URLLC with Massive MIMO:
Analysis and Design at Finite Blocklength

Johan Östman, *Student Member, IEEE*, Alejandro Lancho, *Member, IEEE*, Giuseppe Durisi, *Senior Member, IEEE*, and Luca Sanguinetti, *Senior Member, IEEE*

Abstract—The fast adoption of Massive MIMO for high-throughput communications was enabled by many research contributions mostly relying on infinite-blocklength information-theoretic bounds. This makes it hard to assess the suitability of Massive MIMO for ultra-reliable low-latency communications (URLLC) operating with short-blocklength codes. This paper provides a rigorous framework for the characterization and numerical evaluation (using the saddlepoint approximation) of the error probability achievable in the uplink and downlink of Massive MIMO at finite blocklength. The framework encompasses imperfect channel state information, pilot contamination, spatially correlated channels, and arbitrary linear spatial processing. In line with previous results based on infinite-blocklength bounds, we prove that, with minimum mean-square error (MMSE) processing and spatially correlated channels, the error probability at finite blocklength goes to zero as the number M of antennas grows to infinity, even under pilot contamination. However, numerical results for a practical URLLC network setup involving a base station with $M = 100$ antennas, show that a target error probability of 10^{-5} can be achieved with MMSE processing, uniformly over each cell, only if orthogonal pilot sequences are assigned to all the users in the network. Maximum ratio processing does not suffice.

Index Terms—Massive MIMO, ultra-reliable low-latency communications, finite blocklength information theory, saddlepoint approximation, outage probability, pilot contamination, MR and MMSE processing, asymptotic analysis.

I. INTRODUCTION

Among the new use cases that will be supported by next generation wireless systems [2], some of the most challenging ones fall into the category of ultra-reliable low-latency communications (URLLC). For example, in URLLC for factory automation [3], small payloads on the order of 100 bits must be delivered within hundreds of microseconds and with a reliability no smaller than 99.999%. To achieve such a high reliability, it is crucial to exploit diversity. Unfortunately, the stringent latency requirements prevent the exploitation of diversity in time. Furthermore, the use of frequency diversity is problematic, especially in the uplink where current standardization rules do not allow user equipments (UEs) to spread a packet over independently fading frequency resources. Thus, the spatial diversity offered by multiple antennas becomes critical to achieve the desired reliability. The latest instantiation of multiple antenna technologies is the so-called Massive MIMO (multiple-input multiple-output), which refers to a wireless network where base stations (BS), equipped with a very large number M of antennas, serve a multitude of UEs via linear spatial signal processing [4]. Thanks to the intense research performed since its inception in 2010, the advantages of Massive MIMO in terms of spectral efficiency [5], [6], energy efficiency [7], and power control [8] are well understood, and its key ingredients have made it into the 5G standard [9]. However, all these results have mainly been established in the *ergodic regime*, where the propagation channel evolves according to a block-fading model, and each codeword spans an increasingly large number of independent fading realizations as the codeword length goes to infinity (infinite-blocklength regime). Since these assumptions are highly questionable in URLLC scenarios [10], it remains unclear whether the design guidelines that have been obtained so far for Massive MIMO (see [11], [12] for a detailed review on the topic) apply to URLLC deployments.

A. Prior Art

Unlike the vast majority of literature on Massive MIMO, which focuses on the aforementioned ergodic regime, the authors in [13], [14] assume that the fading channel stays constant during the transmission of a codeword (the so-called quasi-static fading scenario) and use outage capacity [15] as asymptotic performance metric. Although the quasi-static fading scenario is relevant for URLLC, the infinite blocklength assumption may yield incorrect estimates of the error probability. The use of outage capacity in the context of URLLC is often justified by the results reported in [16], where it is proved that short channel codes operate close to the outage capacity for quasi-static fading channels. More specifically, the authors of [16] proved that the difference between the outage capacity and the maximum coding rate, achievable at finite blocklength over quasi-static fading channels, goes to zero much faster than the difference between the capacity and the maximum coding rate achievable over additive white Gaussian noise (AWGN) channels. The intuition is that the dominant sources of errors in quasi-static fading channels are deep-fade events, which cannot be alleviated through the use of channel codes, since channel coding provides protection only against additive noise.

The application of this result to Massive MIMO is problematic since, as M grows, we start observing channel hardening and the underlying effective channel (after precoding/combining) becomes more similar to an AWGN channel.

As a consequence, finite-blocklength effects become more pronounced, since additive noise turns into the dominating impairment. Another unsatisfactory feature of the outage-capacity framework is its inability to account for the channel state information (CSI) acquisition overhead, caused by the transmission of pilot sequences. Indeed, quasi-static fading channels can be learnt perfectly at the receiver in the asymptotic limit of large blocklength with no rate penalty: it is enough to let the number of pilot symbols grow sublinearly with the blocklength. The attempts made so far to include channel-estimation overhead in the outage setup [13], [14] are not convincing from a theoretical perspective. A theoretically satisfying framework must include the use of a mismatch receiver that treats the channel estimate, obtained using a fixed number of pilot symbols, as perfect. One difficulty is that a fundamental result commonly used in the ergodic case to bound the mutual information, by treating the channel estimation error as noise (see, e.g., [17, Lemma B.0.1]), does not apply to the outage case. This is because, in the outage setup, the fading channel stays constant over the entire codeword, and one is interested in computing an outage event over fading realizations. This means that both the channel and its estimate must be treated as deterministic quantities when computing bounds on the instantaneous spectral efficiency.

The limitation of both ergodic and outage setups can be overcome by performing a nonasymptotic analysis of the error probability based on the finite-blocklength information-theoretic bounds introduced in [18] and extended to fading channels in [16], [19], [20]. This approach has been pursued recently in [21], [22]. However, the analysis in these papers relies on the so called *normal approximation* [18, Eq. (291)], whose tightness for the range of error probabilities of interest in URLLC is questionable. Also, the use of the normal approximation for the case of imperfect CSI in both [21], [22] is not convincing, since the approximation does not depend on the instantaneous channel estimation error, but only on its variance. This is not compatible with a scenario in which the channel stays constant over the duration of each codeword.

B. Contributions

To verify if the design guidelines developed for Massive MIMO in the context of non-delay limited, large-throughput, communication links apply also to the URLLC setup, we present a rigorous nonasymptotic characterization of the error probability achievable in Massive MIMO. Specifically, we provide a firm upper bound on the error probability, which is obtained by adapting the random-coding union bound with parameter s (RCUs) introduced in [23] to the case of Massive MIMO communications. The resulting bound applies to Gaussian codebooks, and holds for any linear processing scheme and any pilot-based channel estimation scheme. Since the bound is in terms of integrals that are not known in closed form and need to be evaluated numerically, which is impractical when the targeted error probability is low, we also present an accurate and easy-to-compute approximation, based on the saddlepoint method [24, Ch. XVI].

We then use the bound to evaluate the error probability in the uplink (UL) and downlink (DL) of a Massive MIMO network, with imperfect channel state information, pilot contamination, and spatially correlated channels. Both minimum mean-square error (MMSE) and maximum ratio (MR) processing are considered. We remark that the application of the RCUs bound and saddlepoint approximation to characterize the error probability in this scenario is novel. Furthermore, differently from [25], the proposed saddlepoint approximation involves quantities that can be characterized in closed form. Hence, it can be evaluated efficiently. We prove that the average error probability at finite blocklength with MMSE tends to zero as $M \to \infty$, whereas it converges to a positive number when MR is used. These results are similar in flavor to those about Massive MIMO ergodic rates in the infinite-blocklength regime (see, e.g., [6] and [26]).

Through numerical experiments, we estimate the error probability achievable for finite values of M and quantify the impact of spatial correlation and pilot contamination. Inspired by [27], we use the *network availability* as performance metric, which we define as the fraction of UE placements for which the per-link error probability, averaged over the small-scale fading and the additive noise, is below a given target. In the asymptotic outage setting, this quantity is obtained by characterizing the metadistribution of the signal-to-interference ratio (SIR) [27]. At finite blocklength, the network availability turns out to be related to the metadistribution of the so called *generalized information density* [23, Eq. (3)].

The numerical experiments show that, for finite values of M, it is important to take into account spatial correlation to obtain realistic estimates of the error probability. Furthermore, pilot contamination turns out to have a strong impact on performance. Consider for example a network with four $75\,\text{m} \times 75\,\text{m}$ cells, $K = 10$ UEs, $M = 100$ BS antennas. Furthermore, assume a transmit power of $10\,\text{dBm}$ in UL and DL, an error probability target of 10^{-5} and a fixed frame of 300 symbols, which accommodates pilots and data transmission in UL and DL. Assume also that in each data transmission phase, 160 information bits need to be conveyed with an error probability target of 10^{-5}. For this scenario, a network availability above 90% can be achieved with MMSE processing in UL and DL only if pilot contamination is avoided by allocating as many pilot symbols as the total number of UEs in the network. In contrast, when all cells use the same pilot sequences, a network availability just above 50% is achieved despite the fact that the shorter duration of the pilot sequences allows for a larger number of channel uses in the data phase. With MR processing, the network availability remains below 50% for both UL and DL, even when pilot contamination is avoided. These numerical results suggest the following guidelines for the design of Massive MIMO for URLLC applications: i) Pilot contamination must be avoided; ii) In line with [26], MMSE should be chosen in place of the simpler MR.

C. Paper Outline and Notation

In Section II, we present the finite-blocklength framework that will be used to analyze and design Massive MIMO networks. In Section III, the finite-blocklength framework is used to analyze the impact on the error probability of

pilot contamination, spatial correlation, and of the number of BS antennas, by focusing on a single-cell network with two UEs. The analysis is extended to a general multicell multiuser setting in Section IV. Some conclusions are drawn in Section V.

Lower-case bold letters are used for vectors and upper-case bold letters are used for matrices. The circularly-symmetric Gaussian distribution is denoted by $\mathcal{CN}(0, \sigma^2)$, where σ^2 denotes the variance. We use $\mathbb{E}[\cdot]$ to indicate the expectation operator, and $\mathbb{P}[\cdot]$ for the probability of a set. The natural logarithm is denoted by $\log(\cdot)$, and $Q(\cdot)$ stands for the Gaussian Q-function. The Frobenius and spectral norms of a matrix $\mathbf{X}$ are denoted by $\|\mathbf{X}\|_F$ and $\|\mathbf{X}\|_2$, respectively. The operators $(\cdot)^\mathrm{T}$, $(\cdot)^*$, and $(\cdot)^\mathrm{H}$ denote transpose, complex conjugate, and Hermitian transpose, respectively. Finally, we use $\overset{d}{=}$ to denote equality in distribution while, for two random sequences a_n, b_n, we write $a_n \asymp b_n$ to indicate that $\lim_{n \to \infty}(a_n - b_n) = 0$ almost surely.

D. Reproducible Research

The Matlab code used to obtain the simulation results is available at: https://github.com/infotheorychalmers/URLLC_Massive_MIMO.

II. A FINITE-BLOCKLENGTH UPPER-BOUND ON THE ERROR PROBABILITY

In this section, we present a finite-blocklength upper bound on the error probability and describe an efficient method for its numerical evaluation, based on the saddlepoint approximation [24, Ch. XVI]. We start by considering the simple case in which the received signal is the superposition of a scaled version of the desired signal and additive Gaussian noise. This simple channel model constitutes the building block for the analysis of the error probability achievable in the Massive MIMO networks considered in Sections III and IV.

A. Upper Bound for Deterministic and Random Channels

Consider a discrete AWGN channel given by

$$v[k] = gq[k] + z[k], \quad k = 1, \ldots, n \qquad (1)$$

where $q[k] \in \mathbb{C}$ and $v[k] \in \mathbb{C}$ are the input and output over channel use k, respectively, and n is the codeword length. Furthermore, $g \in \mathbb{C}$ is the channel gain, which is assumed to remain constant during transmission of the n-length codeword. The additive noise variables $\{z[k] \in \mathbb{C}; k = 1, \ldots, n\}$, are independent and identically distributed (i.i.d.), $\mathcal{CN}(0, \sigma^2)$, random variables. In what follows, we assume that:

1) The receiver *does not know* the channel gain g but has an estimate $\widehat{g}$ of g that is treated as perfect.
2) To determine the transmitted codeword $\mathbf{q} = [q[1], \ldots, q[n]]^\mathrm{T}$, the receiver seeks the codeword $\widetilde{\mathbf{q}}$ from the codebook $\mathcal{C}$ that, once scaled by $\widehat{g}$, is the closest to the received vector $\mathbf{v} = [v[1], \ldots, v[n]]^\mathrm{T} \in \mathbb{C}^n$ in Euclidean distance. Mathematically, the estimated codeword $\widehat{\mathbf{q}}$ is obtained as

$$\widehat{\mathbf{q}} = \underset{\widetilde{\mathbf{q}} \in \mathcal{C}}{\arg\min} \|\mathbf{v} - \widehat{g}\widetilde{\mathbf{q}}\|^2. \qquad (2)$$

A receiver operating according to (2) is known as mismatched scaled nearest-neighbor (SNN) decoder [17]. Note that it coincides with the optimal maximum likelihood decoder *if and only if* $\widehat{g} = g$.

We are interested in deriving an upper bound on the error probability $\epsilon = \mathbb{P}[\widehat{\mathbf{q}} \neq \mathbf{q}]$ achieved by the SNN decoding rule (2). To do so, we follow a standard practice in information theory and use a random-coding approach [28]. Specifically, we consider a Gaussian random code ensemble, where the elements of each codeword are drawn independently from a $\mathcal{CN}(0, \rho)$ distribution.[1] Here, ρ can be thought of as the average transmit power. We consider the cases where the channel gain g in (1) can be modelled as a deterministic or a random variable. In the literature, this latter case is commonly referred to as quasi-static fading setting [30, p. 2631].

