IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 28 April 2023, accepted 20 May 2023, date of publication 2 June 2023, date of current version 13 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3282369

RESEARCH ARTICLE

Performance of Quantized Massive MIMO With Fronthaul Rate Constraint Over Quasi-Static Channels

YASAMAN ETTEFAGH[1], (Member, IEEE), SINA REZAEI AGHDAM[1], (Member, IEEE), GIUSEPPE DURISI[1], (Senior Member, IEEE), SVEN JACOBSSON[2], MIKAEL COLDREY[2], AND CHRISTOPH STUDER[3], (Senior Member, IEEE)

[1]Department of Electrical Engineering, Chalmers University of Technology, 412 96 Gothenburg, Sweden
[2]Ericsson Research, 417 56 Gothenburg, Sweden
[3]Department of Information Technology and Electrical Engineering, ETH Zürich, 8092 Zürich, Switzerland

Corresponding author: Giuseppe Durisi (durisi@chalmers.se)

This work was supported in part by the Swedish Foundation for Strategic Research under Grant ID14-0022, and in part by the Swedish Governmental Agency for Innovation Systems (VINNOVA) within the Competence Center ChaseOn.

ABSTRACT We provide a rigorous framework for characterizing and numerically evaluating the error probability achievable in the uplink and downlink of a fully digital quantized multiuser multiple-input multiple-output (MIMO) system. We assume that the system operates over a quasi-static channel that does not change across the finite-length transmitted codewords, and only imperfect channel state information (CSI) is available at the base station (BS) and at the user equipments. The need for the novel framework developed in this paper stems from the fact that, for the quasi-static scenario, commonly used signal-to-interference-and-distortion-ratio expressions that depend on the variance of the channel estimation error are not relatable to any rigorous information-theoretic achievable-rate bound. We use our framework to investigate how the performance of a fully digital massive MIMO system subject to a fronthaul rate constraint, which imposes a limit on the number of samples per second produced by the analog-to-digital and digital-to-analog converters (ADCs and DACs), depends on the number of BS antennas and on the precision of the ADCs and DACs. In particular, we characterize, for a given fronthaul constraint, the trade-off between the number of antennas and the resolution of the data converters, and discuss how this trade-off is influenced by the accuracy of the available CSI. Our framework captures explicitly the cost, in terms of spectral efficiency, of pilot transmission—an overhead that the outage capacity, the classic asymptotic metric used in this scenario, cannot capture. We present extensive numerical results that validate the accuracy of the proposed framework and allow us to characterize, for a given fronthaul constraint, the optimal number of antennas and the optimal resolution of the converters as a function of the transmitted power and of the available CSI.

INDEX TERMS Finite blocklength, fronthaul rate constraint, low-precision converters, multi-user massive multiple-input multiple-output (MIMO), quasi-static scenario.

I. INTRODUCTION

Massive multiple-input multiple-output (MIMO) is a key technology enabler of 5G. Indeed, the large number of active antennas available at the base station (BS) in multiuser massive MIMO architectures results in significant spectral and energy efficiency gains compared to traditional, small-scale MIMO architectures. Furthermore, these gains can be achieved by means of low-complexity spatial processing [2].

In this paper, we focus on fully-digital massive MIMO architectures in which the radio-frequency (RF) frontend, which we will refer to as remote radio head (RRH) and which hosts the RF circuitry, mixers, and data converters, is not colocated with the baseband unit (BBU), in which digital signal processing is performed. Separating these two units is

The associate editor coordinating the review of this manuscript and approving it for publication was Ding Xu.

convenient for accessibility, maintenance, and reconfigurability purposes. We are specifically interested in the scenario in which the RRH is equipped with a massive antenna array.[1] One challenge in such architectures is that the RRH and BBU need to be connected via a finite-capacity fronthaul link—a limitation that is important to take into account when designing massive MIMO systems, especially the ones operating over the large bandwidths available in the millimeter-wave (mm-wave) part of the spectrum. To understand the scale of this interconnect problem, consider for example a massive MIMO base-station (BS) equipped with 100 active antenna branches, each connected to two 16-bit resolution data converters, one for the real and one for the imaginary part of the base-band signal, operating at 1 GS/s. Such an architecture produces 3.2 Tb/s of raw baseband data, which is difficult to transfer using current fronthaul standards. One possible, low-complexity approach to circumvent this issue is to reduce the resolution of the data converters. We are then left with the following natural question. How should one choose the sampling rate, the number of antennas, and the resolution of the data converters, given a constraint on the product of these three quantities, which reflects the fronthaul capacity? In this paper, we will shed light on this question by characterizing the trade-off between the number of antennas and the resolution of the data converters for a fixed sampling rate.

A. PRIOR ART

The problem of designing wireless systems in the presence of a fronthaul constraint has been studied extensively in the context of cloud radio access networks (see, e.g., [4] and references therein). However, the focus of this line of work is on solutions where significant signal-processing capabilities are available at the RRHs, which can then execute sophisticated multiterminal vector compression techniques. In contrast, the focus of this paper is on low-complexity solutions enabling low-cost RRHs. Specifically, we consider a simpler, suboptimal approach for reducing the required fronthaul rate, which involves lowering the precision of the converters at the RRH.

A large body of literature is concerned with the performance achievable with multiuser massive MIMO systems in which the BS is equipped with low-resolution data converters. Existing works include the derivation of information-theoretic achievable rates in the ergodic scenario with Gaussian codebooks [5], [6], [7], [8], [9], [10], [11], the design of channel estimation and data-detection algorithms [12], [13], [14], [15], [16], [17], of linear and nonlinear precoders [8], [18], [19], [20], and of low-complexity equalization techniques [21]. All of these results reveal that satisfactory performance can be achieved even when using 1-bit converters at the RRHs, and that, by using 3-to-5-bit converters, one can approach closely the performance achievable in the infinite-precision case. Extensions of these works to the case of distributed massive MIMO, with focus on the spectral and energy efficiency achievable in the ergodic setting, have been provided in [22], [23], [24], [25], and [26].

The focus of this paper is on the analysis of the rate achievable in the less studied *quasi-static scenario*, in which, differently from the commonly analyzed ergodic setting, the channel remains constant for the duration of each transmitted codeword.[2] This scenario is relevant in propagation conditions with limited time and frequency diversity. It is also relevant when short codewords are transmitted, which occurs in control channels, during the initial-access phase, and in machine-type communications involving stringent latency requirements and limited bandwidth.

For the case in which low-precision data converters are used at the BS, it is crucial to assume that the BS has at its disposal only imperfect channel-state information (CSI), typically acquired in massive MIMO systems via uplink pilot transmission. Indeed, the presence of low-precision converters makes acquiring perfect CSI challenging, even when the number of transmitted pilot symbols is large. As illustrated recently in [27] in the context of short-packet transmissions over infinite-precision massive MIMO links, analyses for the quasi-static case in the presence of imperfect CSI are nontrivial. Indeed, one cannot simply take the ergodic-rate signal-to-interference-and-noise ratio (SINR) expressions reported in, e.g., [28, Thm. 4.1] for the uplink and account for the quasi-static nature of the channel by evaluating the cumulative distribution function of the SINR to determine the outage probability as a function of the transmission rate. This is not correct even in the asymptotic regime of large blocklength, since the resulting expression cannot be related to any rigorous notion of outage probability. Similarly, one cannot insert these ergodic SINR expressions into normal-approximation formulas [29, Eq. 223] to obtain approximations on the rate achievable in the finite-blocklength regime. Unfortunately, both approaches are commonly found in the literature. This highlights the need for a rigorous framework—built from first principles—to analyze the quasi-static scenario.

