-high reliability (packet error rate of 10^{-5} or lower)? What is the cost of

learning the fading coefficients, whose knowledge is required to exploit spatial degrees of freedom, when the packet size is short? Does this cost overcome the benefits of multiple antennas?

Contributions: The tension between reliability, throughput, and channel-estimation overhead in multiple-antenna communications have been investigated previously in the literature [6]–[11], [2], [12]. However, as we shall review in Section III, most of the available results are asymptotic either in the packet length, or in the SNR, or in both. Hence, their relevance in the context of short-packet communication systems is unclear.

In this paper, we address this issue by presenting a more refined *nonasymptotic* analysis of the tradeoff between reliability, throughput, and channel-estimation overhead, which relies on the finite-blocklength and finite-error-probability bounds on the *maximum coding rate* developed in [13], [14]. The results in this paper generalize to the multiple-antenna setting the analysis conducted in [15] for the single-input single-output (SISO) case. Our main contributions are as follows:

- Focusing on the so-called *Rayleigh block-fading* model [16], [8] and on the scenario where neither the transmitter nor the receiver have access to *a priori* channel state information (CSI), we obtain nonasymptotic achievability and converse bounds on the maximum coding rate that are explicit in the SNR, the packet size, and the packet reliability. The achievability bound is based on the dependency test (DT) bound [13, Th. 22] and uses unitary space-time modulation (USTM) [17] as input distribution in the random coding argument. This distribution is known to achieve the high-SNR ergodic capacity of Rayleigh block-fading channels in the no CSI setup when the *coherence interval* of the channel (i.e., the number of time-frequency slots over which the channel stays approximately constant) is larger than the sum of transmit and receive antennas [8], [12]. Since the lower bound involves an optimization over the number of available antennas, it can be used to estimate the optimal number of transmit antennas to be used. The converse bound is based on the meta-converse (MC) theorem [13, Th. 31], and uses the output distribution induced by USTM as auxiliary output distribution.
- We present numerical evidence that the newly derived achievability and converse bounds delimit tightly the maximum coding rate already for packet lengths of 168 channel uses. The bounds allow one to coarsely determine the optimal number of transmit antennas to be used as a function of the available time-frequency diversity, i.e., the ratio of the packet length to the channel coherence interval.
- Comparison with nonasymptotic maximum coding rate bounds, obtained for specific space-time inner codes (such as the Alamouti scheme), allows us to determine for which degree of

time-frequency diversity and for which packet-reliability level, the available transmit antennas should be used to provide spatial diversity, or spatial multiplexing, or should be partly switched off to limit the channel-estimation overhead. For example, for the specific case of a 2×2 Rayleigh block-fading channel with SNR equal to $6\,\mathrm{dB}$, packet size equal to 168 symbols and packet error rate of 10^{-3}, Alamouti is to optimal[1] within 15% when the channel coherence interval is above 42 symbols, whereas switching off one of the two transmit antennas is optimal within 15% when the channel coherence interval is below 12 symbols. For values of coherence interval between 12 and 42 symbols, transmit diversity should be combined with or replaced by spatial multiplexing.

In previous works, researchers have drawn inspiration from the structure of the capacity achieving distribution of MIMO channels to design practical coded-modulation schemes (see e.g., [18]). 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.

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 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 and $[a]^+$ stands for $\max\{a, 0\}$.

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

[1]We say that a scheme is *optimal within* 15% if it achieves a rate within 15% of the meta-converse upper bound on the maximum coding rate.

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.

Throughout the paper, we shall focus on the setting where both the transmitter and the receiver know the distribution of $\mathbb{H}_k$ but not its realization. In other words, *a priori* CSI is not available at the transmitter and at the receiver. The assumption of no *a priori* CSI at the transmitter is reasonable in a high-mobility scenario, where fast channel variations make channel tracking at the transmitter unfeasible. It is also appropriate for transmission over control channels or for time-critical applications. Indeed, in both situations it is desirable to avoid the creation of a feedback link, required to provide CSI at the transmitter. The assumption of no *a priori* CSI at the receiver allows one to characterize the information-theoretic cost of learning the channel at the receiver (for example by pilot transmission followed by channel estimation); see also [15, Sec. I]. As we shall see, this cost may be relevant when the packet size is limited and the channel is rapidly varying.

III. Maximum Coding Rate

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

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

Evidently, (3) implies the per-codeword power constraint[2]

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

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

Since the noise has unit variance, ρ in (4) can be thought 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 $(\ln M)/(l n_{\mathrm{c}})$ for which there exists a $(l, n_{\mathrm{c}}, M, \epsilon, \rho)$ code. Formally,

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

Note 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 [19]–[21].

[2]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).

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 [22], the ergodic capacity $C_{\mathrm{erg}}(\rho)$ does not depend on ϵ. Although $C_{\mathrm{erg}}(\rho)$ is not known in closed form for the case when CSI is not available *a priori* at the receiver, its high-SNR behavior is well understood [16], [17], [23], [8], [12]. Specifically, Zheng & Tse showed that [8]

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

where

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

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 [6], we see from (11) that the multiplexing (a.k.a. *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 the number of pilots per time-frequency slot needed to learn the channel at the receiver when m^* transmit antennas are used. The multiplexing penalty vanishes when n_{c} is large.