Theorem 1: Assume that $g \in \mathbb{C}$ and $\widehat{g} \in \mathbb{C}$ in (1) are deterministic. There exists a coding scheme with $m = 2^b$ codewords of length n operating according to the mismatched SNN decoding rule (2), whose error probability ϵ is upper-bounded by[2]

$$\epsilon = \mathbb{P}[\widehat{\mathbf{q}} \neq \mathbf{q}]$$
$$\leq \mathbb{P}\left[\sum_{k=1}^n \imath_s(q[k], v[k]) + \log(u) \leq \log(m-1)\right] \qquad (3)$$

for all $s > 0$. Here, u is a random variable that is uniformly distributed over the interval $[0, 1]$ and $\imath_s(q[k], v[k])$ is the generalized information density, given by

$$\imath_s(q[k], v[k]) = -s\left|v[k] - \widehat{g}q[k]\right|^2$$
$$+ \frac{s|v[k]|^2}{1 + s\rho|\widehat{g}|^2} + \log\left(1 + s\rho|\widehat{g}|^2\right). \qquad (4)$$

Assume now that $g \in \mathbb{C}$ and $\widehat{g} \in \mathbb{C}$ in (1) are random variables drawn according to an arbitrary joint distribution. Then, for all $s > 0$, the error probability ϵ is upper-bounded by

$$\epsilon = \mathbb{P}[\widehat{\mathbf{q}} \neq \mathbf{q}]$$
$$\leq \mathbb{E}_{g,\widehat{g}}\left[\mathbb{P}\left[\sum_{k=1}^n \imath_s(q[k], v[k]) \leq \log\frac{m-1}{u}\bigg| g, \widehat{g}\right]\right] \qquad (5)$$

where the average is taken over the joint distribution of g and $\widehat{g}$. If $g \in \mathbb{C}$ is a random variable and $\widehat{g} \in \mathbb{C}$ is deterministic,[3] the average in (5) is only taken over the distribution of g.

Proof: The proof for the case of g and $\widehat{g}$ being deterministic, which is given in Appendix A for completeness, follows by particularizing the RCUs bound introduced in [23, Th. 1] to the considered setup. The upper bound for random g and $\widehat{g}$ readily follows by taking an expectation over the joint distribution of g and $\widehat{g}$. ∎

Coarsely speaking, Theorem 1 shows that the error probability in the finite-blocklength regime can be characterized

---

[1]Note that this ensemble is not optimal at finite blocklength, not even if $\widehat{g} = g$. However, it is commonly used to obtain tractable expressions and insights into the performance of communication systems [11], [12], [29]. Our analysis can be extended to other ensembles—see, e.g., [20].

[2]Note that the probability in (3) is computed with respect to the channel inputs $\{q[k]\}_{k=1}^n$, the additive noise $\{z[k]\}_{k=1}^n$, and the random variable u.

[3]This case will turn out important to analyze the DL of Massive MIMO networks.

in terms of the probability that the empirical average of the generalized information density $\imath_s$ is smaller than the chosen rate $R = (\log m)/n$. In contrast, in the infinite-blocklength regime, the error (outage) probability, is given by the probability that the so-called generalized mutual information [17, Sec. III] $I_s = \mathbb{E}[\imath_s(q[1], v[1])]$ is below the chosen rate. If g is known at the receiver, i.e., $\widehat{g} = g$, it follows immediately from the decoding rule (2) that $\epsilon \to 0$ when the SNR grows unboundedly, i.e., $\rho/\sigma^2 \to \infty$. The following lemma shows that this is also true for the upper bounds (3) and (5).

Lemma 1: If $g = \widehat{g}$, then

$$\lim_{\rho/\sigma^2 \to \infty} \mathbb{P}\left[\sum_{k=1}^{n} \imath_s(q[k], v[k]) \leq \log \frac{m-1}{u}\right] = 0. \quad (6)$$

Proof: This result is easily established by setting $v[k] = gq[k]$ and $\widehat{g} = g$ in (4) and by noting that one can make (4) arbitrarily large by choosing s sufficiently large. ∎

We anticipate that Lemma 1 will be important for the characterization of the error probability of Massive MIMO in the asymptotic limit of large antenna arrays, i.e., $M \to \infty$.

The upper bounds in (3) and (5) involve the evaluation of a tail probability, which is not known in closed form and needs to be evaluated numerically. Furthermore, they can be tightened by performing an optimization over the parameter $s > 0$, which also needs to be performed numerically. All this is computational demanding, especially when one targets the low error probabilities required in URLLC applications. In the next section, we discuss how this problem can be alleviated by using a saddlepoint approximation.

B. Saddlepoint Approximation

One possible way to numerically approximate (3) and (5) is to perform a normal approximation on the probability term based on the Berry-Esseen central limit theorem [24, Ch. XVI.5]. This leads to the following expansion:

$$\mathbb{P}\left[\sum_{k=1}^{n} \imath_s(q[k], v[k]) \leq \log \frac{m-1}{u}\right]$$
$$= Q\left(\frac{nI_s - \log(m-1)}{\sqrt{nV_s}}\right) + o\left(\frac{1}{\sqrt{n}}\right) \quad (7)$$

where $I_s = \mathbb{E}[\imath_s(q[1], v[1])]$ is the so-called generalized mutual information [17, Sec. III],

$$V_s = \mathbb{E}\left[|\imath_s(q[1], v[1]) - I_s|^2\right] \quad (8)$$

is the variance of the information density, typically referred to as *channel dispersion* [18, Sec. IV], and $o(1/\sqrt{n})$ accounts for terms that decay faster than $1/\sqrt{n}$ as $n \to \infty$. The so-called *normal approximation* obtained by neglecting the $o(1/\sqrt{n})$ term in (7) is accurate only when $R = (\log m)/n$ is close to I_s [25]. Unfortunately, this is typically not the case in URLLC since one needs to operate at rates much lower than I_s to obtain the required low error probabilities at SNR values of practical interest (see, e.g., [25, Fig. 3]). A more accurate approximation, that holds for all values of R, can be obtained using the saddlepoint method. The

main idea of the saddlepoint method is to perform an exponential tilting [24, Ch. XVI.7] on the random variables $\{\imath_s(q[k], v[k]), k = 1, \ldots, n\}$, which moves their mean close to the desired rate R. This guarantees that a subsequent use of the normal approximation yields small errors.

The saddlepoint method has been applied to obtain accurate approximations of the RCUs in, e.g., [31] and [25]. In the following, we particularize these expressions to the setup considered in Theorem 1 and refer to [25], [31] for further details and proofs. While to obtain (7), it is sufficient to check that the third central moment of $\imath_s(q[k], v[k])$ is bounded (which is indeed the case in our setup), the existence of a saddlepoint approximation requires the more stringent condition that the third derivative of the moment-generating function (MGF) of $-\imath_s(q[k], v[k])$ exists in a neighborhood of zero. Specifically, we require that there exist two values $\underline{\zeta} < 0 < \overline{\zeta}$ such that

$$\sup_{\underline{\zeta} < \zeta < \overline{\zeta}} \frac{d^3}{d\zeta^3}\left|\mathbb{E}\left[e^{-\zeta \imath_s(q[k], v[k])}\right]\right| < \infty. \quad (9)$$

As shown in Appendix B, this condition is verified in our setup. Specifically, we have that

$$\underline{\zeta} = -\frac{\sqrt{(\beta_B - \beta_A)^2 + 4\beta_A\beta_B(1 - \nu)} + \beta_A - \beta_B}{2\beta_A\beta_B(1 - \nu)} \quad (10)$$

$$\overline{\zeta} = \frac{\sqrt{(\beta_B - \beta_A)^2 + 4\beta_A\beta_B(1 - \nu)} - \beta_A + \beta_B}{2\beta_A\beta_B(1 - \nu)} \quad (11)$$

where

$$\beta_A = s(\rho|g - \widehat{g}|^2 + \sigma^2) \quad (12)$$

$$\beta_B = \frac{s}{1 + s\rho|\widehat{g}|^2}\left(\rho|g|^2 + \sigma^2\right) \quad (13)$$

$$\nu = \frac{s^2\left|\rho|g|^2 + \sigma^2 - g^*\widehat{g}\rho\right|^2}{\beta_A\beta_B(1 + s\rho|\widehat{g}|^2)}. \quad (14)$$

The saddlepoint approximation that will be provided in Theorem 2 below depends on the cumulant-generating function (CGF) of $-\imath_s(q[k], v[k])$

$$\kappa(\zeta) = \log \mathbb{E}\left[e^{-\zeta \imath_s(q[k], v[k])}\right] \quad (15)$$

and on its first derivative $\kappa'(\zeta)$ and second derivative $\kappa''(\zeta)$. In our setup, these quantities can be computed in closed form for all $\zeta \in (\underline{\zeta}, \overline{\zeta})$ and are given by (see Appendix B)

$$\kappa(\zeta) = -\zeta \log\left(1 + s\rho|\widehat{g}|^2\right) - \log\left(1 + (\beta_B - \beta_A)\zeta - \beta_A\beta_B(1 - \nu)\zeta^2\right) \quad (16)$$

$$\kappa'(\zeta) = -\log\left(1 + s\rho|\widehat{g}|^2\right) - \frac{(\beta_B - \beta_A) - 2\beta_A\beta_B(1 - \nu)\zeta}{1 + (\beta_B - \beta_A)\zeta - \beta_A\beta_B(1 - \nu)\zeta^2} \quad (17)$$

$$\kappa''(\zeta) = \left[\frac{(\beta_B - \beta_A) - 2\beta_A\beta_B(1 - \nu)\zeta}{1 + (\beta_B - \beta_A)\zeta - \beta_A\beta_B(1 - \nu)\zeta^2}\right]^2 + \frac{2\beta_A\beta_B(1 - \nu)}{1 + (\beta_B - \beta_A)\zeta - \beta_A\beta_B(1 - \nu)\zeta^2}. \quad (18)$$

Note that $-\kappa(\zeta)$ coincides with the so-called Gallager's E_0 function for the mismatched case [23, Eq. (22)]. As a conse-

quence, we have that $I_s = -\kappa'(0)$. Furthermore, the so-called *critical rate* R_s^{cr} (see [28, Eq. (5.6.30)]) is given by

$$R_s^{\mathrm{cr}} = -\kappa'(1). \tag{19}$$

We are now ready to present the saddlepoint expansion of the RCUs bound (3).

Theorem 2: Let $m = e^{nR}$ for some $R > 0$, and let $\zeta \in (\underline{\zeta}, \overline{\zeta})$ be the solution to the equation $R = -\kappa'(\zeta)$.[4] If $\zeta \in [0, 1]$, then $R_s^{\mathrm{cr}} \leq R \leq I_s$ and

$$\mathbb{P}\left[\sum_{k=1}^{n} \imath_s(q[k], v[k]) \leq \log \frac{e^{nR} - 1}{u}\right]$$
$$= e^{n[\kappa(\zeta) + \zeta R]}\left[\Psi_{n,\zeta}(\zeta) + \Psi_{n,\zeta}(1 - \zeta) + o\left(\frac{1}{\sqrt{n}}\right)\right] \tag{20}$$

where

$$\Psi_{n,\zeta}(u) \triangleq e^{n\frac{u^2}{2}\kappa''(\zeta)} Q\left(u\sqrt{n\kappa''(\zeta)}\right) \tag{21}$$

and $o(1/\sqrt{n})$ comprises terms that vanish faster than $1/\sqrt{n}$ and are uniform in ζ.

If $\zeta > 1$, then $R < R_s^{\mathrm{cr}}$ and

$$\mathbb{P}\left[\sum_{k=1}^{n} \imath_s(q[k], v[k]) \leq \log \frac{e^{nR} - 1}{u}\right]$$
$$= e^{n[\kappa(1) + R]}\left[\widetilde{\Psi}_n(1, 1) + \widetilde{\Psi}_n(0, -1) + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)\right] \tag{22}$$

where

$$\widetilde{\Psi}_n(a_1, a_2) = e^{na_1\left[R_s^{\mathrm{cr}} - R + \frac{\kappa''(1)}{2}\right]}$$
$$\times Q\left(a_1\sqrt{n\kappa''(1)} + a_2\frac{n(R_s^{\mathrm{cr}} - R)}{\sqrt{n\kappa''(1)}}\right) \tag{23}$$

and $\mathcal{O}(1/\sqrt{n})$ comprises terms that are of order $1/\sqrt{n}$ and are uniform in ζ. If $\zeta < 0$, then $R > I_s$ and

$$\mathbb{P}\left[\sum_{k=1}^{n} \imath_s(q[k], v[k]) \leq \log \frac{e^{nR} - 1}{u}\right]$$
$$= 1 - e^{n[\kappa(\zeta) + \zeta R]}\left[\Psi_{n,\zeta}(-\zeta) - \Psi_{n,\zeta}(1 - \zeta)\right.$$
$$\left. + o\left(\frac{1}{\sqrt{n}}\right)\right]. \tag{24}$$

Proof: The proof follows along steps similar to [31, App. E] and to [25, App. I], and it is thus omitted because of space limitations. ∎

We will refer to the approximations obtained by ignoring the $o(1/\sqrt{n})$ terms and the $\mathcal{O}(1/\sqrt{n})$ terms in (20), (22), and (24) as *saddlepoint approximations*. Note that the exponential term on the right-hand side of (20) and (22) corresponds to the Gallager error exponent for the mismatch decoding scenario [32]. This means that the saddlepoint approximation provides an estimate of the subexponential factor, thereby allowing one to obtain accurate approximations of error probability values for which the error exponent is inaccurate. In a nutshell, the key idea of the saddlepoint method is to isolate the Gallager

[4]The existence of such a solution for all rates $R \geq 0$ follows from (17).

error-exponent term, i.e., the exponential term in (20), (22), and (24), which governs the exponential decay of the error probability as a function of the blocklength, and then to use the Berry-Esseen central-limit theorem to characterize only the pre-exponential factor, i.e., the factor that multiplies the exponential term. It is also worth highlighting that since all quantities in (20), (22), and (24) are known is closed form, the evaluation of the saddlepoint approximation, for a given ζ and its corresponding rate $R = -\kappa'(\zeta)$, entails a complexity similar to that of the normal approximation (7).

Note that both the saddlepoint approximation and the normal approximation can be tightened by performing an optimization over s, which may be time consuming. One way to avoid this step is to choose an s that is optimal in some asymptotic regime. One can for example set s so as to maximize the generalized mutual information I_s. The corresponding value for s can be obtained in closed form [17, Eq. (64)].