B. CONTRIBUTIONS

In this paper, we provide such a rigorous framework, and use it to characterize the uplink and downlink packet error probability achievable in the quasi-static scenario, for the case in which a BS, equipped with a large antenna array and low-precision converters, serves in the same time-frequency resources multiple user equipments (UEs). Our framework leverages three fundamental ingredients: (i) the random-coding union bound with parameter s (RCUs) from finite-blocklength information theory [30] to capture the finite length of the transmission packets; (ii) a scaled nearest-neighbor mismatch decoder [31] to account for the imperfect CSI available at the BS and at the UEs, as well

[1] We will not consider the distributed massive MIMO scenario in which the BBU is connected to multiple spatially distributed RRHs [3].

[2] This scenario is a special case of the so called "block-fading" model; specifically, we assume that each codeword is entirely contained within a fading block.

as for the low-complexity, suboptimal processing performed at both transmitters and receivers; (iii) Bussgang's decomposition [32] to deal with the nonlinearity introduced by the low-resolution converters. These ingredients have been used separately in the literature before. However, their combination, which is required to address the key design question formulated in this paper, is novel and nontrivial. We also show how to approximate the obtained error probability bounds with simpler asymptotic expressions that are in terms of the so-called generalized mutual information (GMI).[3] The resulting approximation is shown to be very accurate for packet error probabilities in the range $[10^{-3}, 10^{-1}]$.

We then use the packet error-probability bounds developed in the paper to obtain engineering insights into the optimal design of multiuser massive MIMO systems with low-precision converters, operating under a fronthaul constraint. Specifically, focusing on a realistic clustered channel model, and considering the quantization-aware channel-estimation algorithm proposed in [13], we determine, for a given fronthaul constraint, the optimal number of antennas and the resolution of the quantizers that maximize the rate at which uplink and downlink communications can be sustained with a packet error probability not exceeding 10%. For the parameters considered in our numerical simulations, which pertain a scenario with 8 users and a fronthaul constraint of 512 bit/s/Hz, our analysis reveals that the highest performing BS architecture involves a large antenna array (from 64 to 256 antenna elements, depending on the SNR) connected to low-precision data converters (from 1 to 4 bits, depending on the SNR). A solution involving 1-bit data converters and 256 antennas turns out to be optimal from a bi-directional communication perspective for low transmitted-power levels, whereas, for high transmitted-power levels, a solution involving 4-bit data converters and 64 antennas is preferable. Increasing the precision of the converters beyond these values, at the cost of a reduction in the number of antennas, turns out to be deleterious, once the impact of imperfect CSI is accounted for. Interestingly, these conclusions are different from the ones derived by performing a perfect-CSI analysis. Indeed, in the perfect-CSI case, the use of architectures involving higher-precision quantizers and fewer antennas is preferable.

C. PAPER OUTLINE

In Section II, we introduce the system model, the fronthaul constraint, and the linear spatial processing that will be considered in the rest of the paper. In Section III, we present our nonasymptotic framework for the characterization of the packet error probability, as well as asymptotic limits and approximations that will be useful to obtain system-design guidelines. In Section IV, we describe our numerical-simulation setup and conduct experiments to determine (i) the impact of a fronthaul constraint on the channel-estimation accuracy obtainable via pilot transmission, (ii) the optimal number of pilot symbols, and (iii) the effect on performance of nonsubtractive dithering in the 1-bit quantization case. We then shed light on the fronthaul-induced trade-off between number of antennas at the BS and resolution of the quantizers. Some concluding remarks are provided in Section V.

[3]This quantity was previously used in performance analyses of low-precision massive MIMO architectures in, e.g., [10].

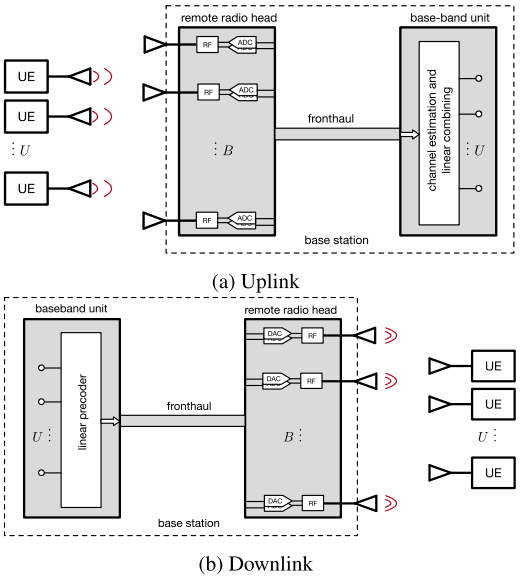


FIGURE 1. Overview of the fully digital BS architecture considered in the paper. A *B*-antenna BS serves *U* users over the same time-frequency resources. Each antenna is connected to a pair of quantizers with *Q*-bit resolution. The BS consists of a BBU and an RRH that are connected via a rate-constrained fronthaul interface.

D. NOTATION

Lower-case bold letters are used for vectors and upper-case bold letters for matrices. We denote by $\mathcal{CN}(\mathbf{0}_N, \mathbf{R})$, where $\mathbf{0}_N$ stands for the all-zero vector of size N, the distribution of an N-dimensional circularly-symmetric complex-valued Gaussian vector with zero mean and $N \times N$ covariance matrix $\mathbf{R}$. We use $\mathbf{I}_N$ to denote the $N \times N$ identity matrix, and $\mathbb{E}[\cdot]$, $\mathbb{V}\mathrm{ar}[\cdot]$, $\mathbb{P}[\cdot]$ to denote the expectation, variance, and probability operators, respectively. The natural logarithm is denoted by $\log(\cdot)$, the Gaussian Q-function by $Q_{\mathrm{G}}(\cdot)$, the indicator function by $\mathbb{1}\{\cdot\}$, and the floor function by $\lfloor\cdot\rfloor$. Finally, the notation $f(n) = \mathcal{O}(g(n))$, $n \to \infty$ means that $\limsup_{n\to\infty} |f(n)/g(n)| < \infty$.

II. SYSTEM MODEL

We consider a single-cell massive multiuser MIMO scenario, in which a BS equipped with B antennas, serves U single-antenna UEs in the same time-frequency resources. As depicted in Fig. 1, the BS consists of a RRH and a BBU that are connected via a rate-constrained fronthaul interface. Each antenna is equipped with a pair of Q-bit data converters, one for the in-phase and one for the quadrature component. We consider a time division duplexing (TDD) scenario. In the uplink, the data-transmission phase is preceded by a pilot-transmission phase, which allows the BS to acquire

(imperfect) CSI. The signal transmitted by the U UEs is quantized at the B BS antennas using the low-precision data converters. The quantized signal is then transferred to the BBU via the fronthaul link, where channel estimation, linear combining, and decoding are performed. In the downlink, the linearly precoded signal is quantized at the BBU and transferred over the fronthaul link, where it is converted into the analog domain and transmitted over the B antennas. It follows that an architecture with B active antennas and Q-bit converters, operating at the Nyquist sampling rate, requires a fronthaul interface able to support a rate of $2BQ$ bit/s/Hz.