By tightening the high-SNR expansion (11) [8], [12], one obtains an accurate approximation of capacity at finite SNR [15], [24]. 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 [8]. 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 [12]. 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 need of channel estimation, and its high-SNR expansion (11) describes compactly how this penalty depends on the channel coherence interval, its asymptotic nature in the blocklength 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 [25]

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

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

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

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 [6] conjectured that the optimal diagonal matrix Q_1 is of the form

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

for some $m \in \{1, \ldots, m_{\mathrm{t}}\}$. This conjecture was proved in [26] 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 coherent interval n_{c} gets large, the cost of learning the channel at the receiver vanishes [27, p. 2632], [14]. 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

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

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

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

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

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

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 [28], 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 [7], [29]

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

The expressions in (18) and in (19) describe elegantly and succinctly the tradeoff between diversity and multiplexing. 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 version of the diversity-multiplexing tradeoff have been proposed in [30], [31]. However, these extensions rely on the outage probability and are, in contrast to the original formulation in [7], 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 diversity-multiplexing tradeoff 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 2: Assume that $n_{\mathrm{c}} \geq m_{\mathrm{t}} + m_{\mathrm{r}}$. Let $\mathbb{X} = \sqrt{\rho n_{\mathrm{c}}/\widetilde{m}_{\mathrm{t}}}\mathbb{U}$ 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. Finally, 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

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

Here, $\sigma_1 > \cdots > \sigma_{m_{\mathrm{r}}}$ denote the m_{r} nonzero singular values of Y, which are positive and distinct almost surely [32], $q = \min\{\widetilde{m}_{\mathrm{t}}, m_{\mathrm{r}}\}$, $\mu = \rho n_{\mathrm{c}}/\widetilde{m}_{\mathrm{t}}$, and

$$\psi_{\widetilde{m}_{\mathrm{t}}}(b_1, \ldots, b_{m_{\mathrm{r}}}) = \frac{\det\{\mathsf{M}\}}{\prod\limits_{i<j}^{m_{\mathrm{r}}}(b_i - b_j)} \prod_{k=1}^{m_{\mathrm{r}}} \frac{\exp\bigl(-b_k/(1+\mu)\bigr)}{b_k^{n_{\mathrm{c}}-m_{\mathrm{r}}}} \tag{21}$$

where M is a $p \times p$ real matrix ($p = \max\{\widetilde{m}_{\mathrm{t}}, m_{\mathrm{r}}\}$) whose entries are given by

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

Here,

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

denotes the regularized incomplete Gamma function.

Proof: The proof, which relies on the Itzykson-Zuber integral [33, Eq. (3.2)] and on repeated use of [34, Lem. 5], can be found, e.g., in [12, App. A] and, more recently, in [35]. ■

Remark 1: A different expression for $f_{\mathbb{Y}}(\mathsf{Y})$ is reported in [23]. The expression in Lemma 2 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 DT bound [13, Th. 22] (maximal error probability version) and makes use of the USTM-induced output distribution given in Lemma 2.

Theorem 3: 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 $\{\mathbb{Z}_k \in \mathbb{C}^{n_{\mathrm{c}} \times m_{\mathrm{r}}}\}_{k=1}^{l}$ are independent complex Gaussian 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

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

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

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

We have

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

Proof: 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 $\{\mathbb{X}_k = \sqrt{\rho n_{\mathrm{c}}/\widetilde{m}_{\mathrm{t}}}\mathbb{U}_k\}_{k=1}^{l}$ 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 $\{\mathbb{Y}_k = \sqrt{\rho n_{\mathrm{c}}/\widetilde{m}_{\mathrm{t}}}\mathbb{U}_k\mathbb{H}_k + \mathbb{W}_k\}_{k=1}^{l}$ 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 [13, Eq. (4)] can be decomposed as follows:

$$\imath\big(\mathsf{U}^l; \mathsf{Y}^l\big) = \sum_{k=1}^{l} \imath(\mathsf{U}_k; \mathsf{Y}_k) = \sum_{k=1}^{l} \ln \frac{f_{\mathbb{Y}\,|\,\mathbb{U}}(\mathsf{Y}_k \,|\, \mathsf{U}_k)}{f_{\mathbb{Y}}(\mathsf{Y}_k)} \tag{28}$$

where

$$f_{\mathbb{Y}\,|\,\mathbb{U}}(\mathsf{Y}_k \,|\, \mathsf{U}_k) = \frac{e^{-\mathrm{tr}\left\{\mathsf{Y}_k^H \left(\mathsf{I}_{n_{\mathrm{c}}} + (\rho n_{\mathrm{c}}/\widetilde{m}_{\mathrm{t}})\mathsf{U}_k\mathsf{U}_k^H\right)^{-1}\mathsf{Y}_k\right\}}}{\pi^{m_{\mathrm{r}} n_{\mathrm{c}}}(1+\rho n_{\mathrm{c}}/\widetilde{m}_{\mathrm{t}})^{\widetilde{m}_{\mathrm{t}} m_{\mathrm{r}}}}. \tag{29}$$

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

$$f_{\mathbb{Y}\,|\,\mathbb{U}}(\mathsf{Y} \,|\, \mathsf{V}^H\mathsf{U}) = f_{\mathbb{Y}\,|\,\mathbb{U}}(\mathsf{V}\mathsf{Y} \,|\, \mathsf{U}) \tag{30}$$

and

$$f_{\mathbb{Y}}(\mathsf{V}\mathsf{Y}) = f_{\mathbb{Y}}(\mathsf{Y}). \tag{31}$$

Consequently, the probability law of the information density $\imath(\mathsf{U}_k;\mathbb{Y}_k)$ in (28), (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

$$\bar{\mathsf{U}} = \begin{bmatrix} \mathsf{I}_{\widetilde{m}_{\mathrm{t}}} \\ \mathbf{0}_{n_{\mathrm{c}}-\widetilde{m}_{\mathrm{t}}\times\widetilde{m}_{\mathrm{t}}} \end{bmatrix}. \tag{32}$$

Using [13, Th. 22], we conclude that there exists a $(l, n_{\mathrm{c}}, M, \epsilon, \rho)$ code satisfying

$$\epsilon \leq \mathbb{E}\left[\exp\left\{-\left[\sum_{k=1}^{l}\imath\big(\bar{\mathsf{U}};\mathbb{Y}_k\big) - \ln(M-1)\right]^{+}\right\}\right] \tag{33}$$

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 (33) 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). ■

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 MC theorem for maximal error probability of error [13, Th. 31] and uses the output distribution induced by the USTM input distribution (see (20)) as auxiliary output distribution.