C. Outage Probability and Normal Approximation

Equipped with the bound (5) and with an efficient method for the numerical evaluation of the probability term within (5), we can now evaluate the error probability achievable for short blocklengths and investigate whether the outage probability is an accurate performance metric in Massive MIMO systems for URLLC applications. For the sake of simplicity, we consider a single-UE multiantenna system in which the BS has a large number M of antennas. We denote by $\mathbf{h} \in \mathbb{C}^M$ the channel between the UE and the BS array and assume that it can be modelled as uncorrelated Rayleigh fading $\mathbf{h} \sim \mathcal{CN}(\mathbf{0}_M, \beta\mathbf{I}_M)$ where β is the large-scale fading gain [11, Sec. 1.3.2]. If perfect CSI is available at the receiver and MR combining is used for detection, the UL channel input-output relation can be expressed as

$$v[k] = \frac{\mathbf{h}^{\mathrm{H}}}{\|\mathbf{h}\|}\mathbf{h}q[k] + \frac{\mathbf{h}^{\mathrm{H}}}{\|\mathbf{h}\|}\mathbf{z}'[k], \quad k = 1, \dots, n \tag{25}$$

where $\mathbf{z}'[k] \sim \mathcal{CN}(\mathbf{0}_M, \sigma^2\mathbf{I}_M)$ is the thermal noise over the antenna array over channel use k. Note that (25) can be mapped into (1) by setting $g = \frac{\mathbf{h}^{\mathrm{H}}}{\|\mathbf{h}\|}\mathbf{h} = \|\mathbf{h}\|$ and $z[k] = \frac{\mathbf{h}^{\mathrm{H}}}{\|\mathbf{h}\|}\mathbf{z}'[k] \sim \mathcal{CN}(0, \sigma^2)$. Since $\mathbf{h}$ is perfectly known at the receiver, we have that $\widehat{g} = g = \|\mathbf{h}\|$. In the limit $n \to \infty$, it can be shown that the probability term in (5), once optimized over the parameter s, is equal to 1 if $\log(1 + \rho|g|^2/\sigma^2) < R$ and 0 otherwise. This means that the bound in (5) converges to the outage probability

$$\mathbb{P}\left[\log\left(1 + \frac{\rho g^2}{\sigma^2}\right) < R\right]. \tag{26}$$

Here, the probability is evaluated with respect to the random variable $g = \|\mathbf{h}\|$.

In Fig. 1, we depict the outage probability in (26) as a function of the number of BS antennas M. Comparisons are made with the upper bound in (5), evaluated by means of both Monte-Carlo integration (exact) and the saddlepoint approximation in Theorem 2. We also depict the normal approximation obtained by averaging (7) over g. In the evaluation
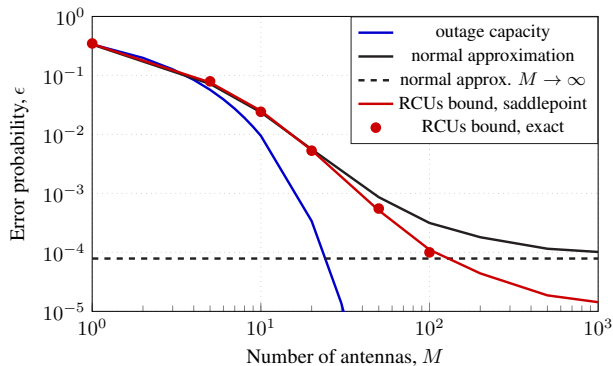
(a) Fixed average received SNR = 1 dB.

(b) Fixed transmit power $\rho = -24\,\mathrm{dBm}$.

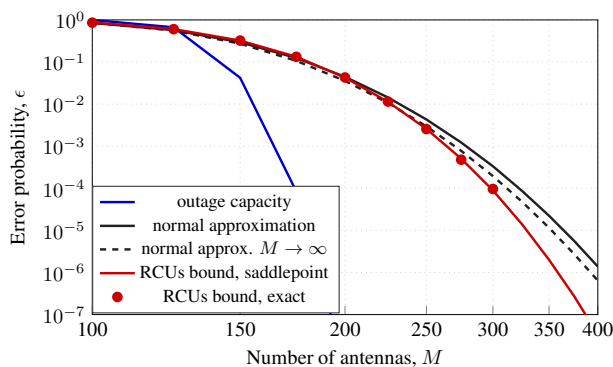Fig. 1: Average error probability in the UL of a single-UE multiantenna system when $\widehat{g} = g = \|\mathbf{h}\|$ with $\mathbf{h} \sim \mathcal{CN}(\mathbf{0}_M, \beta\mathbf{I}_M)$, $n = 100$, and $R = 0.6$ bits per channel use. The UE is assumed to be at a distance from the BS that results in $\beta/\sigma^2 = 1$.

of (5), we set $\widehat{g} = g$ and optimize over the parameter s by means of a bisection search.[5] We assume that $\sigma^2 = -94\,\mathrm{dBm}$ and set $\beta = \sigma^2$ so that $\mathbb{E}\big[g^2\big]/\sigma^2 = \beta M/\sigma^2 = M$.[6] Furthermore, we consider a codeword length $n = 100$ and a rate of $R = 60/100 = 0.6$ bits per channel use.

In Fig. 1a, we illustrate the error probability for a transmit power ρ that decreases as $1/M$. Specifically, we set $\rho = \widetilde{\rho}/M$ with $\widetilde{\rho} = 1\,\mathrm{dB}$. Since $g^2/M \to \beta$ as $M \to \infty$ and we assume $\beta = \sigma^2$, it thus follows that the instantaneous SNR $\rho g^2/\sigma^2$ converges to the deterministic value $\widetilde{\rho}$ as $M \to \infty$. This means that, as $M \to \infty$, the normal approximation for i.i.d. Gaussian inputs given in (7) for a fixed g converges to a deterministic quantity. Specifically, in the limit $M \to \infty$, we have that $I_s = \log(1 + \widetilde{\rho}\beta/\sigma^2)$ (achieved for $s = 1/\sigma^2$) and $V_s = 2\widetilde{\rho}\beta/(\widetilde{\rho}\beta + \sigma^2)$ [29, Eq. (2.55)]. The resulting approximation is of interest because it does not require any Monte-Carlo averaging over the realizations of the fading channel. From Fig. 1a, we see that the outage probability (26) approximates well the exact RCUs bound (5) only when M is small, i.e., $M < 5$, whereas the normal approximation loses accuracy

when $M > 20$. Both approximations are not accurate at the low error probabilities of interest in URLLC. The saddlepoint approximation is instead very accurate for all M values.

In Fig. 1b we report the error probability with no power scaling so that the average received SNR increases as M increases. Specifically, we consider a fixed transmit power $\rho = -24\,\mathrm{dBm}$. Hence, for $M = 320$ the average received SNR in Fig. 1b equals $1\,\mathrm{dB}$, which coincides with the average received SNR in Fig. 1a. With no power scaling, the outage probability (26) is an accurate approximation for the RCUs bound (5) only for very large values of the error probability, whereas the accuracy of the normal approximation (7) is acceptable for ϵ within the range $[10^{-3}, 1]$. The saddlepoint approximation again is on top of the RCUs bound for all values of error probability considered in the figure.

Based on the above results, we conclude that outage probability and the normal approximation do not always provide accurate estimates of the error probability achievable in large-antenna systems with short-packet communications over quasi-static channels. The accuracy of these approximations becomes even more questionable in the presence of imperfect CSI. This problem can be avoided altogether by using the nonasymptotic bound (5) in Theorem 1, which can be efficiently evaluated by means of the saddlepoint approximation in Theorem 2. In the next two sections, we will show how the simple input-output relation (1) can be used as building block for the analysis of practical Massive MIMO networks with imperfect CSI, pilot contamination, spatial correlation among antennas, and both inter-cell and intra-cell interference. Theorem 1 and Theorem 2 will then be used to efficiently evaluate the average error probability also in these more realistic scenarios.

III. A TWO-UE SINGLE-CELL MASSIVE MIMO SCENARIO

We consider a single-cell network where the BS is equipped with M antennas and serves $K = 2$ single-antenna UEs. We denote by $\mathbf{h}_i \in \mathbb{C}^M$ the channel vector between the BS and UE i for $i = 1, 2$. We use a correlated Rayleigh fading model where $\mathbf{h}_i \sim \mathcal{CN}(\mathbf{0}_M, \mathbf{R}_i)$ remains constant for the duration of a codeword transmission. The normalized trace $\beta_i = \mathrm{tr}(\mathbf{R}_i)/M$ determines the average large-scale fading between UE i and the BS, while the eigenstructure of $\mathbf{R}_i$ describes its spatial channel correlation [11, Sec. 2.2]. We assume that $\mathbf{R}_1$ and $\mathbf{R}_2$ are known at the BS; see, e.g., [26], [33] for a description of practical estimation methods. This setup is sufficient to demonstrate the usefulness of the framework developed in Section II for the analysis and design of Massive MIMO networks. A more general setup will be considered in Section IV.

A. Uplink pilot transmission

We consider the standard time-division duplex (TDD) Massive MIMO protocol, where the UL and DL transmissions are assigned n channel uses in total, divided in n_{p} channel uses for UL pilots, n_{ul} channel uses for UL data, and $n_{\mathrm{dl}} = n - n_{\mathrm{p}} - n_{\mathrm{ul}}$ channel uses for DL data. We assume that the n_{p}-length pilot sequence $\boldsymbol{\phi}_i \in \mathbb{C}^{n_{\mathrm{p}}}$ with $\boldsymbol{\phi}_i^{\mathrm{H}}\boldsymbol{\phi}_i = n_{\mathrm{p}}$ is used by UE i for

---

[5]In all numerical simulations presented throughout the paper, we will always evaluate the error probability bound in (5) using the saddlepoint approximation in Theorem 2, and optimize it over the parameter s via a bisection search.

[6]With the distance-dependent pathloss model that will be introduced in (39), this corresponds to a distance of 36.4 m.

channel estimation. The elements of ϕ_i are scaled by the square-root of the pilot power $\sqrt{\rho^{\mathrm{ul}}}$ and transmitted over n_{p} channel uses. When the UEs transmit their pilot sequences, the received pilot signal $\mathbf{Y}^{\mathrm{pilot}} \in \mathbb{C}^{M \times n_{\mathrm{p}}}$ is

$$\mathbf{Y}^{\mathrm{pilot}} = \sqrt{\rho^{\mathrm{ul}}}\mathbf{h}_1\phi_1^{\mathsf{H}} + \sqrt{\rho^{\mathrm{ul}}}\mathbf{h}_2\phi_2^{\mathsf{H}} + \mathbf{Z}^{\mathrm{pilot}} \qquad (27)$$

where $\mathbf{Z}^{\mathrm{pilot}} \in \mathbb{C}^{M \times n_{\mathrm{p}}}$ is the additive noise with i.i.d. elements distributed as $\mathcal{CN}(0, \sigma_{\mathrm{ul}}^2)$. Assuming that $\mathbf{R}_1$ and $\mathbf{R}_2$ are known at the BS, the MMSE estimate of $\mathbf{h}_i$ is [11, Sec. 3.2]

$$\widehat{\mathbf{h}}_i = \sqrt{\rho^{\mathrm{ul}} n_{\mathrm{p}}}\mathbf{R}_i\mathbf{Q}_i^{-1}\left(\mathbf{Y}^{\mathrm{pilot}}\phi_i\right) \qquad (28)$$

for $i = 1, 2$ with

$$\mathbf{Q}_i = \rho^{\mathrm{ul}}\mathbf{R}_1\phi_1^{\mathsf{H}}\phi_i + \rho^{\mathrm{ul}}\mathbf{R}_2\phi_2^{\mathsf{H}}\phi_i + \sigma_{\mathrm{ul}}^2\mathbf{I}_M. \qquad (29)$$

The MMSE estimate $\widehat{\mathbf{h}}_i$ and the estimation error $\widetilde{\mathbf{h}}_i = \mathbf{h}_i - \widehat{\mathbf{h}}_i$ are independent random vectors, distributed as $\widehat{\mathbf{h}}_i \sim \mathcal{CN}(\mathbf{0}, \boldsymbol{\Phi}_i)$ and $\widetilde{\mathbf{h}}_i \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}_i - \boldsymbol{\Phi}_i)$, respectively, with $\boldsymbol{\Phi}_i = \rho^{\mathrm{ul}} n_{\mathrm{p}}\mathbf{R}_i\mathbf{Q}_i^{-1}\mathbf{R}_i$.

It follows from (29) that if the two UEs use orthogonal pilot sequences, i.e., $\phi_1^{\mathsf{H}}\phi_2 = 0$, they do not interfere, whereas they interfere if they use the same pilot sequence, i.e. $\phi_1 = \phi_2$. This interference is known as pilot contamination and has two main consequences in the channel estimation process [11, Sec. 3.2.2]. The first is a reduced estimation quality; the second is that the estimates $\widehat{\mathbf{h}}_1$ and $\widehat{\mathbf{h}}_2$ become correlated. To see this, observe that if $\phi_1 = \phi_2$ then $\mathbf{Y}^{\mathrm{pilot}}\phi_1 = \mathbf{Y}^{\mathrm{pilot}}\phi_2$ and $\mathbf{Q}_1 = \mathbf{Q}_2 = \mathbf{Q}$ with $\mathbf{Q} = \rho^{\mathrm{ul}} n_{\mathrm{p}}\mathbf{R}_1 + \rho^{\mathrm{ul}} n_{\mathrm{p}}\mathbf{R}_2 + \sigma_{\mathrm{ul}}^2\mathbf{I}_M$. Hence, $\widehat{\mathbf{h}}_2$ can be written as $\widehat{\mathbf{h}}_2 = \mathbf{R}_2(\mathbf{R}_1)^{-1}\widehat{\mathbf{h}}_1$ provided that $\mathbf{R}_1$ is invertible. This implies that the two estimates are correlated with cross-correlation matrix given by $\mathbb{E}\left[\widehat{\mathbf{h}}_1\widehat{\mathbf{h}}_2^{\mathsf{H}}\right] = \boldsymbol{\Upsilon}_{12} = \rho^{\mathrm{ul}} n_{\mathrm{p}}\mathbf{R}_1\mathbf{Q}^{-1}\mathbf{R}_2$. This holds even though the underlying channels $\mathbf{h}_1$ and $\mathbf{h}_2$ are statistically independent, which implies that $\mathbb{E}[\mathbf{h}_1\mathbf{h}_2^{\mathsf{H}}] = \mathbf{0}_M$. Observe that if there is no spatial correlation, i.e., $\mathbf{R}_i = \beta_i\mathbf{I}_M$, $i = 1, 2$, then the channel estimates are identical up to a scaling factor, i.e., they are linearly dependent. We will return to the issue of pilot contamination in Section III-E.