We assume uniform, symmetric, mid-rise quantizers with step size Δ and Q-bit resolution. Specifically, let $r \in \mathbb{R}$ be the input of the quantizer. Then, the output $\mathcal{Q}(r)$ is given by

$$\mathcal{Q}(r) = \begin{cases} \frac{\Delta}{2}(1-L), & \text{if } r < -\frac{\Delta}{2}L \\ \Delta\left\lfloor \frac{r}{\Delta} \right\rfloor + \frac{\Delta}{2}, & \text{if } -\frac{\Delta}{2}L \le r < \frac{\Delta}{2}L \\ \frac{\Delta}{2}(L-1), & \text{if } r \ge \frac{\Delta}{2}L. \end{cases} \tag{1}$$

Here, $L = 2^Q$ denotes the number of quantization levels. For a complex-valued input z, we let $\mathcal{Q}(z) = \mathcal{Q}(\Re\{z\}) + j\mathcal{Q}(\Im\{z\})$. For a vector $\mathbf{z}$, we denote by $\mathcal{Q}(\mathbf{z})$ the result of applying $\mathcal{Q}(\cdot)$ entry-wise to its elements.

Note that we assume for simplicity that all converters have the same resolution. However, our analysis can be readily extended to the scenario considered in, e.g., [10], in which the converters may have different resolution.

A. UPLINK TRANSMISSION

We consider a TDD transmission protocol in which an uplink frame consisting of n_{ul} channel uses is followed by a downlink frame of n_{dl} channel uses. The fading process is assumed to stay constant over the duration of the $n_{\text{ul}} + n_{\text{dl}}$ channel uses. Furthermore, we assume that reciprocity holds, so that the channel estimated in the uplink can be used by the BS in downlink transmission.

In the uplink, we model the B-dimensional discrete-time, complex-valued, base-band signal received at the BS at time instant k as follows:

$$\mathbf{y}^{\text{ul}}[k] = \mathbf{H}\mathbf{s}^{\text{ul}}[k] + \mathbf{n}^{\text{ul}}[k], \quad k = 1, \dots, n_{\text{ul}}. \tag{2}$$

Here, $\mathbf{s}^{\text{ul}}[k] = \left[s_1^{\text{ul}}[k], s_2^{\text{ul}}[k], \dots, s_U^{\text{ul}}[k]\right]^T \in \mathbb{C}^U$ is the signal transmitted by the U UEs at time instant k, the $B \times U$ matrix $\mathbf{H}$ represents the fading channel, and $\mathbf{n}^{\text{ul}}[k] \sim \mathcal{CN}(\mathbf{0}_B, N_0\mathbf{I}_B)$ denotes the additive white Gaussian noise at the BS, which we assume to be independent across k, and independent also of the transmitted signal and the fading matrix. The first n_{p} channel uses in the uplink are reserved for the transmission of pilot symbols, used by the BS to estimate $\mathbf{H}$. The remaining $n_{\text{d}} = n_{\text{ul}} - n_{\text{p}}$ channel uses are reserved for data transmission. Note that, so far, we have not provided any statistical model for the fading channel $\mathbf{H}$. This is because the information theoretic bounds we shall provide in Section III hold for arbitrary quasi-static fading models.

At the receiver, the analog signal from which $\mathbf{y}^{\text{ul}}[k]$ is obtained, is passed through an automatic gain control (AGC) circuit, which scales the analog signal so as to match the dynamic range of the quantizer. Then, a linear combiner $\mathbf{W} \in \mathbb{C}^{B\times U}$, which is computed by the BS on the basis of the CSI acquired via the n_{p} pilot symbols, is used to obtain an estimate $\hat{\mathbf{s}}^{\text{ul}}[k] \in \mathbb{C}^U$ of the transmitted signal $\mathbf{s}^{\text{ul}}[k]$ on the basis of the quantizer output. Mathematically, we have the following model:

$$\hat{\mathbf{s}}^{\text{ul}}[k] = \mathbf{W}^H \mathcal{Q}\Big(\mathbf{A}\mathbf{y}^{\text{ul}}[k]\Big), \quad k = n_{\text{p}}+1, \dots, n_{\text{ul}}. \tag{3}$$

Here, the diagonal matrix $\mathbf{A}$ models the AGC operation. Note that we have not specified how pilot transmission is performed or which channel estimator and linear combiner are used. Again, this is because the information-theoretic bounds we shall provide in Section III hold for arbitrary pilot transmission schemes and channel estimators.

B. DOWNLINK TRANSMISSION

The acquired CSI in the uplink phase is used by the BS to compute the linear precoder $\mathbf{P}$. The resulting precoded signal is then passed through a Q-bit quantizer to satisfy the fronthaul-rate requirements. As a consequence, the U-dimensional discrete-time received signal at the UEs can be modeled as follows:

$$\mathbf{y}^{\text{dl}}[k] = \mathbf{H}^T \alpha \mathcal{Q}\Big(\mathbf{P}\mathbf{s}^{\text{dl}}[k]\Big) + \mathbf{n}^{\text{dl}}[k] \tag{4}$$

for $k = n_{\text{ul}} + 1, \dots, n_{\text{ul}} + n_{\text{dl}}$. Here, $\mathbf{s}^{\text{dl}}[k] = \left[s_1^{\text{dl}}[k], \dots, s_U^{\text{dl}}[k]\right]^T$ contains the signal intended to each of the U UEs, and the vector $\mathbf{n}^{\text{dl}}[k] \sim \mathcal{CN}(\mathbf{0}_U, N_0\mathbf{I}_U)$ is the AWGN at the UEs' side. This vector is independent across k and does not depend on the transmitted signal or the fading matrix. In (4), the parameter α is a normalization factor used to enforce the power constraint $\mathbb{E}\left[\|\alpha\mathcal{Q}\big(\mathbf{P}\mathbf{s}^{\text{dl}}[k]\big)\|^2\right] = \rho^{\text{dl}}$. Note that when the resolution of the quantizer is very low (e.g., for a 1-bit quantizer), significant throughput gains can be achieved by adopting more sophisticated *nonlinear* precoders, which depend on the transmitted data (see, e.g., [8], [19], [20]). In this paper, we focus on linear precoders because they are the de-facto standard in commercial massive MIMO BS. Similarly to the uplink, the information-theoretic bounds we shall provide in Section III hold for an arbitrary linear precoder $\mathbf{P}$. We shall assume that the UEs are equipped with high-resolution converters. Hence, we will not model the quantization distortion at the UEs in uplink and downlink.

III. ANALYSIS OF THE ACHIEVABLE ERROR PROBABILITY

We will now provide a nonasymptotic, i.e., finite-blocklength, upper bound on the error probability achievable for a given transmission rate in both the uplink and the downlink for the system model described in Section II. From this result, one can directly obtain a lower bound on the achievable rates for a given target error probability. Indeed, in the finite blocklength regime, there is a fundamental trade-off between

packet error probability and rate: reducing the error probability comes at the cost of a reduction in the transmission rate, which is made explicit by our bounds. Our derivation is based on an information-theoretic random-coding argument. Specifically, we will provide a characterization of the average error probability, averaged over the so-called i.i.d. Gaussian codebook ensemble, in which each symbol of the transmitted codewords is generated independently from the same zero-mean Gaussian distribution. The main components of our analysis are the mismatch-decoding framework [33] and the RCUs from finite-blocklength information theory [30]. Both tools will be described in the following sections.