Theorem 4: 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

$$\mathbb{\Delta}_{k,\widetilde{m}_{\mathrm{t}}} = \mathrm{diag}\{\Delta_{k,\widetilde{m}_{\mathrm{t}},1},\ldots,\Delta_{k,\widetilde{m}_{\mathrm{t}},m_{\mathrm{r}}}\}. \tag{34}$$

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

$$\widetilde{\mathbf{\Sigma}}_k = \begin{bmatrix} \mathsf{I}_{m_{\mathrm{t}}} + \mathbf{\Sigma}_k & \mathbf{0} \\ \mathbf{0} & \mathsf{I}_{n_{\mathrm{c}}-m_{\mathrm{t}}} \end{bmatrix}. \tag{35}$$

Let $\{\mathbb{U}_k \in \mathbb{C}^{n_{\mathrm{c}}\times m_{\mathrm{r}}}\}_{k=1}^{l}$ be i.i.d. isotropically distributed (truncated) unitary matrices, and let $\{\bar{\mathbb{Z}}_k \in \mathbb{C}^{n_{\mathrm{c}}\times m_{\mathrm{r}}}\}_{k=1}^{l}$ be independent complex Gaussian matrices with i.i.d. $\mathcal{CN}(0,1)$ entries. Finally, let

$$\begin{aligned}\bar{c}_{\widetilde{m}_{\mathrm{t}}}(\mathbf{\Sigma}_k) &= \widetilde{m}_{\mathrm{t}}(n_{\mathrm{c}}-\widetilde{m}_{\mathrm{t}})\ln\frac{\rho n_{\mathrm{c}}}{\widetilde{m}_{\mathrm{t}}} - \widetilde{m}_{\mathrm{t}}(n_{\mathrm{c}}-\widetilde{m}_{\mathrm{t}}-m_{\mathrm{r}})\ln\left(1+\frac{\rho n_{\mathrm{c}}}{\widetilde{m}_{\mathrm{t}}}\right)\\ &\quad - m_{\mathrm{r}}\ln\det\widetilde{\mathbf{\Sigma}}_k - \sum_{u=n_{\mathrm{c}}-p+1}^{n_{\mathrm{c}}}\ln\Gamma(u) + \sum_{u=1}^{\widetilde{m}_{\mathrm{t}}}\ln\Gamma(u)\end{aligned} \tag{36}$$

$$T_{k,\widetilde{m}_{\mathrm{t}}}(\mathsf{\Sigma}_k) = \bar{c}_{\widetilde{m}_{\mathrm{t}}}(\mathsf{\Sigma}_k) - \mathrm{tr}\{\mathbb{U}_k \mathbb{\Delta}_{k,\widetilde{m}_{\mathrm{t}}} \mathbb{U}_k^H \widetilde{\mathsf{\Sigma}}_k^{-1})\} - \ln \psi_{\widetilde{m}_{\mathrm{t}}}(\Delta_{k,\widetilde{m}_{\mathrm{t}},1}, \ldots, \Delta_{k,\widetilde{m}_{\mathrm{t}},m_{\mathrm{r}}}) \tag{37}$$

and

$$\bar{S}_{k,\widetilde{m}_{\mathrm{t}}}(\mathsf{\Sigma}_k) = \bar{c}_{\widetilde{m}_{\mathrm{t}}}(\mathsf{\Sigma}_k) - \mathrm{tr}\{\bar{\mathbb{Z}}_k^H \bar{\mathbb{Z}}_k\} - \ln \psi_{\widetilde{m}_{\mathrm{t}}}(\bar{\Lambda}_{k,\widetilde{m}_{\mathrm{t}},1}, \ldots, \bar{\Lambda}_{k,\widetilde{m}_{\mathrm{t}},m_{\mathrm{r}}}) \tag{38}$$

where $\bar{\Lambda}_{k,\widetilde{m}_{\mathrm{t}},1} > \cdots > \bar{\Lambda}_{k,\widetilde{m}_{\mathrm{t}},m_{\mathrm{r}}}$ are the ordered eigenvalues of $\mathbb{Z}_k^H \widetilde{\mathsf{\Sigma}}_k \mathbb{Z}_k$ (which are positive and distinct almost surely), and where $p = \max\{\widetilde{m}_{\mathrm{t}}, m_{\mathrm{r}}\}$, 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

$$R^*(l, n_{\mathrm{c}}, \epsilon, \rho) \leq \min_{1 \leq \widetilde{m}_{\mathrm{t}} \leq m_{\mathrm{t}}} \sup_{\{\mathsf{\Sigma}_k\}_{k=1}^l} \frac{1}{n} \ln \frac{1}{\Pr\left\{\sum_{k=1}^l T_{k,\widetilde{m}_{\mathrm{t}}}(\mathsf{\Sigma}_k) \geq \gamma\right\}} \tag{39}$$

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

$$\Pr\left\{\sum_{k=1}^l \bar{S}_{k,\widetilde{m}_{\mathrm{t}}}(\mathsf{\Sigma}_k) \leq \gamma\right\} = \epsilon. \tag{40}$$

Remark 2: To facilitate its numerical evaluation, the meta-converse upper bound (39) can be relaxed by using [13, Eq. (102)], which yields

$$R^*(l, n_{\mathrm{c}}, \epsilon, \rho) \leq \min_{1 \leq \widetilde{m}_{\mathrm{t}} \leq m_{\mathrm{t}}} \sup_{\{\mathsf{\Sigma}_k\}_{k=1}^l} \inf_{\lambda} \frac{1}{n} \left[\lambda - \ln\left(\left[\Pr\left\{\sum_{k=1}^l \bar{S}_{k,\widetilde{m}_{\mathrm{t}}}(\mathsf{\Sigma}_k) \leq \lambda\right\} - \epsilon\right]^+\right)\right]. \tag{41}$$

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

Remark 3: As for the outage capacity in (13), the symmetry in (41) suggests that the supremum over $\{\mathsf{\Sigma}_k\}_{k=1}^l$, is achieved when

$$\mathsf{\Sigma}_k = \frac{\rho}{m_k} \mathrm{diag}\{\underbrace{1, \ldots, 1}_{m_k}, \underbrace{0, \ldots, 0}_{m_{\mathrm{t}} - m_k}\} \tag{42}$$

for some $m_k \in \{1, \ldots, m_{\mathrm{t}}\}$, $k \in \{1, \ldots, l\}$. We can think of (42) as a finite-blocklength extension of Telatar conjecture. Although far from conclusive, the numerical results reported in Section VII support the validity of this conjecture.