B. Uplink data transmission

During UL data transmission, the received complex baseband signal $\mathbf{r}^{\mathrm{ul}}[k] \in \mathbb{C}^M$ over an arbitrary channel use k, where $k = 1, \dots, n_{\mathrm{ul}}$, is given by

$$\mathbf{r}^{\mathrm{ul}}[k] = \mathbf{h}_1 x_1^{\mathrm{ul}}[k] + \mathbf{h}_2 x_2^{\mathrm{ul}}[k] + \mathbf{z}^{\mathrm{ul}}[k] \qquad (30)$$

where $x_i^{\mathrm{ul}}[k] \sim \mathcal{CN}(0, \rho^{\mathrm{ul}})$ is the information bearing signal[7] transmitted by UE i with ρ^{ul} being the average UL transmit power and $\mathbf{z}^{\mathrm{ul}}[k] \sim \mathcal{CN}(\mathbf{0}, \sigma_{\mathrm{ul}}^2\mathbf{I}_M)$ is the independent additive noise. The BS detects the signal $x_1^{\mathrm{ul}}[k]$ by using the combining vector $\mathbf{u}_1 \in \mathbb{C}^M$, to obtain

$$\begin{aligned} y_1^{\mathrm{ul}}[k] &= \mathbf{u}_1^{\mathsf{H}}\mathbf{r}^{\mathrm{ul}}[k] \\ &= \mathbf{u}_1^{\mathsf{H}}\mathbf{h}_1 x_1^{\mathrm{ul}}[k] + \mathbf{u}_1^{\mathsf{H}}\mathbf{h}_2 x_2^{\mathrm{ul}}[k] + \mathbf{u}_1^{\mathsf{H}}\mathbf{z}^{\mathrm{ul}}[k]. \end{aligned} \qquad (31)$$

[7]As detailed in Section II, we will evaluate the error probability for a Gaussian random code ensemble, where the elements of each codeword are drawn independently from a $\mathcal{CN}(0, \rho^{\mathrm{ul}})$ distribution.

Note that (31) has the same form as (1) with $v[k] = y_1^{\mathrm{ul}}[k]$, $q[k] = x_1^{\mathrm{ul}}[k]$, $g = \mathbf{u}_1^{\mathsf{H}}\mathbf{h}_1$, and $z[k] = \mathbf{u}_1^{\mathsf{H}}\mathbf{h}_2 x_2^{\mathrm{ul}}[k] + \mathbf{u}_1^{\mathsf{H}}\mathbf{z}^{\mathrm{ul}}[k]$. Furthermore, given $\{\mathbf{h}_1, \mathbf{u}_1, \mathbf{h}_2\}$, the random variables $\{z[k] : k = 1, \dots, n_{\mathrm{ul}}\}$ are conditionally i.i.d. and $z[k] \sim \mathcal{CN}(0, \sigma^2)$ with $\sigma^2 = \rho^{\mathrm{ul}}|\mathbf{u}_1^{\mathsf{H}}\mathbf{h}_2|^2 + \|\mathbf{u}_1\|^2\sigma_{\mathrm{ul}}^2$.

We assume that the BS treats the acquired (noisy) channel estimate $\widehat{\mathbf{h}}_1$ as perfect. This implies that, to recover the transmitted codeword, which we assume to be drawn from a codebook $\mathcal{C}^{\mathrm{ul}}$, it performs mismatched SNN decoding with $\widehat{g} = \mathbf{u}_1^{\mathsf{H}}\widehat{\mathbf{h}}_1$. Specifically, the estimated codeword $\widehat{\mathbf{x}}_1^{\mathrm{ul}}$ is obtained as

$$\widehat{\mathbf{x}}_1^{\mathrm{ul}} = \underset{\widetilde{\mathbf{x}}_1^{\mathrm{ul}} \in \mathcal{C}^{\mathrm{ul}}}{\arg\min}\|\mathbf{y}_1^{\mathrm{ul}} - (\mathbf{u}_1^{\mathsf{H}}\widehat{\mathbf{h}}_1)\widetilde{\mathbf{x}}_1^{\mathrm{ul}}\|^2 \qquad (32)$$

with $\mathbf{y}_1^{\mathrm{ul}} = [y_1^{\mathrm{ul}}[1], \dots, y_1^{\mathrm{ul}}[n_{\mathrm{ul}}]]^{\mathsf{T}}$ and $\widetilde{\mathbf{x}}_1^{\mathrm{ul}} = [\widetilde{x}_1^{\mathrm{ul}}[1], \dots, \widetilde{x}_1^{\mathrm{ul}}[n_{\mathrm{ul}}]]^{\mathsf{T}}$. It thus follows that (3) provides a bound on the conditional error probability for UE 1 given g and $\widehat{g}$. To obtain the average error probability, we need to take an expectation over $g = \mathbf{u}_1^{\mathsf{H}}\mathbf{h}_1$, $\widehat{g} = \mathbf{u}_1^{\mathsf{H}}\widehat{\mathbf{h}}_1$, and $\sigma^2 = \rho^{\mathrm{ul}}|\mathbf{u}_1^{\mathsf{H}}\mathbf{h}_2|^2 + \|\mathbf{u}_1\|^2\sigma_{\mathrm{ul}}^2$, which results in

$$\begin{aligned} &\epsilon_1^{\mathrm{ul}} \\ &\leq \mathbb{E}\left[\mathbb{P}\left[\sum_{k=1}^{n_{\mathrm{ul}}} \imath_s(y_1^{\mathrm{ul}}[k], x_1^{\mathrm{ul}}[k]) \leq \log\frac{m-1}{u}\,\bigg|\,g, \widehat{g}, \sigma^2\right]\right]. \end{aligned} \qquad (33)$$

The saddlepoint approximation in Theorem 2 can be applied verbatim to efficiently compute the conditional probability in (33). The average error probability for UE 2 can be evaluated similarly.

The combining vector $\mathbf{u}_1$ is selected at the BS based on the channel estimates $\widehat{\mathbf{h}}_1$ and $\widehat{\mathbf{h}}_2$. The simplest choice is to use MR combining: $\mathbf{u}_1^{\mathrm{MR}} = \widehat{\mathbf{h}}_1/M$. A more computationally intensive choice is MMSE combining:

$$\mathbf{u}_1^{\mathrm{MMSE}} = \left(\sum_{i=1}^{2}\widehat{\mathbf{h}}_i\widehat{\mathbf{h}}_i^{\mathsf{H}} + \mathbf{Z}\right)^{-1}\widehat{\mathbf{h}}_1 \qquad (34)$$

where $\mathbf{Z} = \sum_{i=1}^{2}\boldsymbol{\Phi}_i + \frac{\sigma_{\mathrm{ul}}^2}{\rho^{\mathrm{ul}}}\mathbf{I}_M$.

C. Downlink data transmission

Assume that, to transmit to UE i with $i = 1, 2$, the BS uses the precoding vector $\mathbf{w}_i \in \mathbb{C}^M$, which determines the spatial directivity of the transmission and satisfies the normalization $\mathbb{E}\left[\|\mathbf{w}_i\|^2\right] = 1$. During DL data transmission, the received signal $y_1^{\mathrm{dl}}[k] \in \mathbb{C}$ at UE 1 over channel use k, where $k = 1, \dots, n_{\mathrm{dl}}$, is

$$y_1^{\mathrm{dl}}[k] = \mathbf{h}_1^{\mathsf{H}}\mathbf{w}_1 x_1^{\mathrm{dl}}[k] + \mathbf{h}_1^{\mathsf{H}}\mathbf{w}_2 x_2^{\mathrm{dl}}[k] + z_1^{\mathrm{dl}}[k] \qquad (35)$$

where $x_i^{\mathrm{dl}}[k] \sim \mathcal{CN}(0, \rho^{\mathrm{dl}})$ is the data signal intended for UE i and $z_1^{\mathrm{dl}}[k] \sim \mathcal{CN}(0, \sigma_{\mathrm{dl}}^2)$ is the receiver noise at UE 1. Again, we can put (35) in the same form as (1) by setting $v[k] = y_1^{\mathrm{dl}}[k]$, $q[k] = x_1^{\mathrm{dl}}[k]$, $g = \mathbf{h}_1^{\mathsf{H}}\mathbf{w}_1$ and $z[k] = \mathbf{h}_1^{\mathsf{H}}\mathbf{w}_2 x_2^{\mathrm{dl}}[k] + z_1^{\mathrm{dl}}[k]$. Note that, given $\{\mathbf{h}_1, \mathbf{w}_1, \mathbf{w}_2\}$, the random variables $\{z[k] : k = 1, \dots, n_{\mathrm{dl}}\}$ are conditional i.i.d. and $z[k] \sim \mathcal{CN}(0, \sigma^2)$ with $\sigma^2 = \rho^{\mathrm{dl}}|\mathbf{h}_1^{\mathsf{H}}\mathbf{w}_2|^2 + \sigma_{\mathrm{dl}}^2$.

Since no pilots are transmitted in the DL, the UE does not know the precoded channel $g = \mathbf{h}_1^{\mathsf{H}}\mathbf{w}_1$ in (35). Instead, we

assume that the UE has access its expected value $\mathbb{E}\left[\mathbf{h}_1^{\mathsf{H}}\mathbf{w}_1\right]$ and uses this quantity to perform mismatched SNN decoding. Specifically, we have that $\widehat{g} = \mathbb{E}\left[\mathbf{h}_1^{\mathsf{H}}\mathbf{w}_1\right]$ and

$$\widehat{\mathbf{x}}_1^{\mathrm{dl}} = \arg\min_{\widetilde{\mathbf{x}}_1^{\mathrm{dl}} \in \mathcal{C}^{\mathrm{dl}}} \|\mathbf{y}_1^{\mathrm{dl}} - \widehat{g}\widetilde{\mathbf{x}}_1^{\mathrm{dl}}\|^2 \tag{36}$$

with $\mathbf{y}_1^{\mathrm{dl}} = [y_1^{\mathrm{dl}}[1], \ldots, y_1^{\mathrm{dl}}[n_{\mathrm{dl}}]]^{\mathsf{T}}$ and $\widetilde{\mathbf{x}}_1^{\mathrm{dl}} = [\widetilde{x}_1^{\mathrm{dl}}[1], \ldots, \widetilde{x}_1^{\mathrm{dl}}[n_{\mathrm{dl}}]]^{\mathsf{T}}$. Obviously, channel hardening is critical for this choice to result in good performance [11, Sec. 2.5.1]. Since $\widehat{g} = \mathbb{E}\left[\mathbf{h}_1^{\mathsf{H}}\mathbf{w}_1\right]$ is deterministic, the error probability at UE 1 in the DL can be evaluated as follows:

$$\epsilon_1^{\mathrm{dl}} \leq \mathbb{E}\left[\mathbb{P}\left[\sum_{k=1}^{n_{\mathrm{dl}}} \imath_s(y_1^{\mathrm{dl}}[k], x_1^{\mathrm{dl}}[k]) \leq \log\frac{m-1}{u}\,\bigg|\,g, \sigma^2\right]\right]. \tag{37}$$

Similarly to (33), the saddlepoint approximation in Theorem 2 can be used to evaluate the conditional probability in (37) efficiently.

Similar to the UL, the upper bound (37) holds for any precoder vector that is selected on the basis of the channel estimates available at the BS. Different precoders yield different tradeoffs between the error probability achievable at the UEs. A common heuristic comes from UL-DL duality [11, Sec. 4.3.2], which suggests to choose the precoding vectors $\mathbf{w}_i$ as the following function of the combining vectors: $\mathbf{w}_i = \mathbf{u}_i/\sqrt{\mathbb{E}[\|\mathbf{u}_i\|^2]}$. By selecting $\mathbf{u}_i$ as one of the uplink combining schemes described earlier, the corresponding precoding scheme is obtained; that is, $\mathbf{u}_i = \mathbf{u}_i^{\mathrm{MR}}$ yields MR precoding and $\mathbf{u}_i = \mathbf{u}_i^{\mathrm{MMSE}}$ yields MMSE precoding.

D. Numerical Analysis

In this section, we use the finite blocklength bound in Theorem 1 to study the impact of imperfect CSI, pilot contamination, and spatial correlation in both UL and DL. We assume that the $K = 2$ UEs are within a square area of $75\,\mathrm{m} \times 75\,\mathrm{m}$, with the BS at the center of the square. The BS is equipped with a horizontal uniform linear array (ULA) with antenna elements separated by half a wavelength. The antennas and the UEs are located in the same horizontal plane, thus the azimuth angle is sufficient to determine the directivity. We assume that the scatterers are uniformly distributed in the angular interval $[\varphi_i - \Delta, \varphi_i + \Delta]$, where φ_i is the nominal angle-of-arrival (AoA) of UE i and Δ is the angular spread. Hence, the (m_1, m_2)th element of $\mathbf{R}_i$ is equal to [11, Sec. 2.6]

$$[\mathbf{R}_i]_{m_1,m_2} = \frac{\beta_i}{2\Delta}\int_{-\Delta}^{\Delta} e^{\mathrm{j}\pi(m_1-m_2)\sin(\varphi_i+\bar{\varphi})}d\bar{\varphi}. \tag{38}$$

We assume $\Delta = 25°$ and let the large-scale fading coefficient, measured in dB, be

$$\beta_i = -35.3 - 37.6\log_{10}\left(\frac{d_i}{1\,\mathrm{m}}\right) \tag{39}$$

where d_i is the distance between the BS and UE i. The communication takes place over a $20\,\mathrm{MHz}$ bandwidth with a total receiver noise power of $\sigma_{\mathrm{ul}}^2 = \sigma_{\mathrm{dl}}^2 = -94\,\mathrm{dBm}$ (consisting of thermal noise and a noise figure of $7\,\mathrm{dB}$ in the receiver hardware) at both the BS and UEs. The UL and DL
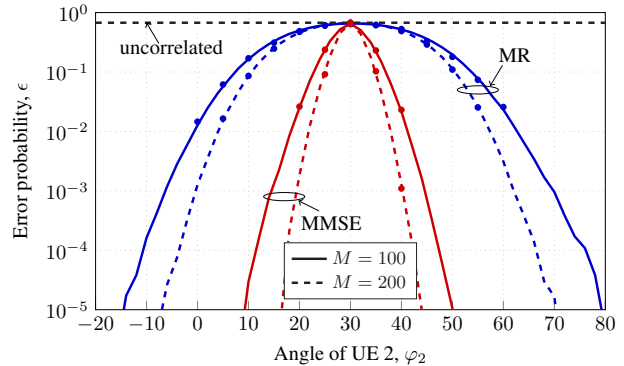
(a) Uplink transmission.
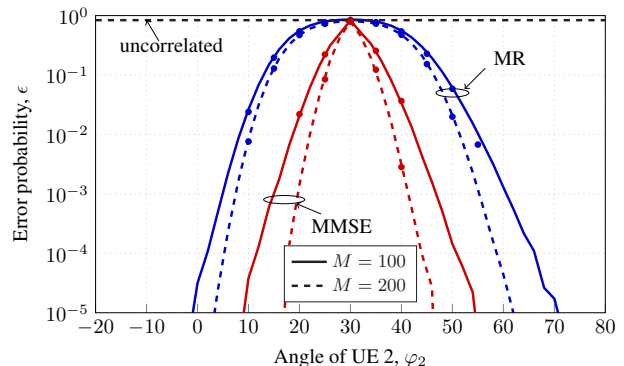
(b) Downlink transmission.