A. PRELIMINARY RESULT

As a preliminary result, we shall state the desired bound on the error probability for the following simpler infinite-precision, single-antenna, single-UE nonfading channel model:

$$v[k] = gq[k] + z[k], \quad k = 1, \dots, n. \tag{5}$$

Here, g is a deterministic complex-valued coefficient and $\{z[k]\}_{k=1}^{n}$ is a sequence of i.i.d. $\mathcal{CN}(0, \sigma^2)$ random variables. Note that we allow σ^2 to depend on g. This will turn out to be important to apply the bound on the error probability obtained when analyzing (5) to the input-output relations of interest in this paper, i.e., (3) and (4). The bound reported in this section is derived under the following crucial assumptions:

(i) The receiver has access to a noisy estimate $\hat{g}$ of g, which the receiver treats as perfect. Specifically, since the additive noise is Gaussian, the receiver operates according to the so-called scaled-nearest neighbor principle [31], i.e., it seeks the n-dimensional transmitted codeword $[\hat{q}[1], \dots, \hat{q}[n]]^T$ that, after scaling by $\hat{g}$, is closest to the receiver vector $[v[1], \dots, v[n]]^T$ in Euclidean norm.

(ii) The average packet error probability is averaged over the ensemble of Gaussian i.i.d. codebooks. Specifically, the input signals $q[k]$ in (5) are drawn independently from a $\mathcal{CN}(0, \rho)$ distribution. Here, ρ denotes the transmit power.

Under these assumptions, as proven for example in [27, Thm. 1], one can establish the existence of a coding scheme with rate R and packet error probability $\epsilon = \mathbb{P}\Big[\big[\hat{q}[1], \dots, \hat{q}[n]\big]^T \neq \big[q[1], \dots, q[n]\big]^T\Big]$ that is upper-bounded by

$$\epsilon \leq \inf_{s>0} \mathbb{P}\left[\frac{\log(f)}{n} + \frac{1}{n}\sum_{k=1}^{n} \imath_s(q[k], v[k]) \leq R\right]. \tag{6}$$

Here, f is a random variable that is uniformly distributed on the interval $[0, 1]$ and

$$\imath_s(q[k], v[k]) = -s\left|v[k] - \hat{g}q[k]\right|^2 + \frac{s\left|v[k]\right|^2}{1 + s\rho\left|\hat{g}\right|^2} + \log\left(1 + s\rho\left|\hat{g}\right|^2\right) \tag{7}$$

is the so-called *generalized information density* [30]. The bound in (6) is an instantiation (for a given channel model and a given mismatch-decoding rule) of a more general bound, commonly referred to as RCUs [30]. This bound is, in turn, a relaxation and generalization to mismatch decoding of the random-coding union bound proposed in [29, Thm. 16]. The bound in (6) is optimized over the parameter $s > 0$, which originates from the Chernoff-bound step used to relax the random-coding union bound. Note, though, that any choice of $s > 0$ results in a valid (although potentially looser) bound.

B. LINEARIZATION VIA BUSSGANG'S THEOREM

One obstacle in the direct application of the bound (7) to the uplink and downlink channel input-output-relations (3) and (4) is the presence of the nonlinear operator $\mathcal{Q}(\cdot)$, which prevents the direct use of the mismatch-decoding framework. Indeed, the mismatch-decoding operation that results in the information density (7) relies on the linearity of (5). Bussgang's theorem [32], which has been used extensively in the massive MIMO literature to analyze the impact of hardware impairments [6], [11], [34], [35], provides a simple approach to overcome this issue. Specifically, Bussgang's theorem yields a simple way to compute the correlation between two Gaussian vectors, after one of the two vectors is passed through a nonlinearity (in our case, the quantization operation (1)). This theorem, combined with a standard linear minimum mean square error (LMMSE) decomposition, allows us to obtain the desired linearization.

1) UPLINK

Let us start by considering the uplink input-output relation after spatial combining given in (3). Throughout, we shall assume, in agreement with what stated in Section III-A, that the input signals $s_u^{\mathrm{ul}}[k]$, $u = 1, \dots, U$, $k = n_{\mathrm{p}} + 1, \dots, n_{\mathrm{ul}}$, are drawn independently from a $\mathcal{CN}(0, \rho^{\mathrm{ul}})$ distribution, where ρ^{ul} denotes the uplink transmit power, which we assume being the same for all UEs. Assume that the channel matrix $\mathbf{H}$ is fixed. It is convenient to write the output $\mathbf{r}^{\mathrm{ul}}[k] = \mathcal{Q}\big(\mathbf{A}\mathbf{y}^{\mathrm{ul}}[k]\big)$ of the quantizer as the sum of the LMMSE estimate of $\mathbf{r}^{\mathrm{ul}}[k]$ given the input $\mathbf{A}\mathbf{y}^{\mathrm{ul}}[k]$ of the quantizer, plus the uncorrelated, non-Gaussian estimation error $\mathbf{d}^{\mathrm{ul}}[k]$ as follows:

$$\mathbf{r}^{\mathrm{ul}}[k] = \mathbf{G}^{\mathrm{ul}}\mathbf{y}^{\mathrm{ul}}[k] + \mathbf{d}^{\mathrm{ul}}[k] \quad k = n_{\mathrm{p}} + 1, \dots, n_{\mathrm{ul}}. \tag{8}$$

Here, $\mathbf{G}^{\mathrm{ul}}$ is the LMMSE-filter matrix. Since the input $\mathbf{A}\mathbf{y}^{\mathrm{ul}}[k]$ of the quantizer is conditionally Gaussian given the channel matrix $\mathbf{H}$, this filter takes on a particularly simple form. Specifically, it follows from Bussgang's theorem that $\mathbf{G}^{\mathrm{ul}}$ is diagonal and given by [18]

$$\mathbf{G}^{\mathrm{ul}} = \frac{\Delta}{\sqrt{\pi}} \operatorname{diag}\big(\mathbf{A}\mathbf{C}_{\mathbf{y}^{\mathrm{ul}}}\mathbf{A}\big)^{-1/2} \times \sum_{i=1}^{L-1} \exp\big(-\Delta^2(i - L/2)^2 \operatorname{diag}\big(\mathbf{A}\mathbf{C}_{\mathbf{y}^{\mathrm{ul}}}\mathbf{A}\big)^{-1}\big) \tag{9}$$

where $\mathbf{C}_{\mathbf{y}^{\mathrm{ul}}} = \mathbb{E}\left[\mathbf{y}^{\mathrm{ul}}(\mathbf{y}^{\mathrm{ul}})^H\right]$. Substituting (9) into (8) and then (8) into (3), we obtain the desired linearized input-output relation, to which we can apply the error-probability bound (6).

2) DOWNLINK

We shall assume that the downlink input signals $\mathbf{s}^{\mathrm{dl}}[k]$ are drawn independently from a $\mathcal{CN}(\mathbf{0}_U, \mathbf{I}_U)$ distribution. Assume that the precoding matrix $\mathbf{P}$ is fixed. Since $\mathbf{P}\mathbf{s}^{\mathrm{dl}}[k]$ is conditionally Gaussian given $\mathbf{P}$, it follows from Bussgang's theorem that (4) can be equivalently expressed as

$$\mathbf{y}^{\mathrm{dl}}[k] = \mathbf{H}^T \alpha \left(\mathbf{G}^{\mathrm{dl}}\mathbf{P}\mathbf{s}^{\mathrm{dl}}[k] + \mathbf{d}^{\mathrm{dl}}[k]\right) + \mathbf{n}^{\mathrm{dl}}[k] \tag{10}$$

for $k = n_{\mathrm{ul}} + 1, \ldots, n_{\mathrm{ul}} + n_{\mathrm{dl}}$, where $\mathbf{d}^{\mathrm{dl}}[k]$ is the non-Gaussian quantization noise, which is uncorrelated with $\mathbf{s}^{\mathrm{dl}}[k]$, and where the LMMSE-filter matrix $\mathbf{G}^{\mathrm{dl}}$ has the same form as $\mathbf{G}^{\mathrm{ul}}$ in (9), with $\mathbf{A}\mathbf{C}_{\mathbf{y}^{\mathrm{ul}}}\mathbf{A}$ replaced by $\mathbf{P}\mathbf{P}^H$.