Remark 4: 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

$$\mathrm{tr}\big\{\mathsf{C}_m^H \mathsf{C}_m\big\} \leq l n_{\mathrm{c}} \rho \tag{43}$$

can be obtained by evaluating the supremum in (39) and (41) over all $\{\mathsf{\Sigma}_k\}_{k=1}^{l}$ that satisfy

$$\sum_{k=1}^{l} \mathrm{tr}\{\mathsf{\Sigma}_k\} \leq l n_{\mathrm{c}} \rho \tag{44}$$

Proof: Fix $1 \leq \widetilde{m}_{\mathrm{t}} \leq m_{\mathrm{t}}$. To bound $R^*(l, n_{\mathrm{c}}, \epsilon, \rho)$, we use the meta-converse theorem for maximal error probability [13, Th. 31] with auxiliary pdf

$$q_{\mathbb{Y}^l}(\mathsf{Y}^l) = \prod_{k=1}^{l} f_{\mathbb{Y}}(\mathsf{Y}_k) \tag{45}$$

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

$$R^*(l, n_{\mathrm{c}}, \epsilon, \rho) \leq \sup_{\mathsf{X}^l} \frac{1}{n} \ln \frac{1}{\beta_{1-\epsilon}\big(\mathsf{X}^l, q_{\mathbb{Y}^l}\big)} \tag{46}$$

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 [13, Eq. (105)].[3] Using the Neyman-Pearson lemma, we have that

$$\beta_{1-\epsilon}(\mathsf{X}^l, q_{\mathbb{Y}^l}) = \Pr\big\{\imath\big(\mathsf{X}^l; \mathbb{Y}^l\big) \geq \gamma\big\}, \quad \mathbb{Y}^l \sim q_{\mathbb{Y}^l} \tag{47}$$

where γ is the solution of

$$\Pr\big\{\imath\big(\mathsf{X}^l; \mathbb{Y}^l\big) \leq \gamma\big\} = \epsilon, \quad \mathbb{Y}^l \sim f_{\mathbb{Y}^l \,|\, \mathbb{X}^l}(\cdot \,|\, \mathsf{X}^l) \tag{48}$$

and where $\imath(\cdot;\cdot)$ is defined as in (28).

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

$$\mathsf{X}_k \mathsf{X}_k^H = \mathsf{V}_k \mathsf{\Sigma}_k \mathsf{V}_k^H, \quad k = 1, \ldots, l. \tag{49}$$

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 (30) and (31) 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 (46) 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 (38); 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 (37). Finally, (39) follows by minimizing over $\widetilde{m}_{\mathrm{t}} \in \{1, \ldots, m_{\mathrm{t}}\}$. ■

[3]To be precise, the second argument of $\beta_{1-\epsilon}(\cdot,\cdot)$ in [13, 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.

VI. How to Use the Available Transmit Antennas?

Insights on how the available transmit antennas should be used to maximize $R^*(l, n_{\mathrm{c}}, \epsilon, \rho)$ can be obtained by considering a setup where an outer code defined along the same lines as Definition 1 is combined with a specific 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 3 and 4. In Sections VI-A and VI-B, we provide achievability and converse bound obtained by assuming that a diversity-exploiting orthogonal space-time inner code is used. We shall focus on the 2×2 and 4×4 MIMO configurations.

A. Alamouti

For the 2×2 case, we consider an Alamouti inner space-time code [36]. When CSI is available at the receiver, this scheme provides a diversity gain of 4 (per coherence interval) but no multiplexing gain [37, Sec. 9.1.5].

In order to analyze the finite blocklength performance of such a scheme when CSI is not *a priori* available at the receiver, we proceed similarly to 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

$$\mathsf{X}_k = [\boldsymbol{a}_k \quad e(\boldsymbol{a}_k)] \tag{50}$$

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 [36]:

$$[\boldsymbol{b}]_{2l-1} = [e(\boldsymbol{a})]_{2l-1} = [\boldsymbol{a}]^*_{2l}, \quad l = 1, 2, \ldots, n_{\mathrm{c}}/2 \tag{51a}$$

$$[\boldsymbol{b}]_{2l} = [e(\boldsymbol{a})]_{2l} = -[\boldsymbol{a}]^*_{2l-1}, \quad l = 1, 2, \ldots, n_{\mathrm{c}}/2. \tag{51b}$$

In Lemma 5 below, we provide the pdf of the channel output $\mathbb{Y}$ induced by an input matrix $\mathbb{X}$ constructed as in (50), 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 5: Assume that $m_{\mathrm{t}} = m_{\mathrm{r}} = 2$ and that n_{c} is even and larger or equal to 4. Let

$$\mathbb{X} = [\boldsymbol{A} \quad e(\boldsymbol{A})] \tag{52}$$

where $e(\cdot)$ is defined in (51) and where $\boldsymbol{A} = \sqrt{\rho n_{\rm c}/2}\,\boldsymbol{U}$, with $\boldsymbol{U}$ being an isotropically distributed unit-norm $n_{\rm c}$-dimensional complex random vector. Let $\mathbb{Y} = \begin{bmatrix}\boldsymbol{Y}_1 & \boldsymbol{Y}_2\end{bmatrix} = \mathbb{X}\mathbb{H}+\mathbb{W}$, where $\mathbb{H} \in \mathbb{C}^{2\times 2}$ and $\mathbb{W} \in \mathbb{C}^{n_{\rm c}\times 2}$ are defined similarly as in (1). Furthermore, let

$$\hat{\mathbb{Y}} = [\boldsymbol{Y}_1 \quad e(\boldsymbol{Y}_1) \quad \boldsymbol{Y}_2 \quad e(\boldsymbol{Y}_2)] \tag{53}$$