Fig. 2: Average error probability ϵ for UE 1 versus the nominal angle of UE 2 when $\phi_1 = \phi_2$. Here, $\rho^{\mathrm{ul}} = \rho^{\mathrm{dl}} = 10$ dBm, $\Delta = 25°$, $\varphi_1 = 30°$, $b = 160$, $n_{\mathrm{p}} = 2$, and $n = 300$. The curves are obtained using the saddlepoint approximation; the circles indicate the values of the RCUs bound, computed directly via (5).

transmit powers are equal and given by $\rho^{\mathrm{ul}} = \rho^{\mathrm{dl}} = 10\,\mathrm{mW}$. We assume a total of $n = 300$ channel uses, out of which n_{p} channel uses are allocated for pilot transmission and $n_{\mathrm{ul}} = n_{\mathrm{dl}} = (n - n_{\mathrm{p}})/2$ channel uses are assigned to the UL and DL data transmissions, respectively. In each data-transmission phase, $b = 160$ information bits are to be conveyed. These parameters are in agreement with the stringent low-latency setups described in [3, App. A.2.3.1].

Fig. 2 shows the UL and DL error probability ϵ of UE 1 with MR and MMSE combining, when the two UEs use the same pilot sequence (i.e., pilot contamination is present) and $M = 100$ or 200. The uncorrelated Rayleigh-fading case where $\mathbf{R}_i = \beta_i \mathbf{I}_M$, $i = 1, 2$, is also reported as reference. The nominal angle of UE 1 is fixed at $\varphi_1 = 30°$ while the angle of UE 2 varies from $-20°$ to $80°$. We let $d_1 = d_2 = 36.4\,\mathrm{m}$, which leads to $\beta_1 = \beta_2 = -94$ dB. Fig. 2 reveals that a low error probability can be achieved if the UEs are well-separated in the angle domain, even when the channel estimates are affected by pilot contamination. MMSE combining/precoding achieves a much lower error probability for a given angle separation. These results are in agreement with the findings reported in the asymptotic regime of large packet size in [6], [26].

Fig. 2 shows that the error probability with MR combining in the UL is worse than that of MR precoding in the DL. This

phenomenon can be clarified by comparing the input-otput relations in (31) and (35) for the case of perfect CSI at both BS and UEs. Specifically, when the desired signal experiences a deep fade, the magnitude of the UL interference is unaffected whereas the DL interference becomes small. This results in a larger error probability in the UL compared to the DL. The same argument holds also for the case of imperfect CSI with and without pilot contamination. Note that this phenomenon does not occur when MMSE combining/precoding is used. On the contrary, with MMSE combining/precoding the DL performs slightly worse than the UL because DL decoding relies on channel hardening.

Assume now that the 2 UEs are positioned independently and uniformly at random within the square area of $75\,\text{m} \times 75\,\text{m}$, with a minimum distance from the BS of $5\,\text{m}$. Fig. 3 shows the UL and DL network availability η with both MR and MMSE when $M = 100$. We define η as

$$\eta = \mathbb{P}[\epsilon \leq \epsilon_{\text{target}}] \tag{40}$$

and represents the probability that the target error probability ϵ_{target} is achieved on a link between a randomly positioned UE and its corresponding BS, in the presence of randomly positioned interfering UEs (in this case, just one). Note that the error probability ϵ is averaged with respect to the small-scale fading and the additive noise, given the UEs location, whereas the network availability is computed with respect to the random UEs locations. We consider both the scenario in which the UEs use orthogonal pilot sequences, i.e., $\phi_1^{\text{H}}\phi_2 = 0$, and the one in which $\phi_1 = \phi_2$.

The results of Fig. 3 show that pilot contamination reduces significantly the network availability irrespective of the processing scheme. MR performs better in the DL than in the UL when orthogonal pilot sequences are used. This is in agreement with what stated when discussing Fig. 2. However, in the case of pilot contamination, the UL achieves better performance than the DL when the UE relies on channel hardening (and slightly worse performance than the DL when the UE has access to perfect CSI). Note that this does not contradict what stated after Fig. 2. Indeed, due to the random UE placements, the correlation matrix may have low rank. This affects channel hardening and, consequently, results in a deterioration of the DL performance. For MMSE processing, the UL is always superior to the DL because the DL relies on channel hardening.

E. Asymptotic Analysis as $M \to \infty$

It is well known that, for spatially uncorrelated Rayleigh fading channels, the interference caused by pilot contamination limits the spectral efficiency of Massive MIMO in the large-blocklength ergodic setup as $M \to \infty$ and the number of UEs K is fixed, for both MR and MMSE combining/precoding [4], [34]. However, it was recently shown in [6] that Massive MIMO with MMSE combining/precoding is not asymptotically limited by pilot contamination when the spatial correlation exhibited by practically relevant channels is taken into consideration.

We show next that a similar conclusion holds for the *average error probability* in the *finite-blockength* regime when
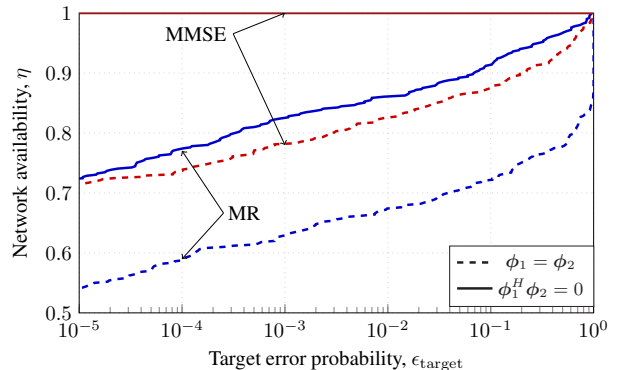
(a) Uplink transmission.
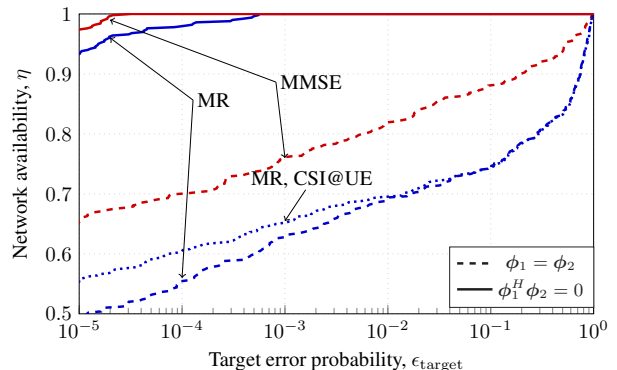
(b) Downlink transmission.

Fig. 3: Network availability η with and without pilot contamination with $M = 100$, $\rho^{\text{ul}} = \rho^{\text{dl}} = 10$ dBm, $n_{\text{p}} = 2$, $b = 160$, $n = 300$, and $\Delta = 25°$.

$M \to \infty$ and $K = 2$.[8] Specifically, we prove that, in the presence of spatial correlation, the error probability vanishes as $M \to \infty$, provided that MMSE combining/precoding is used. To this end, we will proceed similarly as in [6] and make the following two assumptions.

Assumption 1: For $i = 1, 2$, $\liminf_M \frac{1}{M}\text{tr}(\mathbf{R}_i) > 0$ and $\limsup_M \|\mathbf{R}_i\|_2 < \infty$.

Assumption 2: For $(\lambda_1, \lambda_2) \in \mathbb{R}^2$ and $i = 1, 2$,

$$\liminf_M \inf_{\{(\lambda_1, \lambda_2):\lambda_i=1\}} \frac{1}{M}\|\lambda_1\mathbf{R}_1 + \lambda_2\mathbf{R}_2\|_F^2 > 0. \tag{41}$$

The first condition in Assumption 1 implies that the array gathers an amount of signal energy that is proportional to M. The second condition implies that the increased signal energy is spread over many spatial dimensions, i.e., the rank of $\mathbf{R}_i$ must be proportional to M. These two conditions are commonly invoked in the asymptotic analysis of Massive MIMO [34]. Assumption 2 requires $\mathbf{R}_1$ and $\mathbf{R}_2$ to be asymptotically linearly independent [26].

In Theorem 3 below, we establish that, with MR combining, the probability of error vanishes as $M \to \infty$ if the two UEs transmit orthogonal pilot sequences. However, it converges to a positive constant if they share the same pilot sequence.

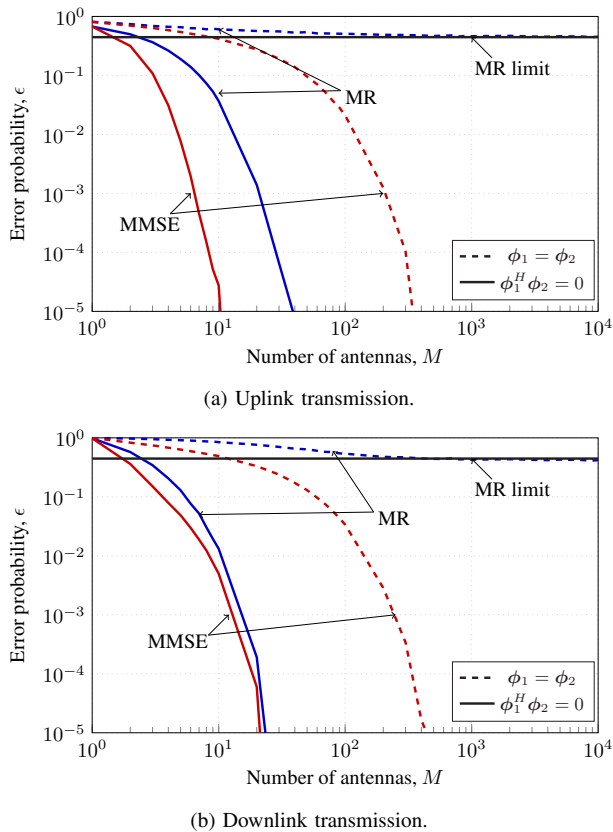[8]We consider the case $K = 2$ for simplicity, although a similar result can be obtained for arbitrary K using the same approach.

(a) Uplink transmission.

(b) Downlink transmission.

Fig. 4: Average error probability ϵ of UE 1 versus number of antennas M with and without pilot contamination. Here, $\rho^{\mathrm{ul}} = \rho^{\mathrm{dl}} = 10$ dBm, $n_{\mathrm{p}} = 2$, $b = 160$, $n = 300$, $\Delta = 25°$, $\varphi_1 = 30°$, and $\varphi_2 = 40°$.

Theorem 3: Let $c > 0$ be a positive real-valued scalar. If MR combining is used with $\mathbf{u}_1^{\mathrm{MR}} = \frac{1}{M}\widehat{\mathbf{h}}_1$, then under Assumption 1,

$$\lim_{M\to\infty} \epsilon_1^{\mathrm{ul}} = 0, \text{ if } \boldsymbol{\phi}_1^{\mathrm{H}}\boldsymbol{\phi}_2 = 0, \tag{42}$$

$$\lim_{M\to\infty} \epsilon_1^{\mathrm{ul}} = c, \text{ if } \boldsymbol{\phi}_1 = \boldsymbol{\phi}_2. \tag{43}$$

Proof: See Appendix C. ∎

Next, we show that, if MMSE combining is used, the error probability vanishes as $M \to \infty$ even in the presence of pilot contamination.

Theorem 4: If MMSE combining is used with $\mathbf{u}_1^{\mathrm{MMSE}}$ given by (34), then under Assumption 1 and Assumption 2, the average error probability ϵ_1^{ul} goes to zero as $M \to \infty$, both when $\boldsymbol{\phi}_1^{\mathrm{H}}\boldsymbol{\phi}_2 = 0$ and when $\boldsymbol{\phi}_1 = \boldsymbol{\phi}_2$.

Proof: The proof is given in Appendix D. It makes use of the asymptotic analysis presented in [6, App. B] to show that $y_1^{\mathrm{ul}} \asymp x_1^{\mathrm{ul}}$ as $M \to \infty$, even in the presence of pilot contamination. Once this is proved, the result follows by applying Lemma 1 from Section II. ∎

Note that Theorem 3 and Theorem 4 can be extended to the DL with a similar methodology. Details are omitted due to space limitations.

To validate the asymptotic analysis provided by Theorems 3 and 4 and to quantify the impact of pilot contamination for values of M of practical interest, we numerically evaluate the UL error probability when the 2 UEs transmit at the same

power, are at the same distance from the BS, and use the same pilot sequence. Furthermore, we assume that their nominal angles are $\varphi_1 = 30°$ and $\varphi_2 = 40°$. Note that the angle between the two UEs is small. Hence, we expect pilot contamination to have a significant impact on the error probability. As in Fig. 2, we assume that $\sigma_{\mathrm{ul}}^2 = \sigma_{\mathrm{dl}}^2 = -94\,\mathrm{dBm}$ and that the UEs are located 36.4 m away from the BS so that $\beta_1 = \beta_2 = -94\,\mathrm{dB}$. In Fig. 4a, we illustrate the average error probability as a function of M with MR and MMSE. We see that, in the presence of pilot contamination, the error probability with MR converges to a nonzero constant as M grows, in accordance with Theorem 3. In contrast, the error probability with MMSE goes to 0 as $M \to \infty$, in accordance with Theorem 4. However, a comparison with the orthogonal-pilot case reveals that, for fixed M, pilot contamination has a significant impact on the error probability of MMSE. As shown in Fig. 4b, similar conclusions can be drawn for the DL.

IV. MASSIVE MIMO NETWORK

We will now extend the analysis in Section III to a Massive MIMO network with L cells, each comprising a BS with M antennas and K UEs. We denote by $\mathbf{h}_{li}^j \sim \mathcal{CN}(\mathbf{0}_M, \mathbf{R}_{li}^j)$ the channel between UE i in cell l and the BS in cell j. The n_{p}-length pilot sequence of UE i in cell j is denoted by the vector $\boldsymbol{\phi}_{ji} \in \mathbb{C}^{n_{\mathrm{p}}}$ and satisfies $\|\boldsymbol{\phi}_{ji}\|^2 = n_{\mathrm{p}}$. We assume that the K UEs in a cell use mutually orthogonal pilot sequences and these pilot sequences are reused in a fraction $1/f$ of the L cells with $n_{\mathrm{p}} = Kf$. The channel vectors are estimated using the MMSE estimator given in [11, Sec. 3.2].