C. THE ACTUAL ERROR-PROBABILITY BOUND

1) UPLINK

We let $\hat{s}_u^{\mathrm{ul}}[k]$ denote the uth entry of the vector $\hat{\mathbf{s}}^{\mathrm{ul}}[k]$ in (3), and the vectors $\mathbf{w}_u$ and $\mathbf{h}_u$ denote the uth columns of the $B \times U$ combining matrix $\mathbf{W}$ and channel matrix $\mathbf{H}$. Substituting (8) into (3), we can write the estimate $\hat{s}_u^{\mathrm{ul}}[k]$ of the kth data symbol from user u as

$$\hat{s}_u^{\mathrm{ul}}[k] = \mathbf{w}_u^H\mathbf{G}^{\mathrm{ul}}\mathbf{A}\mathbf{h}_u s_u^{\mathrm{ul}}[k] + \sum_{\substack{v=1\\v\neq u}}^{U}\mathbf{w}_u^H\mathbf{G}^{\mathrm{ul}}\mathbf{A}\mathbf{h}_v s_v^{\mathrm{ul}}[k] + \mathbf{w}_u^H\mathbf{G}^{\mathrm{ul}}\mathbf{A}\mathbf{n}^{\mathrm{ul}}[k] + \mathbf{w}_u^H\mathbf{d}^{\mathrm{ul}}[k] \tag{11}$$

for $k = n_{\mathrm{p}} + 1, \ldots, n_{\mathrm{ul}}$ and $u = 1, \ldots, U$. The first term in (11) denotes the useful signal. The remaining terms comprise the residual multiuser interference, the additive noise, and the quantization noise. The BS is not aware of the effective channel gain $g_u^{\mathrm{ul}} = \mathbf{w}_u^H\mathbf{G}^{\mathrm{ul}}\mathbf{A}\mathbf{h}_u$. However, it can use the n_{p} pilot symbols to obtain the estimate $\hat{g}^{\mathrm{ul}} = \mathbf{w}_u^H\mathbf{G}^{\mathrm{ul}}\mathbf{A}\hat{\mathbf{h}}_u$, where $\hat{\mathbf{h}}_u$ denotes the uth column of the channel estimate matrix $\widehat{\mathbf{H}}$.

Since $\mathbf{H}$, and, hence $\widehat{\mathbf{H}}$, are assumed to stay constant over the entire transmission duration, which involves $n_{\mathrm{ul}} + n_{\mathrm{dl}}$ channel uses, we can obtain a mismatch-decoding upper bound on the per-user error probability ϵ_u^{ul}, $u = 1, \ldots, U$, by applying (6) to the linearized input-output relation (11) for each realization of $\mathbf{H}$ and $\widehat{\mathbf{H}}$ and then by averaging over $\mathbf{H}$ and $\widehat{\mathbf{H}}$. Specifically, by setting $q[k] = s_u^{\mathrm{ul}}[k]$, $v[k] = \hat{s}_u^{\mathrm{ul}}[k]$, $g = g_u^{\mathrm{ul}}$, $\hat{g} = \hat{g}_u^{\mathrm{ul}}$, and $\rho = \rho^{\mathrm{ul}}$ in (6) and (7), and by accounting for the pilot-transmission overhead, we can upper-bound the uplink per-user error probability as

$$\epsilon_u^{\mathrm{ul}} \leq \mathbb{E}_{\mathbf{H},\widehat{\mathbf{H}}}\left[\inf_{s>0}\mathbb{P}\left[\frac{\log(f)}{n_{\mathrm{ul}}} + \frac{1}{n_{\mathrm{ul}}}\sum_{k=n_{\mathrm{p}}+1}^{n_{\mathrm{ul}}}\imath_s(s_u^{\mathrm{ul}}[k], \hat{s}_u^{\mathrm{ul}}[k]) \leq R \,\middle|\, \mathbf{H}, \widehat{\mathbf{H}}\right]\right]. \tag{12}$$

In (12), the probability inside the expectation is computed with respect to the transmitted symbols, the additive noise, and the uniform random variable f.

2) DOWNLINK

We let $s_u^{\mathrm{dl}}[k]$, $n_u^{\mathrm{dl}}[k]$, and $y_u^{\mathrm{dl}}[k]$ denote the uth entry of the symbol vector $\mathbf{s}^{\mathrm{dl}}[k]$, noise vector $\mathbf{n}^{\mathrm{dl}}[k]$, and received vector $\mathbf{y}^{\mathrm{dl}}[k]$, respectively. Furthermore, let $\mathbf{p}_u$ denote the uth column of the precoding matrix $\mathbf{P}$. It then follows from (10) that

$$y_u^{\mathrm{dl}}[k] = \alpha\mathbf{h}_u^T\mathbf{G}^{\mathrm{dl}}\mathbf{p}_u s_u^{\mathrm{dl}}[k] + \sum_{\substack{v=1\\v\neq u}}^{U}\alpha\mathbf{h}_u^T\mathbf{G}^{\mathrm{dl}}\mathbf{p}_v s_v^{\mathrm{dl}}[k] + \alpha\mathbf{h}_u^T\mathbf{d}^{\mathrm{dl}}[k] + n_u^{\mathrm{dl}}[k] \tag{13}$$

for $k = n_{\mathrm{ul}} + 1, \ldots, n_{\mathrm{ul}} + n_{\mathrm{dl}}$, and $u = 1, \ldots, U$. We assume that the uth UE is not aware of the effective channel $g_u^{\mathrm{dl}} = \alpha\mathbf{h}_u^T\mathbf{G}^{\mathrm{dl}}\mathbf{p}_u$ but it is aware of its mean $\hat{g}_u^{\mathrm{dl}} = \alpha\,\mathbb{E}\left[\mathbf{h}_u^T\mathbf{G}^{\mathrm{dl}}\mathbf{p}_u\right]$. This setup is often considered in the massive MIMO literature and yields, in the asymptotic ergodic setting, the so-called hardening bound [28, Thm. 4.6]. In our quasi-static setup, we will treat $\hat{g}_u^{\mathrm{dl}}$ simply as the imperfect CSI used by the scaled-nearest neighbor decoder at the uth UE, hence providing an operational interpretation to the hardening bound. By setting $q[k] = s_u^{\mathrm{dl}}[k]$, $v[k] = y_u^{\mathrm{dl}}[k]$, $g = g_u^{\mathrm{dl}}$, $\hat{g} = \hat{g}_u^{\mathrm{dl}}$, and $\rho = 1$ in (6) and (7), we can upper-bound the downlink per-user error probability as

$$\epsilon_u^{\mathrm{dl}} \leq \mathbb{E}_{\mathbf{H},\widehat{\mathbf{H}}}\left[\inf_{s>0}\mathbb{P}\left[\frac{\log(f)}{n_{\mathrm{dl}}} + \frac{1}{n_{\mathrm{dl}}}\sum_{k=n_{\mathrm{ul}}+1}^{n_{\mathrm{ul}}+n_{\mathrm{dl}}}\imath_s(s_u^{\mathrm{dl}}[k], y_u^{\mathrm{dl}}[k]) \leq R \,\middle|\, \mathbf{H}, \widehat{\mathbf{H}}\right]\right]. \tag{14}$$

The probability in (14) is computed with respect to the symbols transmitted by the BS and the additive noise. This probability depends on the channel estimate $\widehat{\mathbf{H}}$ indirectly through the precoding matrix $\mathbf{P}$.