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

$$f_{\mathbb{Y}}(\mathsf{Y}) = \frac{\exp\big(\mathrm{tr}\big\{\mathsf{Y}^H\mathsf{Y}\big\}\big)}{\pi^{2n_{\rm c}}(1+\rho n_{\rm c}/2)^{2n_{\rm c}}}\frac{\Gamma(n_{\rm c})}{(\sigma_1-\sigma_3)^4}\det\{\mathsf{M}(\sigma_1,\sigma_3)\} \tag{54}$$

where the 4×4 matrix M is given by

$$\mathsf{M} = \begin{bmatrix} e^{\sigma_1}\widetilde{\gamma}(n_{\rm c}-5,\sigma_1) & (n_{\rm c}-2)\sigma_1^{n_{\rm c}-3} & (n_{\rm c}-3)\sigma_1^{n_{\rm c}-4} & (n_{\rm c}-4)\sigma_1^{n_{\rm c}-5} \\ e^{\sigma_1}\widetilde{\gamma}(n_{\rm c}-4,\sigma_1) & \sigma_1^{n_{\rm c}-2} & \sigma_1^{n_{\rm c}-3} & \sigma_1^{n_{\rm c}-4} \\ e^{\sigma_3}\widetilde{\gamma}(n_{\rm c}-5,\sigma_3) & (n_{\rm c}-2)\sigma_3^{n_{\rm c}-3} & (n_{\rm c}-3)\sigma_3^{n_{\rm c}-4} & (n_{\rm c}-4)\sigma_3^{n_{\rm c}-5} \\ e^{\sigma_3}\widetilde{\gamma}(n_{\rm c}-4,\sigma_3) & \sigma_3^{n_{\rm c}-2} & \sigma_3^{n_{\rm c}-3} & \sigma_3^{n_{\rm c}-4} \end{bmatrix} \tag{55}$$

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

$$\mathsf{M} = \begin{bmatrix} e^{\sigma_1} & 2\sigma_1 & 1 & 0 \\ e^{\sigma_1} & \sigma_1^2 & \sigma_1 & 1 \\ e^{\sigma_3} & 2\sigma_3 & 1 & 0 \\ e^{\sigma_3} & \sigma_3^2 & \sigma_3 & 1 \end{bmatrix} \tag{56}$$

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

Remark 5: Note that although $\boldsymbol{A}$ in (52) is isotropically distributed, the matrix $\mathbb{X}$ is not. Hence, $\mathbb{X}$ does not follow a USTM distribution.

Proof: The proof follows along the same lines as the proof of Lemma 2. ■

Treating the Alamouti space-time inner code as part of the channel, we next report lower and upper bounds on the maximum coding rate $R^*_{\rm ala}(l, n_{\rm c}, \epsilon, \rho)$ achievable when an Alamouti space-time inner code is used. These bounds rely on the closed-form expression for $f_{\mathbb{Y}}(\cdot)$ reported in (54).

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

1) DT lower bound: We provide first an achievability bound, which is based on the DT bound [13, Th. 22].

Theorem 6: Let $\{\mathbb{Z}_k \in \mathbb{C}^{n_\mathrm{c}\times 2}\}_{k=1}^{l}$ be independent complex Gaussian matrices with i.i.d. $\mathcal{CN}(0,1)$ entries. Let

$$\mathsf{D} = \mathrm{diag}\left\{1+\frac{\rho n_\mathrm{c}}{2},\, 1+\frac{\rho n_\mathrm{c}}{2},\, \underbrace{1,\,\ldots,\,1}_{n_\mathrm{c}-2}\right\} \tag{57}$$

$\mathbb{V}_k = \begin{bmatrix}\boldsymbol{V}_{k,1} & \boldsymbol{V}_{k,2}\end{bmatrix} = \mathsf{D}^{1/2}\mathbb{Z}_k$, and

$$\hat{\mathbb{V}}_k = \begin{bmatrix}\boldsymbol{V}_{k,1} & e(\boldsymbol{V}_{k,1}) & \boldsymbol{V}_{k,2} & e(\boldsymbol{V}_{k,2})\end{bmatrix} \tag{58}$$

where the function $e(\cdot)$ was defined in (51). 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

$$S_k = \mathrm{tr}\{\mathbb{Z}_k^H\mathsf{D}\mathbb{Z}_k\} - \mathrm{tr}\{\mathbb{Z}_k^H\mathbb{Z}\} - \ln\Gamma(n_\mathrm{c}) + \ln\det\{\mathsf{M}(\Sigma_{k,1},\Sigma_{k,3})\} - 4\ln(\Sigma_{k,1}-\Sigma_{k,3}) \tag{59}$$

and let

$$\epsilon_\mathrm{ala}(M) = \mathbb{E}\left[\exp\left\{-\left[\sum_{k=1}^{l} S_k - \ln(M-1)\right]^+\right\}\right]. \tag{60}$$

Then

$$R^*_\mathrm{ala}(l,n_\mathrm{c},\epsilon,\rho) \geq \max\left\{\frac{\ln M}{n_\mathrm{c} l}\ :\ \epsilon_\mathrm{ala}(M)\leq\epsilon\right\}. \tag{61}$$

Proof: The proof follows along the same lines as the proof of Theorem 3. ■

2) MC upper bound: Using [13, Th. 22] and [13, Eq. (102)] we obtain the following converse bound on $R^*_\mathrm{ala}(l,n_\mathrm{c},\epsilon,\rho)$.