A. Uplink

The data signal from UE i' in cell l over an arbitrary time instant k is denoted by $x_{li'}^{\mathrm{ul}}[k] \sim \mathcal{CN}(0, \rho^{\mathrm{ul}})$, with ρ^{ul} being the transmit power. To detect $x_{ji}^{\mathrm{ul}}[k]$, BS j selects the combining vector $\mathbf{u}_{ji} \in \mathbb{C}^M$, which is multiplied with the received signal $\mathbf{r}_j^{\mathrm{ul}}[k]$ to obtain

$$y_{ji}^{\mathrm{ul}}[k] = \mathbf{u}_{ji}^{\mathrm{H}}\mathbf{r}_j^{\mathrm{ul}}[k] = \overbrace{\underbrace{\mathbf{u}_{ji}^{\mathrm{H}}\mathbf{h}_{ji}^j x_{ji}^{\mathrm{ul}}[k]}_{\text{Desired signal}}}^{gq[k]} + \overbrace{\underbrace{\sum_{i'=1,i'\neq i}^{K} \mathbf{u}_{ji}^{\mathrm{H}}\mathbf{h}_{ji'}^j x_{ji'}^{\mathrm{ul}}[k]}_{\text{Intra-cell interference}}}^{z[k]}$$

$$+ \overbrace{\underbrace{\sum_{l=1,l\neq j}^{L}\sum_{i'=1}^{K} \mathbf{u}_{ji}^{\mathrm{H}}\mathbf{h}_{li'}^j x_{li'}^{\mathrm{ul}}[k]}_{\text{Inter-cell interference}}}^{z[k]} + \overbrace{\underbrace{\mathbf{u}_{ji}^{\mathrm{H}}\mathbf{z}_j^{\mathrm{ul}}[k]}_{\text{Noise}}}^{z[k]} \tag{44}$$

for $k = 1, \ldots, n_{\mathrm{ul}}$. We note that (44) can be put in the same form as (1) if we set $v[k] = y_{ji}^{\mathrm{ul}}[k]$, $q[k] = x_{ji}^{\mathrm{ul}}[k]$, $g = \mathbf{u}_{ji}^{\mathrm{H}}\mathbf{h}_{ji}^j$, $\widehat{g} = \mathbf{u}_{ji}^{\mathrm{H}}\widehat{\mathbf{h}}_{ji}^j$, and $z[k] = \sum_{i'=1,i'\neq i}^{K} \mathbf{u}_{ji}^{\mathrm{H}}\mathbf{h}_{ji'}^j x_{ji'}^{\mathrm{ul}}[k] + \sum_{l=1,l\neq j}^{L}\sum_{i'=1}^{K} \mathbf{u}_{ji}^{\mathrm{H}}\mathbf{h}_{li'}^j x_{li'}^{\mathrm{ul}}[k] + \mathbf{u}_{ji}^{\mathrm{H}}\mathbf{z}_j^{\mathrm{ul}}[k]$. Given all channels and combining vectors, the random variables $\{z[k] : k = 1, \ldots, n_{\mathrm{ul}}\}$ are conditionally i.i.d. and $z[k] \sim \mathcal{CN}(0, \sigma^2)$ with $\sigma^2 = \sigma_{\mathrm{ul}}^2\|\mathbf{u}_{ji}\|^2 + \rho^{\mathrm{ul}}\sum_{i'=1,i'\neq i}^{K}|\mathbf{u}_{ji}^{\mathrm{H}}\mathbf{h}_{ji'}^j|^2 +$

$\rho^{\mathrm{ul}} \sum_{l=1, l \neq j}^{L} \sum_{i'=1}^{K} |\mathbf{u}_{ji}^{\mathsf{H}} \mathbf{h}_{li'}^{j}|^2$. An upper bound on the error probability $\epsilon_{ji}^{\mathrm{ul}}$ then follows by applying (3) in Theorem 1 and then by averaging over g, $\widehat{g}$ and σ^2. This bound holds for any choice of $\mathbf{v}_{ji}$. In the numerical results, we will consider multicell MMSE and MR combining.
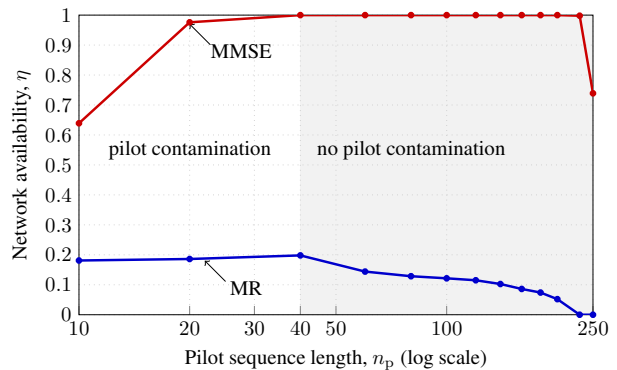
B. Downlink

The BS in cell j transmits the DL signal $\mathbf{x}_j^{\mathrm{dl}}[k] = \sum_{ji'=1}^{K} \mathbf{w}_{ji'} x_{ji'}^{\mathrm{dl}}[k]$ where $x_{ji'}^{\mathrm{dl}}[k] \sim \mathcal{CN}(0, \rho^{\mathrm{dl}})$ is the DL data signal intended for UE i' in cell j over the time index k, assigned to a precoding vector $\mathbf{w}_{ji'} \in \mathbb{C}^M$ that satisfies $\|\mathbf{w}_{ji'}\|^2 = 1$ so that ρ^{dl} represents the transmit power. The received signal $y_{ji}^{\mathrm{dl}}[k] \in \mathbb{C}$ for $k = 1, \ldots, n_{\mathrm{dl}}$ at UE i in cell j is given by

$$
y_{ji}^{\mathrm{dl}}[k] = \overbrace{(\mathbf{h}_{ji}^{j})^{\mathsf{H}} \mathbf{w}_{ji} x_{ji}^{\mathrm{dl}}[k]}^{gq[k]} + \overbrace{\sum_{i'=1, i' \neq i}^{K} (\mathbf{h}_{ji}^{j})^{\mathsf{H}} \mathbf{w}_{ji'} x_{ji'}^{\mathrm{dl}}[k]}^{z[k]}
$$

Desired signal   Intra-cell interference

$$
+ \underbrace{\sum_{l=1, l \neq j}^{L} \sum_{i'=1}^{K} (\mathbf{h}_{ji}^{l})^{\mathsf{H}} \mathbf{w}_{li'} x_{li'}^{\mathrm{dl}}[k]}_{\text{Inter-cell interference}} + \underbrace{z_{ji}^{\mathrm{dl}}[k]}_{\text{Noise}} \quad (45)
$$

where $z_{ji}^{\mathrm{dl}}[k] \sim \mathcal{CN}(0, \sigma_{\mathrm{dl}}^2)$ is the receiver noise. The desired signal to UE i in cell j propagates over the precoded channel $g_{ji} = (\mathbf{h}_{ji}^{j})^{\mathsf{H}} \mathbf{w}_{ji}$. The UE does not know g_{ji} and relies on channel hardening to approximate it with its mean value $\mathbb{E}[g_{ji}] = \mathbb{E}\left[(\mathbf{h}_{ji}^{j})^{\mathsf{H}} \mathbf{w}_{ji}\right]$. As in the UL, we note that (45) can be put in the same form as (1) if we set $v[k] = y_{ji}^{\mathrm{dl}}[k]$, $q[k] = x_{ji}^{\mathrm{dl}}[k]$, $g = (\mathbf{h}_{ji}^{j})^{\mathsf{H}} \mathbf{w}_{ji}$, $\widehat{g} = \mathbb{E}\left[(\mathbf{h}_{ji}^{j})^{\mathsf{H}} \mathbf{w}_{ji}\right]$, and $z[k] = \sum_{i'=1, i' \neq i}^{K} (\mathbf{h}_{ji}^{j})^{\mathsf{H}} \mathbf{w}_{ji'} x_{ji'}^{\mathrm{dl}}[k] + \sum_{l=1, l \neq j}^{L} \sum_{i'=1}^{K} (\mathbf{h}_{ji}^{l})^{\mathsf{H}} \mathbf{w}_{li'} x_{li'}^{\mathrm{dl}}[k] + z_{ji}^{\mathrm{dl}}[k]$. Given all channels and precoding vectors, the random variables $\{z[k] : k = 1, \ldots, n_{\mathrm{dl}}\}$ are conditionally i.i.d. and $z[k] \sim \mathcal{CN}(0, \sigma^2)$ with $\sigma^2 = \sigma_{\mathrm{dl}}^2 + \rho^{\mathrm{dl}} \sum_{i'=1, i' \neq i'}^{K} |(\mathbf{h}_{ji}^{j})^{\mathsf{H}} \mathbf{w}_{ji'}|^2 + \rho^{\mathrm{dl}} \sum_{l=1, l \neq j}^{L} \sum_{i'=1}^{K} |(\mathbf{h}_{ji}^{l})^{\mathsf{H}} \mathbf{w}_{li'}|^2$. An upper bound on the error probability $\epsilon_{ji}^{\mathrm{dl}}$ then follows by applying (3) in Theorem 1 and then by averaging over g and σ^2. As for the UL, the above results hold for any choice of $\mathbf{w}_{ji}$. In the numerical simulations, we will consider both multicell MMSE and MR precoding.

C. Numerical Analysis

The simulation setup consists of $L = 4$ square cells, each of size 75 m $\times$ 75 m, containing $K = 10$ UEs each, independently and uniformly distributed within the cell, at a distance of at least 5 m from the BS. As in Section III-D, we consider a horizontal ULA with $M = 100$ antennas and half-wavelength spacing. The correlation matrix and large-scale fading coefficient associated with each UE follow the models given in (38) and (39), respectively. Furthermore, we employ a wrap-around topology as in [11, Sec. 4.1.3]. As in

(a) Uplink transmission.
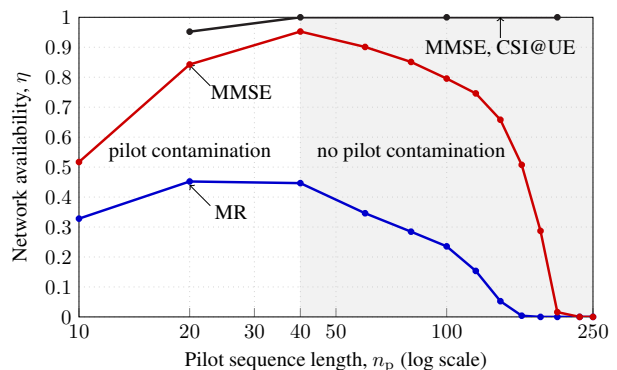
(b) Downlink transmission.

Fig. 5: Network availability for $\epsilon_{\mathrm{target}} = 10^{-5}$. Here, $L = 4$, $K = 10$, $\Delta = 25°$, the cell size is 75×75 m, $\rho^{\mathrm{ul}} = \rho^{\mathrm{dl}} = 10$ dBm, $M = 100$, $b = 160$, and $n = 300$.

Section III-D, we assume $n = 300$, $n_{\mathrm{ul}} = n_{\mathrm{dl}} = (n - n_{\mathrm{p}})/2$, $b = 160$ and $\rho^{\mathrm{ul}} = \rho^{\mathrm{dl}} = 10$ dBm.

In Fig. 5, we plot the network availability (40) for a fixed $\epsilon_{\mathrm{target}} = 10^{-5}$ (which is in agreement with the URLLC requirements) versus the number of pilot symbols $n_{\mathrm{p}} = fK$, where we recall that f is the pilot reuse factor. The results presented in Fig. 3 suggest that pilot contamination should be avoided and that MMSE should be preferred to MR. The results presented in Fig. 5 confirm these design guidelines. With multicell MMSE, a network availability above 90% can be achieved in UL and DL by setting a pilot reuse factor $f = 4$ such that $n_{\mathrm{p}} = fK = 40$. This is the minimum value of n_{p} that results in no pilot contamination in a network with $L = 4$ cells. Increasing n_{p} further has a deleterious effect on the network availability, especially in the DL. Indeed, the corresponding reduction in the number of channel uses $n_{\mathrm{dl}} = (300 - n_{\mathrm{p}})/2$ available for data transmission in the DL overcomes the benefits of a more accurate CSI. As already discussed, the difference in performance between UL and DL with multicell MMSE processing is due to the assumption that the UE has no CSI and performs mismatched decoding by relying on channel hardening. Indeed, when the UEs are provided with perfect CSI (black curve), the network availability achievable in UL and DL is the same. If needed, additional network-availability gains can be achieved by, e.g., increasing the number of BS antennas, by reducing the number

of UEs that are served simultaneously, or by using scheduling to avoid serving at the same time UEs that are difficult to separate spatially via linear precoding. For example, in the scenario considered in Fig. 5b, a network availability above 98% can be achieved by halving the number of scheduled UEs. Finally, note that the network availability achievable with MR is below 50% even when pilot contamination is avoided. This implies that in practical scenarios, MR is too sensitive to interference to achieve the low error probability targets required in URLLC.

V. CONCLUSIONS

We presented guidelines on the design of Massive MIMO systems supporting the transmission of short information packets under the high reliability targets demanded in URLLC. Specifically, we showed that, for a BS equipped with up to 100 antennas, it is imperative to avoid pilot contamination and to use MMSE spatial processing in place of the computationally less intensive MR spatial processing. Our guidelines were based on a firm nonasymptotic bound on the error probability, which is based on recent results in finite-blocklength information theory, and applies to a realistic Massive MIMO network, with imperfect channel state information, pilot contamination, spatially correlated channels, arbitrary linear spatial processing, and randomly positioned UEs. We provided an accurate approximation for this bound, based on the saddlepoint method, which makes its evaluation computationally efficient for the low error probabilities targeted in URLLC. Finally, we showed that analyses based on performance metrics such as outage probability and normal approximation, although appealing because of the simplicity of the underlying mathematical formulas, may result in a significant underestimation of the error probability, which is clearly undesirable when designing URLLC links. Results relied on the assumption that the channel covariance matrix is perfectly known to the receiver. If no such knowledge is available, the BS can perform instead least-square channel estimation followed by regularized zero forcing, at the cost of a performance loss recently quantified in [35].

APPENDIX A - PROOF OF THEOREM 1

Let $\mathbf{q} = [q[1], \ldots, q[n]] \sim \mathcal{CN}(\mathbf{0}, \rho\mathbf{I}_n)$ be the transmitted codeword and $\mathbf{v} = [v[1], \ldots, v[n]]$ be the corresponding channel output obtained via the input-output relation (1). Finally, let $\bar{\mathbf{q}} = [\bar{q}[1], \ldots, \bar{q}[n]]$ be a vector of i.i.d. $\mathcal{CN}(0, \rho)$ random variables, independent of both $\mathbf{q}$ and $\mathbf{v}$. Intuitively, $\bar{\mathbf{q}}$ stands for any codeword different from the transmitted one.