D. ASYMPTOTIC LIMITS AND USEFUL APPROXIMATIONS

The error-probability bounds provided in (12) and (14) are difficult to evaluate. Indeed not only the expectations, but also the conditional probability terms within the bounds can in general not be obtained in closed form, and need to be evaluated numerically, which is especially challenging if one targets low error probabilities. Furthermore, the minimization over s, which is required to tighten the bounds, needs also to be performed numerically for each realization of $\mathbf{H}$ and $\widehat{\mathbf{H}}$.[4]

As we shall discuss next, it turns out that, when one operates in the moderate error-probability regime (i.e., packet error probability in the range $[10^{-3}, 10^{-1}]$), one can obtain asymptotic approximations on the error-probability bounds in (12) and (14) that are much simpler to evaluate numerically.

[4]In practice, one can loosen the bounds by moving the minimization outside the expectation, which alleviates somewhat the numerical complexity of this optimization step.

After demonstrating their accuracy, we will use these approximations in Section IV to provide insights on the optimal system design.

1) A NORMAL APPROXIMATION

To introduce the first approximation, we start with the uplink bound (12). Note that, given $\mathbf{H}$ and $\widehat{\mathbf{H}}$, the conditional probability term in (12) involves the linear combination of $n_{\mathrm{ul}} - n_{\mathrm{p}} + 1$ independent random variables (the $n_{\mathrm{ul}} - n_{\mathrm{p}}$ information-density terms, which are actually also identically distributed, and the $\log(f)$ term). Fix an arbitrary integer $k \in [n_{\mathrm{p}} + 1, n_{\mathrm{ul}}]$. It will turn out convenient to let

$$I_{s,u}^{\mathrm{ul}} = \mathbb{E}\left[\imath_s(s_u^{\mathrm{ul}}[k], \hat{s}_u^{\mathrm{ul}}[k])\right] \tag{15}$$

$$= -s\left(\left|g_u^{\mathrm{ul}} - \hat{g}_u^{\mathrm{ul}}\right|^2 \rho^{\mathrm{ul}} + \sigma_{u,\mathrm{ul}}^2\right) + s\frac{\rho^{\mathrm{ul}}\left|g_u^{\mathrm{ul}}\right|^2 + \sigma_{u,\mathrm{ul}}^2}{1 + s\rho^{\mathrm{ul}}\left|\hat{g}_u^{\mathrm{ul}}\right|^2} + \log\left(1 + s\rho^{\mathrm{ul}}\left|\hat{g}_u^{\mathrm{ul}}\right|^2\right) \tag{16}$$

where the expectation in (15) is computed only with respect to the transmitted symbols and the additive noise (i.e., $\mathbf{H}$ and $\widehat{\mathbf{H}}$ are fixed). In (16), we let $\sigma_{u,\mathrm{ul}}^2$ denote the conditional variance, given $\mathbf{H}$ and $\widehat{\mathbf{H}}$, of the total additive noise in (11), which includes also residual multiuser interference and quantization noise. Specifically, we have that

$$\sigma_{u,\mathrm{ul}}^2 = \sum_{\substack{v=1 \\ v\neq u}}^{U}\left|\mathbf{w}_u^H\mathbf{G}^{\mathrm{ul}}\mathbf{A}\mathbf{h}_v\right|^2 + N_0\|\mathbf{w}_u^H\mathbf{G}^{\mathrm{ul}}\mathbf{A}\|^2 + \mathbf{w}_u^H\mathbf{C}_{\mathbf{d}^{\mathrm{ul}}}\mathbf{w}_u \tag{17}$$

where $\mathbf{C}_{\mathbf{d}^{\mathrm{ul}}} = \mathbb{E}\left[\mathbf{d}^{\mathrm{ul}}\left(\mathbf{d}^{\mathrm{ul}}\right)^H\right]$ denotes the correlation matrix of the uplink quantization distortion. To obtain (17) we have used that the transmitted symbols are independent across users and that the quantization noise and the transmitted symbols are uncorrelated as a consequence of the LMMSE decomposition. The random variable $I_{s,u}^{\mathrm{ul}}$ in (16) is usually referred to as the GMI.

Let us also set $V_{s,u}^{\mathrm{ul}} = \mathbb{V}\mathrm{ar}\left[\imath_s(s_u^{\mathrm{ul}}[k], \hat{s}_u^{\mathrm{ul}}[k])\right]$. This quantity is a generalization to the mismatch-decoding setup of the so-called *channel dispersion* in finite-blocklength information theory (see, e.g., [29, Def. 1]).

Let us now assume that $n_{\mathrm{ul}} - n_{\mathrm{p}} \gg 1$. Since the random variable $\log(f)$ has finite moments, we can approximate the error probability in (12) using the Berry-Essen central-limit theorem [36, Ch. XVI.5] and conclude that

$$\mathbb{P}\left[\frac{\log(f)}{n_{\mathrm{ul}}} + \frac{1}{n_{\mathrm{ul}}}\sum_{k=n_{\mathrm{p}}+1}^{n_{\mathrm{ul}}}\imath_s(s_u^{\mathrm{ul}}[k], \hat{s}_u^{\mathrm{ul}}[k]) \le R \,\middle|\, \mathbf{H}, \widehat{\mathbf{H}}\right] = Q_{\mathrm{G}}\left(\frac{\left(1 - \frac{n_{\mathrm{p}}}{n_{\mathrm{ul}}}\right)I_{u,s}^{\mathrm{ul}} - R}{\sqrt{\left(1 - \frac{n_{\mathrm{p}}}{n_{\mathrm{ul}}}\right)\frac{V_{u,s}^{\mathrm{ul}}}{n_{\mathrm{ul}}}}}\right) + \mathcal{O}\left(\frac{1}{\sqrt{n_{\mathrm{ul}} - n_{\mathrm{p}}}}\right). \tag{18}$$

We shall refer to the approximation on the conditional error probability on the left-hand side of (18) obtained by neglecting the $\mathcal{O}\left(1/\sqrt{n_{\mathrm{ul}} - n_{\mathrm{p}}}\right)$ term on the right-hand side of (18) as *normal approximation*.

A similar approximation can be derived for the conditional probability term in the downlink error probability bound provided in (14). Specifically, let[5] for an arbitrary integer $k \in [n_{\mathrm{ul}} + 1, n_{\mathrm{ul}} + n_{\mathrm{dl}}]$

$$I_{s,u}^{\mathrm{dl}} = \mathbb{E}\left[\imath_s(s_u^{\mathrm{dl}}[k], y_u^{\mathrm{dl}}[k])\right] \tag{19}$$

$$= -s\left(\left|g_u^{\mathrm{dl}} - \hat{g}_u^{\mathrm{dl}}\right|^2 + \sigma_{u,\mathrm{dl}}^2\right) + s\frac{\left|g_u^{\mathrm{dl}}\right|^2 + \sigma_{u,\mathrm{dl}}^2}{1 + s\left|\hat{g}_u^{\mathrm{dl}}\right|^2} + \log\left(1 + s\left|\hat{g}_u^{\mathrm{dl}}\right|^2\right) \tag{20}$$

where

$$\sigma_{u,\mathrm{dl}}^2 = \sum_{v\neq u}\alpha^2\left|\mathbf{h}_u^T\mathbf{G}^{\mathrm{dl}}\mathbf{p}_v\right| + \alpha^2\mathbf{h}_u^T\mathbf{C}_{\mathbf{d}^{\mathrm{dl}}}\mathbf{h}_u^* + N_0 \tag{21}$$