Theorem 7: Let S_k be defined as in (59). Then

$$R^*_\mathrm{ala}(l,n_\mathrm{c},\epsilon,\rho) \leq \inf_\lambda \frac{1}{n}\left[\lambda - \ln\left(\Pr\left\{\sum_{k=1}^{l} S_k \leq \lambda\right\} - \epsilon\right)\right]. \tag{62}$$

Proof: The proof follows along the same lines as the proof of Theorem 4. ■

B. Frequency Switched Transmit Diversity

Since no generalization of the Alamouti space-time inner code exists beyond the 2×2 configuration [38], we consider instead for the 4×4 case the combination of Alamouti and frequency-switched transmit diversity (FSTD) used in LTE [39, 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

$$\begin{bmatrix} a_1 & a_2 & 0 & 0 \\ 0 & 0 & b_1 & b_2 \\ -a_2^* & a_1^* & 0 & 0 \\ 0 & 0 & -b_2^* & b_1^* \end{bmatrix} \tag{63}$$

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 $n_{\mathrm{c}}/2$. Hence, when CSI is available at the receiver, Alamouti combined with FSTD provides a diversity gain of 8 (per coherence interval) and no multiplexing gain. The finite-blocklength performance of this scheme can be analyzed using a similar approach as for the 2×2 case.

VII. Numerical Results

We consider the same setup as in [2], which is based on the 3GPP LTE standard [39]. Specifically, the packet length is set to $n = 168$ symbols (14 OFDM symbols each consisting of 12 tones, which corresponds to an LTE resource block). In typical LTE scenarios, by spreading the tones uniformly over the available bandwidth of 20 MHz, one can obtain 12 frequency diversity branches [2]. In our analysis, we shall also consider the case when $l > 12$. 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].

Control signaling: In Fig. 1 we plot[5] the DT lower bound (27) and the MC upper bound (41) for the case $m_{\mathrm{t}} = m_{\mathrm{r}} = 2$ and $\epsilon = 10^{-3}$. 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 [15] for the single-antenna case. For $n_{\mathrm{c}} < n_{\mathrm{c}}^*$, the cost of estimating the channel overcomes the gain due to time-frequency diversity. For $n_{\mathrm{c}} > n_{\mathrm{c}}^*$, the bottleneck is the limited time-frequency diversity offered by the channel.

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

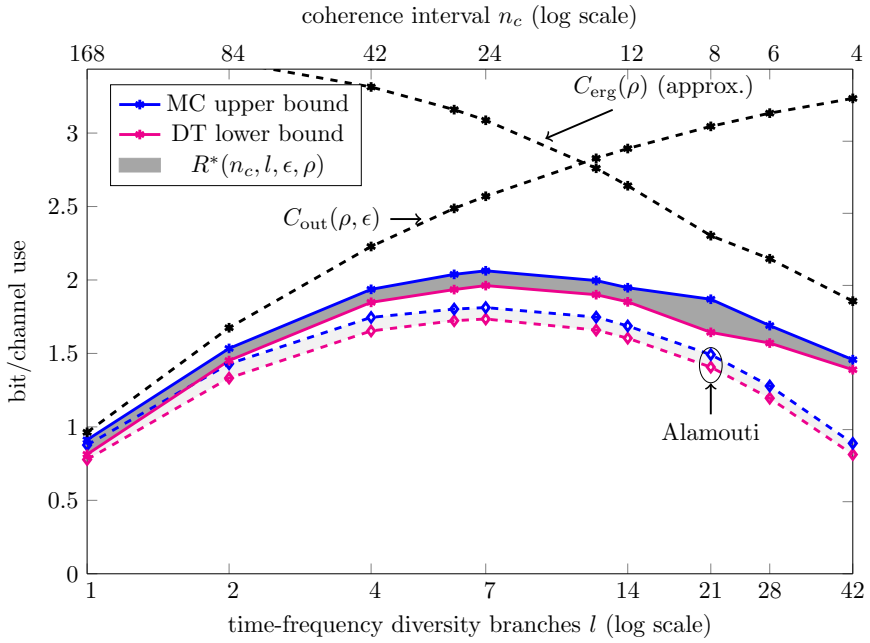

Fig. 1. $m_{\text{t}} = m_{\text{r}} = 2$, $n = 168$, $\epsilon = 10^{-3}$. In the MC upper bound (41) the supremum over $\{\mathbf{\Sigma}_k\}_{k=1}^{l}$ is restricted to $\{\mathbf{\Sigma}_k\}_{k=1}^{l}$ of the form given in (42) when $l > 7$, because of computational complexity.

For the parameters considered in Fig. 1, the optimal coherence interval length is $n_{\text{c}}^* \approx 24$, which corresponds to about 7 time-frequency diversity branches.

In the figure, we also plot the outage capacity $C_{\text{out}}(\rho, \epsilon)$ in (13) as a function of the number of time-frequency diversity branches $l = n/n_{\text{c}}$ (with $n = 168$), and a lower bound on the ergodic capacity $C_{\text{erg}}(\rho)$, as a function of the coherence interval n_{c}. This lower bound on $C_{\text{erg}}(\rho)$, 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 $C_{\text{erg}}(\rho)$ accurately already at moderate SNR values [24].