A simple generalization of the random coding union bound in [18, Th. 16] to the mismatched SNN decoder (2) results in the following bound

$$\epsilon \leq \mathbb{E}[\min\{1, (m-1)f(\mathbf{q}, \mathbf{v})\}] \qquad (46)$$

where $f(\mathbf{q}, \mathbf{v}) = \Pr\{\|\mathbf{v} - \widehat{g}\bar{\mathbf{q}}\|^2 \leq \|\mathbf{v} - \widehat{g}\mathbf{q}\|^2 | \mathbf{q}, \mathbf{v}\}$. The bound (46) is obtained by observing that, when the mismatched SNN decoder (2) is used, an error occurs if, after being scaled by $\widehat{g}$, the codeword $\bar{\mathbf{q}}$ is closer in Euclidean distance to $\mathbf{v}$ than $\mathbf{q}$, and then by using a tightened version of the union bound. We next apply the Chernoff bound to $f(\mathbf{q}, \mathbf{v})$ and obtain that

$$f(\mathbf{q}, \mathbf{v}) \leq \frac{\mathbb{E}_{\bar{\mathbf{q}}}\left[\exp\left(-s\|\mathbf{v} - \widehat{g}\bar{\mathbf{q}}\|^2\right)\right]}{\exp(-s\|\mathbf{v} - \widehat{g}\mathbf{q}\|^2)} \qquad (47)$$

for $s > 0$. Substituting (47) into (46), we conclude that

$$\epsilon \leq \mathbb{E}\left[\min\left\{1, \exp\left(\log(m-1)\right.\right.\right.$$
$$\left.\left.\left. + \log \frac{\mathbb{E}_{\bar{\mathbf{q}}}\left[\exp\left(-s\|\mathbf{v} - \widehat{g}\bar{\mathbf{q}}\|^2\right)\right]}{\exp(-s\|\mathbf{v} - \widehat{g}\mathbf{q}\|^2)}\right)\right\}\right] \qquad (48)$$

$$= \mathbb{E}\left[\min\left\{1, \exp\left(\log(m-1)\right.\right.\right.$$
$$\left.\left.\left. - \sum_{k=1}^{n} \log \frac{\exp\left(-s|v[k] - \widehat{g}q[k]|^2\right)}{\mathbb{E}_{\bar{q}[k]}[\exp(-s|v[k] - \widehat{g}\bar{q}[k]|^2)]}\right)\right\}\right]. \qquad (49)$$

Let now the generalized information density be defined as

$$\imath_s(q[k], v[k]) = \log \frac{\exp\left(-s|v[k] - \widehat{g}q[k]|^2\right)}{\mathbb{E}_{\bar{q}[k]}[\exp(-s|v[k] - \widehat{g}\bar{q}[k]|^2)]}. \qquad (50)$$

Using (50), we can rewrite (49) as

$$\epsilon \leq \mathbb{E}\left[\min\left\{1, \exp\left(\log(m-1)\right.\right.\right.$$
$$\left.\left.\left. - \sum_{k=1}^{n} \imath_s(q[k], v[k])\right)\right\}\right]. \qquad (51)$$

The desired bound (3) follows by observing that, for every positive random variable w, we have that $\mathbb{E}[\min\{1, w\}] = \mathbb{P}[w \geq u]$ where u is uniformly distributed on $[0, 1]$.

To conclude the proof, it remains to show that the generalized information density defined in (50) can be expressed as in (4). Since $\bar{q}[k] \sim \mathcal{CN}(0, \rho)$, it follows that, for a given $v[k]$

$$|v[k] - \widehat{g}\bar{q}[k]|^2 \stackrel{d}{=} \frac{|\widehat{g}|^2\rho}{2}\left(\left(\frac{|v[k]|\sqrt{2}}{|\widehat{g}|\sqrt{\rho}} + u_1\right)^2 + u_2^2\right)$$
$$\stackrel{d}{=} \frac{|\widehat{g}|^2\rho}{2}\theta \qquad (52)$$

where u_1 and u_2 are independent $\mathcal{N}(0, 1)$ random variables and θ follows a noncentral chi-squared distribution with 2 degrees of freedom and noncentrality parameter $\lambda = 2|v[k]|^2/(\rho|\widehat{g}|^2)$. The MGF of θ is given by

$$\mathbb{E}\left[e^{\zeta\theta}\right] = \frac{\exp\left(\frac{\lambda\zeta}{1-2\zeta}\right)}{(1-2\zeta)}, \quad \zeta < \frac{1}{2}. \qquad (53)$$

Using (53) in (50) with $\zeta = -s|\widehat{g}|^2\rho/2$, we conclude that (50) coincides with (4).

APPENDIX B - PROOF OF (10) AND (11)

In this appendix, we prove that (9) holds for every $\zeta \in [\underline{\zeta}, \overline{\zeta}]$, where $\underline{\zeta}$ and $\overline{\zeta}$ are given in (10) and (11), respectively. Let $q \sim \mathcal{CN}(0, \rho)$ and $v = gq + z$ where $z \sim \mathcal{CN}(0, \sigma^2)$, so that $v \sim \mathcal{CN}(0, \sigma_v^2)$ with $\sigma_v^2 = \rho|g|^2 + \sigma^2$ Furthermore, set

$A = s|v - \widehat{g}q|^2$ and $B = \gamma|v|^2$ with $\gamma = s/(1 + s\rho|\widehat{g}|^2)$. We can then rewrite the information density (4) as[9]

$$\imath_s(q,v) = B - A + \log\left(1 + s\rho|\widehat{g}|^2\right) \tag{54}$$

It then follows that A and B are dependent exponentially-distributed random variables with rate parameter $1/\beta_A$ defined in (12) and $1/\beta_B$ defined in (13), respectively. This implies that the random variable $\imath_s(q,v)$ involves the difference between two dependent exponentially-distributed random variables. Let $\Delta = B - A$. The probability density function (PDF) of Δ is [36, Cor. 8]

$$f_\Delta(\delta) = \frac{1}{\sqrt{(\beta_B - \beta_A)^2 + 4\beta_A\beta_B(1 - \nu)}}$$
$$\times \exp\left(-\frac{|\delta|\sqrt{(\beta_B - \beta_A)^2 + 4\beta_A\beta_B(1 - \nu)}}{2\beta_A\beta_B(1 - \nu)}\right)$$
$$\times \exp\left(\frac{\delta(\beta_B - \beta_A)}{2\beta_A\beta_B(1 - \nu)}\right) \tag{55}$$

where $\nu = \text{Cov}(A, B)/\sqrt{\text{Var}(A)\,\text{Var}(B)}$ is the correlation coefficient between A and B. Using (55), we can express the MGF of $-\imath_s(q, v)$ as follows:

$$\mathbb{E}\left[e^{-\zeta\imath_s(q,v)}\right] = \frac{1}{(1 + s\rho|\widehat{g}|^2)^\zeta} \int_{-\infty}^{\infty} \exp(-\zeta\delta)\, f_\Delta(\delta)\mathrm{d}\delta$$
$$= \frac{(1 + s\rho|\widehat{g}|^2)^{-\zeta}}{1 + (\beta_B - \beta_A)\zeta - \beta_A\beta_B(1 - \nu)\zeta^2} \tag{56}$$

where the last step holds for all $\zeta \in (\underline{\zeta}, \overline{\zeta})$, with $\underline{\zeta}$ and $\overline{\zeta}$ given in (10) and (11), respectively. The desired result in (9) follows because the right-hand side of (56) is infinitely differentiable.

To conclude the proof, we need to show that ν in (55) is given by (14). By definition, $\text{Cov}(A, B) = \mathbb{E}[AB] - \mathbb{E}[A]\,\mathbb{E}[B]$ where

$$\mathbb{E}[AB] = s\gamma\,\mathbb{E}\left[|v - \widehat{g}q|^2|v|^2\right]. \tag{57}$$

To compute this correlation, it turns out convenient to set $x = v - \widehat{g}q$ and to express x as the MMSE estimate of v given x plus the uncorrelated estimation error e:

$$x = \alpha v + e. \tag{58}$$

Here, α is the MMSE coefficient, given by $\alpha = \mathbb{E}[v^*x]/\sigma_v^2$, and $e \sim \mathcal{CN}(0, \sigma_e^2)$ where $\sigma_e^2 = \sigma_x^2 - |\mathbb{E}[v^*x]|^2/\sigma_v^2$, with $\sigma_x^2 = \mathbb{E}\left[|x|^2\right] = |g - \widehat{g}|^2\rho + \sigma^2$. Note that since e is Gaussian and uncorrelated with v, then e and v are independent. Using (58), we can rewrite the expectation on the right-hand side of (57) as follows:

$$\mathbb{E}\left[|v - \widehat{g}q|^2|v|^2\right] = \mathbb{E}\left[|x|^2|v|^2\right]$$
$$= \mathbb{E}\left[|\alpha v + e|^2|v|^2\right]$$
$$= |\alpha|^2\,\mathbb{E}\left[|v|^4\right] + \mathbb{E}\left[|v|^2\right]\mathbb{E}\left[|e|^2\right]$$
$$= 2|\alpha|^2\sigma_v^4 + \sigma_v^2\sigma_e^2 \tag{59}$$
$$= |\mathbb{E}[v^*x]|^2 + \sigma_v^2\sigma_x^2. \tag{60}$$

Here, in (59) we used that $\mathbb{E}\left[|v|^4\right] = 2\sigma_v^4$. Furthermore, (60) follows by the definition of α and of σ_e^2. Note now that $\beta_A =$

$s\sigma_x^2$ and $\beta_B = \gamma\sigma_v^2$. Recall also that $\mathbb{E}[A] = \beta_A$, $\text{Var}(A) = \beta_A^2$, $\mathbb{E}[B] = \beta_B$, $\text{Var}(B) = \beta_B^2$. Hence, we conclude that

$$\nu = \frac{s\gamma(|\mathbb{E}[v^*x]|^2 + \sigma_v^2\sigma_x^2) - \beta_A\beta_B}{\beta_A\beta_B} = \frac{s\gamma|\mathbb{E}[v^*x]|^2}{\beta_A\beta_B}. \tag{61}$$

To obtain (14), we use that $\gamma = s/(1 + s\rho|\widehat{g}|^2)$ and that $\mathbb{E}[v^*x] = \sigma_v^2 - g^*\widehat{g}\rho$.

APPENDIX C - PROOF OF THEOREM 3

Substituting $\mathbf{u}_1^{\text{MR}} = \frac{1}{M}\widehat{\mathbf{h}}_1$ into (31), we obtain

$$y_1^{\text{ul}}[k] = \frac{1}{M}\widehat{\mathbf{h}}_1^{\text{H}}\mathbf{h}_1 x_1^{\text{ul}}[k] + \frac{1}{M}\widehat{\mathbf{h}}_1^{\text{H}}\mathbf{h}_2 x_2^{\text{ul}}[k] + \frac{1}{M}\widehat{\mathbf{h}}_1^{\text{H}}\mathbf{z}^{\text{ul}}[k]. \tag{62}$$

Under Assumption 1 and using [6, Lem. 3], we have that, in the limit $M \to \infty$,[10]

$$\frac{1}{M}\widehat{\mathbf{h}}_1^{\text{H}}\mathbf{h}_1 \overset{(a)}{\asymp} \frac{1}{M}\widehat{\mathbf{h}}_1^{\text{H}}\widehat{\mathbf{h}}_1 \asymp \frac{1}{M}\text{tr}(\mathbf{\Phi}_1) \text{ and } \frac{1}{M}\widehat{\mathbf{h}}_1^{\text{H}}\mathbf{z}^{\text{ul}}[k] \overset{(b)}{\asymp} 0. \tag{63}$$

Here, (a) and (b) follow because $\widehat{\mathbf{h}}_1$ and the pair $(\widetilde{\mathbf{h}}_1, \mathbf{z}^{\text{ul}}[k])$ are independent. Similarly, we have that

$$\frac{1}{M}\widehat{\mathbf{h}}_1^{\text{H}}\mathbf{h}_2 \asymp 0, \text{ if } \phi_1 \neq \phi_2 \text{ with } \phi_1^{\text{H}}\phi_2 = 0, \tag{64}$$

$$\frac{1}{M}\widehat{\mathbf{h}}_1^{\text{H}}\mathbf{h}_2 \asymp \frac{1}{M}\widehat{\mathbf{h}}_1^{\text{H}}\widehat{\mathbf{h}}_2 \overset{(c)}{\asymp} \frac{1}{M}\text{tr}(\mathbf{\Upsilon}_{12}), \text{ if } \phi_1 = \phi_2 \tag{65}$$

where (c) follows from the fact that $\widehat{\mathbf{h}}_1$ and $\widehat{\mathbf{h}}_2$ are correlated under pilot contamination. Using (63), (64), and (65) in (62), we conclude that

$$y_1^{\text{ul}}[k] \asymp \frac{1}{M}\text{tr}(\mathbf{\Phi}_1)x_1^{\text{ul}}[k], \text{ if } \phi_1 \neq \phi_2 \text{ with } \phi_1^{\text{H}}\phi_2 = 0, \tag{66}$$

$$y_1^{\text{ul}}[k] \asymp \frac{\text{tr}(\mathbf{\Phi}_1)x_1^{\text{ul}}[k] + \text{tr}(\mathbf{\Upsilon}_{12})x_2^{\text{ul}}[k]}{M}, \text{ if } \phi_1 = \phi_2. \tag{67}$$

This implies that, in the absence of pilot contamination, the input-output relation becomes that of a deterministic noiseless channel as $M \to \infty$, while it converges to that of an AWGN channel with transmit power $\lim_{M\to\infty}[\frac{1}{M}\text{tr}(\mathbf{\Phi}_1)]^2$ and noise variance $\lim_{M\to\infty}[\frac{1}{M}\text{tr}(\mathbf{\Upsilon}_{12})]^2$ when the two UEs use the same pilot sequence. Note also that $g \asymp \widehat{g} \asymp \frac{1}{M}\text{tr}(\mathbf{\Phi}_1)$ where g and $\widehat{g}$ are defined in (33). The desired result then follows from (6).

APPENDIX D - PROOF OF THEOREM 4

We only consider the case in which the two UEs use the same pilot sequence, i.e., $\phi_1 = \phi_2$. By applying the matrix inversion lemma (see, e.g., [6, Lem. 4]) we can rewrite (34) as

$$\mathbf{u}_1^{\text{MMSE}} = \left(\sum_{i=1}^{2}\widehat{\mathbf{h}}_i\widehat{\mathbf{h}}_i^{\text{H}} + \mathbf{Z}\right)^{-1}\widehat{\mathbf{h}}_1 \tag{68}$$

$$= \frac{1}{1 + \gamma_1^{\text{ul}}}\left(\widehat{\mathbf{h}}_2\widehat{\mathbf{h}}_2^{\text{H}} + \mathbf{Z}\right)^{-1}\widehat{\mathbf{h}}_1 \tag{69}$$

---

[9]We drop the indices in q and v because immaterial for the proof.