with $\mathbf{C}_{\mathbf{d}^{\mathrm{dl}}} = \mathbb{E}\left[\mathbf{d}^{\mathrm{dl}}\left(\mathbf{d}^{\mathrm{dl}}\right)^H\right]$. Let $V_{s,u}^{\mathrm{dl}} = \mathbb{V}\mathrm{ar}\left[\imath_s(s_u^{\mathrm{dl}}[k], y_u^{\mathrm{dl}}[k])\right]$. Then,

$$\mathbb{P}\left[\frac{\log(f)}{n_{\mathrm{dl}}} + \frac{1}{n_{\mathrm{dl}}}\sum_{k=n_{\mathrm{ul}}+1}^{n_{\mathrm{ul}}+n_{\mathrm{dl}}}\imath_s(s_u^{\mathrm{dl}}[k], y_u^{\mathrm{dl}}[k]) \le R \,\middle|\, \mathbf{H}, \widehat{\mathbf{H}}\right] = Q_{\mathrm{G}}\left(\frac{I_{u,s}^{\mathrm{dl}} - R}{\sqrt{\frac{V_{u,s}^{\mathrm{dl}}}{n_{\mathrm{dl}}}}}\right) + \mathcal{O}\left(\frac{1}{\sqrt{n_{\mathrm{dl}}}}\right). \tag{22}$$

2) THE OUTAGE-PROBABILITY LIMIT

Neglecting the $\mathcal{O}(\cdot)$ terms in (18) and (22) and then substituting (18) into (12) and (22) into (14), one obtains approximations on the uplink and downlink error probabilities that are accurate when $n_{\mathrm{ul}} - n_{\mathrm{p}} \gg 1$ and $n_{\mathrm{dl}} \gg 1$, and easier to evaluate, since the probability term is given in closed form.

However, the resulting expressions are still challenging to evaluate numerically. Indeed, the first issue is that both $V_{u,s}^{\mathrm{ul}}$ and $V_{u,s}^{\mathrm{dl}}$ depend on the expectation of the product of powers of the input signal and the quantization distortion in the LMMSE decomposition of the quantized signal. These terms do not admit, in general, an analytical characterization. The second issue is that, even after the above-mentioned substitutions, the optimization over s, which is needed to tighten the bounds, cannot be performed analytically.

To avoid both of these issues, we next present an alternative, looser, asymptotic approximation in terms of outage probability. Starting from the uplink, we assume that $n_{\mathrm{ul}} \to \infty$ and that $\lim_{n_{\mathrm{ul}}\to\infty} n_{\mathrm{p}}/n_{\mathrm{ul}} = p \in [0, 1]$, where p is the rate

[5] Note that (20) depends on the transmit power ρ^{dl} indirectly, through the normalization parameter α, which appears in the definitions of g_u^{dl}, $\hat{g}_u^{\mathrm{dl}}$, and $\sigma_{u,\mathrm{dl}}^2$.

penalty due to pilot transmission. It follows from (18) that

$$\lim_{n_{\mathrm{ul}}\to\infty} \mathbb{P}\left[\frac{\log(f)}{n_{\mathrm{ul}}}+\frac{1}{n_{\mathrm{ul}}}\sum_{k=n_{\mathrm{p}}+1}^{n_{\mathrm{ul}}} \imath_s(s_u^{\mathrm{ul}}[k],\hat{s}_u^{\mathrm{ul}}[k]) \le R \,\middle|\, \mathbf{H},\widehat{\mathbf{H}}\right]$$
$$= \begin{cases} 1, & \text{if } (1-p)I_{u,s}^{\mathrm{ul}} < R \\ 0, & \text{otherwise} \end{cases} \tag{23}$$
$$= \mathbb{1}\{(1-p)I_{u,s}^{\mathrm{ul}} < R\}. \tag{24}$$

Using (24) in (12), we conclude that

$$\lim_{n_{\mathrm{ul}}\to\infty} \mathbb{E}_{\mathbf{H},\widehat{\mathbf{H}}}\left[\inf_{s>0} \mathbb{P}\left[\frac{\log(f)}{n_{\mathrm{ul}}} + \frac{1}{n_{\mathrm{ul}}}\sum_{k=n_{\mathrm{p}}+1}^{n_{\mathrm{ul}}} \imath_s(s_u^{\mathrm{ul}}[k],\hat{s}_u^{\mathrm{ul}}[k]) \le R \,\middle|\, \mathbf{H},\widehat{\mathbf{H}}\right]\right]$$
$$= \mathbb{E}\left[\inf_{s>0} \mathbb{1}\{(1-p)I_{u,s}^{\mathrm{ul}} < R\}\right] \tag{25}$$
$$= \mathbb{P}\left[(1-p)\left(\sup_{s>0} I_{u,s}^{\mathrm{ul}}\right) < R\right]. \tag{26}$$

We shall refer to (26) as uplink GMI-based outage bound. It turns out that the maximization over $s > 0$ in (26) can be performed analytically. Specifically, let s_{opt} the value of s that maximizes the GMI $I_{u,s}^{\mathrm{ul}}$. Then, proceeding similar to [31, App. A], one can show that

$$s_{\mathrm{opt}} = \frac{-2c+b+\sqrt{b^2+4ac}}{2bc} \tag{27}$$

where $a = \rho^{\mathrm{ul}}\left|g_u^{\mathrm{ul}}\right|^2 + \sigma_{u,\mathrm{ul}}^2$, $b = \rho^{\mathrm{ul}}\left|\hat{g}_u^{\mathrm{ul}}\right|^2$, and $c = \rho^{\mathrm{ul}}\left|g_u^{\mathrm{ul}} - \hat{g}_u^{\mathrm{ul}}\right|^2 + \sigma_{u,\mathrm{ul}}^2$.

It is worth noting that in the perfect CSI case, in which $\hat{g}_u^{\mathrm{ul}} = g_u^{\mathrm{ul}}$, we have that $a = b + c$. Using this equality in (27), we obtain that

$$s_{\mathrm{opt}} = \frac{1}{c} = \frac{1}{\sigma_{u,\mathrm{ul}}^2}. \tag{28}$$

This implies that, in the perfect CSI case,

$$\sup_{s>0} I_{u,s}^{\mathrm{ul}} = \log\left(1+\frac{\rho^{\mathrm{ul}}\left|g_u^{\mathrm{ul}}\right|^2}{\sigma_{u,\mathrm{ul}}^2}\right) \tag{29}$$

and (26) reduces to the familiar outage-probability formula

$$\mathbb{P}\left[\log\left(1+\frac{\rho^{\mathrm{ul}}\left|g_u^{\mathrm{ul}}\right|^2}{\sigma_{u,\mathrm{ul}}^2}\right) < R\right]. \tag{30}$$

This provides further evidence that (26) is the natural extension of (30) to the case of imperfect CSI. It is worth stressing that, in the imperfect CSI case, the expression for the outage probability obtained by substituting in (26) the optimal value of s given in (28), does not take the form given in (30), with the SINR in (30) replaced by a SINR term including also the variance of the channel estimation error. This means that the ergodic SINR expression for the imperfect CSI case reported in, e.g., [28, Thm. 4.1] should not be used in the quasi-static setting considered in this paper.