As shown in the figure, $C_{\text{out}}(\epsilon, \rho)$ provides a good approximation for $R^*(l, n_{\text{c}}, \epsilon, \rho)$ only when l is small ($n_{\text{c}} \approx n$), i.e., when the fading channel is essentially constant over the duration of the packet (quasi-static scenario). Furthermore, $C_{\text{out}}(\epsilon, \rho)$ fails to capture the loss in throughput due to the channel estimation overhead, which is relevant for small n_{c}. The number of active transmit antennas $\widetilde{m}_{\text{t}}$ that maximizes the DT achievability bound is $\widetilde{m}_{\text{t}} = 2$ (both antennas active) for $1 \leq l \leq 21$, whereas it is $\widetilde{m}_{\text{t}} = 1$ (only one antenna active) for $l > 21$. The lower bound on $C_{\text{erg}}(\rho)$ plotted in the figure, which also involves a maximization over the number of active antennas, exhibits a similar behavior (see Table I), although its gap from $R^*(l, n_{\text{c}}, \epsilon, \rho)$ is large, especially for small

l values. 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 (41) is again $\widetilde{m}_{\mathrm{t}}=2$ for $1\leq l\leq 21$ and $\widetilde{m}_{\mathrm{t}}=1$ for $l>21$. Furthermore, the optimal $\{\mathbf{\Sigma}_k\}_{k=1}^{l}$ take all the same value and are equal to a 2×2 scaled identity matrix for $1\leq l\leq 14$ 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 outage capacity, for $R^*(l,n_{\mathrm{c}},\epsilon,\rho)$ 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 both diversity and multiplexing gains.

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 noncoherent diversity-multiplexing tradeoff (19) do predict that transmit antennas must be 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 $n_{\mathrm{c}}\leq 3$.

Note that 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 in the MC bound). For example, one could drop the assumption that the input 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 to extend the optimization in both (33) and (41) over fractional values of $\widetilde{m}_{\mathrm{t}}$.

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

TABLE I

ESTIMATES ON THE OPTIMAL NUMBER OF TRANSMIT ANTENNAS $\tilde{m}_t$ FOR A 2×2 MIMO SYSTEM; $\rho = 6\,\mathrm{dB}$, $\epsilon = 10^{-3}$, $n = 168$.

Bound	Optimal $\tilde{m}_t$	Diversity branches l
DT	1	$21 < l \leq 48$
	2	$1 \leq l \leq 21$
MC	1	$21 < l \leq 48$
	2	$1 \leq l \leq 21$
USTM bound	1	$21 \leq l \leq 48$
on C_{erg}	2	$1 \leq l < 21$

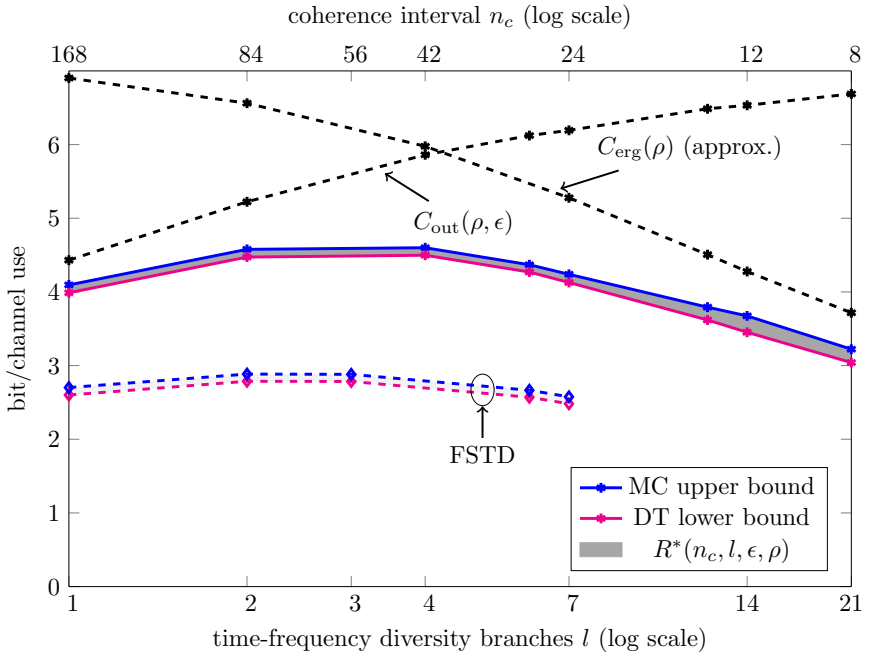

Fig. 2. $m_{\mathrm{t}} = m_{\mathrm{r}} = 4$, $n = 168$, $\epsilon = 10^{-3}$. In the MC upper bound (41), the supremum over $\{\mathbf{\Sigma}_k\}_{k=1}^{l}$ is restricted to $\{\mathbf{\Sigma}_k\}_{k=1}^{l}$ of the form given in (42), because of computational complexity.

characterization of $R^*(l, n_{\mathrm{c}}, \epsilon, \rho)$. 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 diversity gain is suboptimal even when the amount of time-frequency diversity is limited (i.e., l is small). As for 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 ($\widetilde{m}_{\mathrm{t}} = 4$) when $1 \leq l < 12$, by using 3 antennas when $l = 12$, and by

TABLE II

ESTIMATES ON THE OPTIMAL NUMBER OF TRANSMIT ANTENNAS $\tilde{m}_t$ FOR A 4×4 MIMO SYSTEM. $\rho = 6\,\mathrm{dB}$, $\epsilon = 10^{-3}$, $n = 168$.

Bound	Optimal $\tilde{m}_t$	Diversity branches l
DT	2	$12 < l \leq 21$
	3	$l = 12$
	4	$1 \leq l < 12$
MC	2	$14 \leq l \leq 21$
	3	$7 \leq l < 14$
	4	$1 \leq l < 7$
USTM bound on C_{erg}	2	$14 \leq l \leq 21$
	3	$7 \leq l < 14$
	4	$1 \leq l < 7$