[10]Under Assumption 1, $\mathbf{R}_k\mathbf{Q}^{-1}\mathbf{R}_i$ has uniformly bounded spectral norm—a result that follows from [6, Lem. 4].

where we have set $\gamma_1^{\text{ul}} = \widehat{\mathbf{h}}_1^{\text{H}} \left(\widehat{\mathbf{h}}_2 \widehat{\mathbf{h}}_2^{\text{H}} + \mathbf{Z} \right)^{-1} \widehat{\mathbf{h}}_1$. Substituting (69) into (31) we obtain, after multiplying and dividing each term by M,

$$
\begin{aligned}
y_1^{\text{ul}} = {}& \frac{1}{\frac{1}{M} + \frac{\gamma_1^{\text{ul}}}{M}} \frac{1}{M} \widehat{\mathbf{h}}_1^{\text{H}} \left(\widehat{\mathbf{h}}_2 \widehat{\mathbf{h}}_2^{\text{H}} + \mathbf{Z} \right)^{-1} \mathbf{h}_1 x_1^{\text{ul}} \\
& + \frac{1}{\frac{1}{M} + \frac{\gamma_1^{\text{ul}}}{M}} \frac{1}{M} \widehat{\mathbf{h}}_1^{\text{H}} \left(\widehat{\mathbf{h}}_2 \widehat{\mathbf{h}}_2^{\text{H}} + \mathbf{Z} \right)^{-1} \mathbf{h}_2 x_2^{\text{ul}} \\
& + \frac{1}{\frac{1}{M} + \frac{\gamma_1^{\text{ul}}}{M}} \frac{1}{M} \widehat{\mathbf{h}}_1^{\text{H}} \left(\widehat{\mathbf{h}}_2 \widehat{\mathbf{h}}_2^{\text{H}} + \mathbf{Z} \right)^{-1} \mathbf{z}^{\text{ul}}. \quad (70)
\end{aligned}
$$

We begin by considering the first term. Under Assumption 1 and using [6, Lem. 3], we obtain

$$
\frac{1}{M} \widehat{\mathbf{h}}_1^{\text{H}} \left(\widehat{\mathbf{h}}_2 \widehat{\mathbf{h}}_2^{\text{H}} + \mathbf{Z} \right)^{-1} \mathbf{h}_1 \asymp \frac{\gamma_1^{\text{ul}}}{M} \quad (71)
$$

since $\mathbf{h}_1 = \widehat{\mathbf{h}}_1 + \widetilde{\mathbf{h}}_1$ with $\widetilde{\mathbf{h}}_1$ being independent from $\widehat{\mathbf{h}}_1$ and $\widehat{\mathbf{h}}_2$. We note that, under Assumptions 1 and 2, $\frac{\gamma_1^{\text{ul}}}{M}$ converges to a finite value as $M \to \infty$ [6, App. B]. This ensures that

$$
\frac{\frac{\gamma_1^{\text{ul}}}{M}}{\frac{1}{M} + \frac{\gamma_1^{\text{ul}}}{M}} x_1^{\text{ul}} \asymp x_1^{\text{ul}}. \quad (72)
$$

By applying [6, Lem. 5] to the second term in (70), we obtain

$$
\begin{aligned}
& \frac{1}{\frac{1}{M} + \frac{\gamma_1^{\text{ul}}}{M}} \frac{1}{M} \widehat{\mathbf{h}}_1^{\text{H}} \left(\widehat{\mathbf{h}}_2 \widehat{\mathbf{h}}_2^{\text{H}} + \mathbf{Z} \right)^{-1} \mathbf{h}_2 \\
& = \frac{1}{\frac{1}{M} + \frac{\gamma_1^{\text{ul}}}{M}} \left(\frac{1}{M} \widehat{\mathbf{h}}_1^{\text{H}} \mathbf{Z}^{-1} \mathbf{h}_2 \right. \\
& \left. \qquad - \frac{\frac{1}{M} \widehat{\mathbf{h}}_1^{\text{H}} \mathbf{Z}^{-1} \widehat{\mathbf{h}}_2 \frac{1}{M} \widehat{\mathbf{h}}_2^{\text{H}} \mathbf{Z}^{-1} \mathbf{h}_2}{\frac{1}{M} + \frac{1}{M} \widehat{\mathbf{h}}_2^{\text{H}} \mathbf{Z}^{-1} \widehat{\mathbf{h}}_2} \right). \quad (73)
\end{aligned}
$$

Under Assumption 1 and using [6, Lem. 3], we have that[11], as $M \to \infty$,

$$
\frac{1}{M} \widehat{\mathbf{h}}_1^{\text{H}} \mathbf{Z}^{-1} \mathbf{h}_2 \asymp \frac{1}{M} \widehat{\mathbf{h}}_1^{\text{H}} \mathbf{Z}^{-1} \widehat{\mathbf{h}}_2 \asymp \frac{1}{M} \text{tr}(\mathbf{\Upsilon}_{12} \mathbf{Z}^{-1}) \triangleq \beta_{12} \quad (74)
$$

$$
\frac{1}{M} \widehat{\mathbf{h}}_2^{\text{H}} \mathbf{Z}^{-1} \widehat{\mathbf{h}}_2 \asymp \frac{1}{M} \text{tr}(\mathbf{\Phi}_2 \mathbf{Z}^{-1}) \triangleq \beta_{22}. \quad (75)
$$

Substituting (74) and (75) in (73) and using Assumption 2, we conclude that

$$
\begin{aligned}
\frac{\frac{1}{M}}{\frac{1}{M} + \frac{\gamma_1^{\text{ul}}}{M}} \widehat{\mathbf{h}}_1^{\text{H}} \left(\widehat{\mathbf{h}}_2 \widehat{\mathbf{h}}_2^{\text{H}} + \mathbf{Z} \right)^{-1} \mathbf{h}_2 & \asymp \frac{M}{\gamma_1^{\text{ul}}} \left(\beta_{12} - \frac{\beta_{12} \beta_{22}}{\beta_{22}} \right) \\
& = 0 \quad (76)
\end{aligned}
$$

since $\frac{\gamma_1^{\text{ul}}}{M}$ converges to a finite value as $M \to \infty$ [6, App. B]. For the third term in (70), we have

$$
\begin{aligned}
& \frac{1}{\frac{1}{M} + \frac{\gamma_1^{\text{ul}}}{M}} \frac{1}{M} \widehat{\mathbf{h}}_1^{\text{H}} \left(\widehat{\mathbf{h}}_2 \widehat{\mathbf{h}}_2^{\text{H}} + \mathbf{Z} \right)^{-1} \mathbf{z}^{\text{ul}} \\
& = \frac{1}{\frac{1}{M} + \frac{\gamma_1^{\text{ul}}}{M}} \left(\frac{1}{M} \widehat{\mathbf{h}}_1^{\text{H}} \mathbf{Z}^{-1} \mathbf{z}^{\text{ul}} \right. \\
& \left. \qquad - \frac{\frac{1}{M} \widehat{\mathbf{h}}_1^{\text{H}} \mathbf{Z}^{-1} \widehat{\mathbf{h}}_2 \frac{1}{M} \widehat{\mathbf{h}}_2^{\text{H}} \mathbf{Z}^{-1} \mathbf{z}^{\text{ul}}}{\frac{1}{M} + \frac{1}{M} \widehat{\mathbf{h}}_2^{\text{H}} \mathbf{Z}^{-1} \widehat{\mathbf{h}}_2} \right). \quad (77)
\end{aligned}
$$

Under Assumption 1 and using [6, Lem. 3], we have that, as $M \to \infty$,

$$
\frac{1}{M} \widehat{\mathbf{h}}_1^{\text{H}} \mathbf{Z}^{-1} \mathbf{z}^{\text{ul}} \asymp 0 \quad \text{and} \quad \frac{1}{M} \widehat{\mathbf{h}}_2^{\text{H}} \mathbf{Z}^{-1} \mathbf{z}^{\text{ul}} \asymp 0 \quad (78)
$$

where we have used that $(\widehat{\mathbf{h}}_1, \widehat{\mathbf{h}}_2)$ and $\mathbf{z}^{\text{ul}}$ are independent. Therefore, we have that

$$
\frac{1}{\frac{1}{M} + \frac{\gamma_1^{\text{ul}}}{M}} \frac{1}{M} \widehat{\mathbf{h}}_1^{\text{H}} \left(\widehat{\mathbf{h}}_2 \widehat{\mathbf{h}}_2^{\text{H}} + \mathbf{Z} \right)^{-1} \mathbf{z}^{\text{ul}} \asymp 0. \quad (79)
$$

Combining all the above results, we conclude that $y_1^{\text{ul}} \asymp x_1^{\text{ul}}$. This implies that, as $M \to \infty$, the input-output relation (70) converges to that of a deterministic noiseless channel. The desired result then follows from (6).

REFERENCES

[1] J. Östman, A. Lancho, and G. Durisi, "Short-packet transmission over a bidirectional massive MIMO link," in *Proc. Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, USA, Dec. 2019.

[2] 3GPP, "NR; NR and NG-RAN Overall description; Stage-2," 3rd Generation Partnership Project (3GPP), Technical Specification (TS), 9 2019, version 15.7.0.

[3] ——, "Service requirements for cyber-physical control applications in vertical domains," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 22.104, 12 2019, version 17.2.0.

[4] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[5] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 61, no. 4, pp. 1436–1449, Apr. 2013.

[6] E. Björnson, J. Hoydis, and L. Sanguinetti, "Massive MIMO has unlimited capacity," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 574–590, Jan. 2018.

[7] E. Björnson, L. Sanguinetti, J. Hoydis, and M. Debbah, "Optimal design of energy-efficient multi-user MIMO systems: Is Massive MIMO the answer?" *IEEE Trans. Wireless Commun.*, vol. 14, no. 6, pp. 3059–3075, Jun. 2015.

[8] H. V. Cheng, E. Björnson, and E. G. Larsson, "Optimal pilot and payload power control in single-cell massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 65, no. 9, pp. 2363–2378, May 2017.

[9] E. Björnson, L. Sanguinetti, H. Wymeersch, J. Hoydis, and T. L. Marzetta, "Massive MIMO is a reality—what is next?: Five promising research directions for antenna arrays," *Digital Signal Processing*, vol. 94, pp. 3–20, Nov. 2019.

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

[11] E. Björnson, J. Hoydis, and L. Sanguinetti, "Massive MIMO networks: Spectral, energy, and hardware efficiency," *Foundations and Trends® in Signal Processing*, vol. 11, no. 3-4, pp. 154–655, Nov. 2017.

[12] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, *Fundamentals of massive MIMO*. London, U.K.: Cambridge Univ. Press, 2016.

[13] M. Karlsson, E. Björnsson, and E. G. Larsson, "Performance of in-band transmission of system information in massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 1700–1712, Mar. 2018.

[14] A. Bana, G. Xu, E. D. Carvalho, and P. Popovski, "Ultra reliable low latency communications in massive multi-antenna systems," in *Proc. Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, USA, Oct. 2018, pp. 188–192.

[15] 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.

[16] 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.

[17] 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.

[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.

---

[11]Under Assumption 1, $\mathbf{Q}^{-1} \mathbf{R}_i \mathbf{Z}^{-1} \mathbf{R}_k$ has uniformly bounded spectral norm, which can be proved using in [6, Lem. 4].

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

[20] J. Östman, G. Durisi, E. G. Ström, M. C. Coskun, and G. Liva, "Short packets over block-memoryless fading channels: Pilot-assisted or noncoherent transmission?" *IEEE Trans. Commun.*, vol. 67, no. 2, pp. 1521–1536, Feb. 2019.

[21] J. Zeng, T. Lv, R. P. Liu, X. Su, Y. J. Guo, and N. C. Beaulieu, "Enabling ultra-reliable and low-latency communications under shadow fading by massive MU-MIMO," *IEEE Internet of Things J.*, vol. 7, no. 1, pp. 234–246, Jan. 2020.

[22] H. Ren, C. Pan, Y. Deng, M. Elkashlan, and A. Nallanathan, "Joint pilot and payload power allocation for massive-MIMO-enabled URLLC IIoT networks," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 5, pp. 816–830, May 2020.

[23] A. Martinez and A. Guillén i Fàbregas, "Saddlepoint approximation of random–coding bounds," in *Proc. Inf. Theory Applicat. Workshop (ITA)*, San Diego, CA, USA, Feb. 2011.

[24] W. Feller, *An Introduction to Probability Theory and Its Applications*, 2nd ed., New York, NY, USA, 1971, vol. II.

[25] A. Lancho, J. Östman, G. Durisi, T. Koch, and G. Vazquez-Vilar, "Saddlepoint approximations for short-packet wireless communications," *IEEE Trans. Wireless Commun.*, vol. 19, no. 7, pp. 4831–4846, Jul. 2020.

[26] L. Sanguinetti, E. Björnsson, and J. Hoydis, "Towards massive MIMO 2.0: Understanding spatial correlation, interference suppression, and pilot contamination," *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 232–257, Jan. 2020.

[27] M. Haenggi, "The meta distribution of the SIR in Poisson bipolar and cellular networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 4, pp. 2577–2589, Apr. 2016.

[28] R. G. Gallager, *Information Theory and Reliable Communication*. New York, NY, U.S.A.: John Wiley & Sons, 1968.

[29] E. MolavianJazi, "A unified approach to Gaussian channels with finite blocklength," Ph.D. dissertation, University of Notre Dame, Notre Dame, IN, 2014.

[30] 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.

[31] J. Scarlett, A. Martinez, and A. Guillén i Fàbregas, "Mismatched decoding: Error exponents, second-order rates and saddlepoint approximations," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2647–2666, May 2014.

[32] G. Kaplan and S. Shamai, "Information rates and error exponents of compound channels with application to antipodal signaling in a fading environment," *Arch. Elek. Über*, vol. 47, no. 4, pp. 228–239, 1993.

[33] E. Björnson, L. Sanguinetti, and M. Debbah, "Massive MIMO with imperfect channel covariance information," in *2016 50th Asilomar Conference on Signals, Systems and Computers*, 2016, pp. 974–978.

[34] J. Hoydis, S. ten Brink, and M. Debbah, "Massive MIMO in the UL/DL of cellular networks: How many antennas do we need?" *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 574–590, Feb. 2013.

[35] A. Lancho, J. Östman, G. Durisi, and L. Sanguinetti, "A finite-blocklength analysis for URLLC with massive MIMO," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Montreal, Canada, Jun. 2021.

[36] H. Holm and M.-S. Alouini, "Sum and difference of two squared correlated Nakagami variates in connection with the McKay distribution," *IEEE Trans. Commun.*, vol. 52, no. 8, pp. 1367–1376, Aug. 2004.