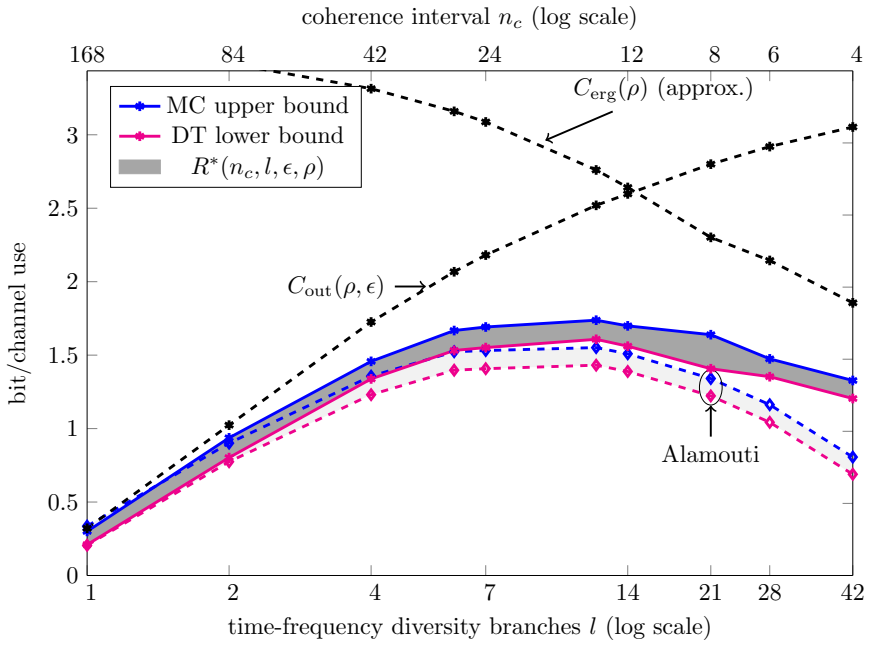

Fig. 3. $m_{\mathrm{t}} = m_{\mathrm{r}} = 2$, $n = 168$, $\epsilon = 10^{-5}$. In the MC upper bound (41) the supremum over $\{\mathbf{\Sigma}_k\}_{k=1}^{l}$ is restricted to $\{\mathbf{\Sigma}_k\}_{k=1}^{l}$ of the form given in (42) when $l > 7$, because of computational complexity.

using only two antennas when $12 < l \leq 21$. Also in this case, the lower bound on $C_{\mathrm{erg}}(\rho)$ and the MC upper bound exhibit a similar behavior (see Table).

Ultra-reliable communication: In Fig. 3 and Fig. 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

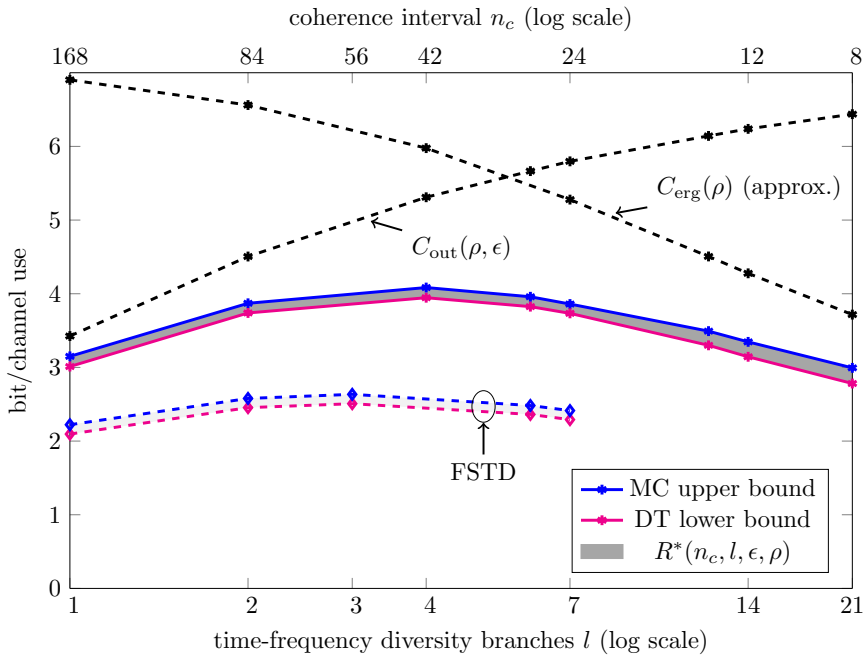

Fig. 4. $m_{\text{t}} = m_{\text{r}} = 4$, $n = 168$, $\epsilon = 10^{-5}$. In the MC upper bound (41), the supremum over $\{\mathbf{\Sigma}_k\}_{k=1}^{l}$ is performed only over $\{\mathbf{\Sigma}_k\}_{k=1}^{l}$ values of the form given in (42), because of computational complexity.

optimal schemes and the diversity-based schemes (Alamouti for the 2×2 configuration, and FSTD for the 4×4 case) gets smaller. This comes as no surprise, since the higher reliability requirement makes the exploitation of transmit diversity advantageous.

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, which are explicit in the packet error rate ϵ, the coherence interval n_{c}, and the number of time-frequency diversity branches l, 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 trade optimally the rate gain resulting from the exploitation of the available time-frequency-spatial resources against the rate loss due to 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.

For the short-packet size applications that are in focus in this paper, traditional infinite-blocklength performance metrics such as the outage and the ergodic capacity, are shown to provide inaccurate

estimates on the maximum coding rate, and also to fail in capturing the fundamental tradeoff between diversity/multiplexing on the one hand and channel estimation overhead on the other hand, which occurs in the short-packet size regime. Hence, our results suggest that the optimal design of the novel low-latency/ultra-reliable services that will be provided by next-generation (5G) 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.

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," Oct. 2014. [Online]. Available: http://arxiv.org/abs/1410.4330

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

[7] L. Zheng and D. 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.

[8] L. Zheng and D. N. C. Tse, "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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] ——, "Diversity versus channel knowledge at finite block-length," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Lausanne, Switzerland, Sep. 2012, pp. 572–576.

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

[21] 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.

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

[23] 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.

[24] 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," Dec. 2014. [Online]. Available: ArXiv

[25] 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.

[26] 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.

[27] 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.

[28] 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.

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

[30] 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.

[31] 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.

[32] 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.

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

[34] 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.

[35] G. Alfano, C.-F. Chiasserini, A. Nordio, and S. Zhou, "Closed-form output statistics of MIMO block-fading channels," *IEEE Trans. Inf. Theory*, 2014, to appear.

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

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

[38] 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.

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