With steps similar to the ones leading to (26), one can show that, in the downlink,

$$\lim_{n_{\mathrm{dl}}\to\infty} \mathbb{E}_{\mathbf{H},\widehat{\mathbf{H}}}\left[\inf_{s>0} \mathbb{P}\left[\frac{\log(f)}{n_{\mathrm{dl}}} + \frac{1}{n_{\mathrm{dl}}}\sum_{k=n_{\mathrm{ul}}+1}^{n_{\mathrm{ul}}+n_{\mathrm{dl}}} \imath_s(s_u^{\mathrm{dl}}[k], y_u^{\mathrm{dl}}[k]) \le R \,\middle|\, \mathbf{H},\widehat{\mathbf{H}}\right]\right]$$
$$= \mathbb{P}\left[\left(\sup_{s>0} I_{u,s}^{\mathrm{dl}}\right) < R\right] \tag{31}$$

where the value of s maximizing $I_{u,s}^{\mathrm{dl}}$ is given by (27), with $a = \left|g_u^{\mathrm{dl}}\right| + \sigma_{u,\mathrm{dl}}^2$, $b = \left|\hat{g}_u^{\mathrm{dl}}\right|^2$, and $c = \left|g_u^{\mathrm{dl}} - \hat{g}_u^{\mathrm{dl}}\right|^2 + \sigma_{u,\mathrm{dl}}^2$. We shall refer to (31) as downlink GMI-based outage bound. Similar to the uplink, if we assume that the UEs have perfect knowledge of the effective channel g_u^{dl}, then

$$\sup_{s>0} I_{u,s}^{\mathrm{dl}} = \log\left(1+\frac{\left|g_u^{\mathrm{dl}}\right|^2}{\sigma_{u,\mathrm{dl}}^2}\right) \tag{32}$$

and (31) reduces to the familiar outage formula

$$\mathbb{P}\left[\log\left(1+\frac{\left|g_u^{\mathrm{dl}}\right|^2}{\sigma_{u,\mathrm{dl}}^2}\right) < R\right]. \tag{33}$$

IV. NUMERICAL RESULTS

We present numerical simulations to demonstrate the accuracy of the approximations on the error-probability bounds presented in Section III. We will then use these approximations—more specifically (26) for the uplink and (31) for the downlink—to investigate the trade-off between the number of antennas and the resolution of the converters in the fully digital massive MIMO architecture described in Section II. We will also study how this trade-off is influenced by the quality of the available CSI. Before presenting our numerical experiments, we detail in the next section the scenario that will be considered throughout, as well as the system parameters and the algorithm used for channel estimation.

A. SIMULATION SETUP

1) PROPAGATION SCENARIO

We consider a small-cell scenario where a BS, which is equipped with a uniform linear array of B equispaced antennas, serves $U = 8$ users, uniformly distributed within a disc centered around the BS, with inner radius of 5 meters and outer radius of 150 meters. The propagation channel between each UE and the BS is modeled according to a standard clustered channel model (previously considered within the context of massive MIMO architectures with low precision quantizers in, e.g., in [37], [38], and [39]). This channel model involves N_{cl} clusters, with each cluster contributing to N_{rmc} resolvable multipath components. According to this model, each column $\mathbf{h}_u$, $u = 1,\ldots,U$, of the channel matrix $\mathbf{H}$ can

be written as

$$\mathbf{h}_u = \sqrt{\frac{1}{N_{\text{cl}} N_{\text{rmc}}}} \sum_{n=1}^{N_{\text{cl}}} \sum_{m=1}^{N_{\text{rmc}}} \alpha_{n,m}^u \mathbf{a}(\theta_{n,m}^u). \tag{34}$$

Here, the fading coefficients $\alpha_{n,m}^u$ are i.i.d. $\mathcal{CN}(0, \sigma_u^2)$ complex random variables, with σ_u^2 modeling the path loss experienced by the uth UE. Furthermore, $\mathbf{a}(\theta_{n,m}^u)$ is the array response vector of the uniform linear array at the BS in far field

$$\mathbf{a}(\theta_{n,m}^u) = \left[1, e^{-j2\pi\theta_{n,m}^u}, \ldots, e^{-j2\pi(B-1)\theta_{n,m}^u}\right]^T \tag{35}$$

where $\theta_{n,m}^u = d \sin(\phi_{n,m}^u)/\lambda$, with d being the antenna spacing, $\phi_{n,m}^u$ the angle of arrival or spatial angle measured from the boresight of the uniform linear array, and λ is the wavelength.

In our simulations, we assume $N_{\text{cl}} = 2$ and $N_{\text{rmc}} = 4$, and fix the antenna spacing to $\lambda/2$. As pathloss model, we assume that $10 \log_{10} \sigma_u^2 = -72 - 29.2 \log_{10}(d_u/d_0)$, where d_u denotes the distance in meters between the uth UE and the BS and $d_0 = 1$ m. The angle of arrival $\phi_{n,m}^u$ is modeled as $\phi_{n,m}^u = \phi_n^u + \phi_{\text{offset}}$, where $\phi_n^u \sim \mathcal{U}[-\pi/3, \pi/3]$ and $\phi_{\text{offset}} \sim \mathcal{U}[-\pi/24, \pi/24]$.

2) SYSTEM PARAMETERS

The noise spectral density N_0 is assumed to be -174 dBm/Hz, the carrier frequency is 30 GHz, and the transmitted signal has a bandwidth equal to 50 MHz. In the channel estimation phase, we assume that each user transmits concurrently orthogonal pilot sequences, obtained by cyclically shifting a Zadoff-Chu sequence [13] of length n_{p}. Let $\mathbf{T} \in \mathbb{C}^{U \times n_{\text{p}}}$ with $\mathbf{T}\mathbf{T}^H = n_{\text{p}} \rho^{\text{ul}} \mathbf{I}_U$ be the matrix containing the pilot symbols transmitted by all UEs. Let the $B \times n_{\text{p}}$ received signal at the BS during the pilot phase be

$$\mathbf{Y}^{\text{p}} = \mathbf{H}\mathbf{T} + \mathbf{N}^{\text{p}} \tag{36}$$

where $\mathbf{N}^{\text{p}} \in \mathbb{C}^{B \times n_{\text{p}}}$ denotes the additive noise. We use the signal received in the pilot phase to determine the AGC diagonal matrix $\mathbf{A}$ when computing $\hat{g}_u^{\text{ul}}$. Specifically, we set

$$\mathbf{A} = \text{diag}\left(\widehat{\mathbf{C}}_{\mathbf{y}^{\text{ul}}}\right)^{-1/2} \tag{37}$$

where

$$\widehat{\mathbf{C}}_{\mathbf{y}^{\text{ul}}} = \frac{1}{n_{\text{p}}} \mathbf{Y}^{\text{p}} (\mathbf{Y}^{\text{p}})^H. \tag{38}$$

We also use $\widehat{\mathbf{C}}_{\mathbf{y}^{\text{ul}}}$ as an estimate of $\mathbf{C}_{\mathbf{y}^{\text{ul}}}$ when evaluating the Bussgang filter $\mathbf{G}^{\text{ul}}$ in the computation of $\hat{g}_u^{\text{ul}}$.

To set the parameter Δ in (1), we treat the input of the quantizer as a complex Gaussian random variable of zero mean and unit variance and we require that the clipping probability does not exceed 10^{-4}. This yields

$$\Delta = \frac{\sqrt{2}}{L} Q_{\text{G}}^{-1}\left(\frac{10^{-4}}{2}\right). \tag{39}$$

Let now, $\tau_i = \Delta(i - L/2)$ for $i = 1, \ldots, L-1$ and $\tau_0 = -\infty$, $\tau_L = \infty$ denote the quantization thresholds. Furthermore, let $\ell_i = \Delta(i - L/2 + 1/2)$ for $i = 0, \ldots, L-1$ be the quantization labels. Treating again the input of the quantizer as a complex Gaussian random variable, we set the downlink power-normalization parameter α in (4) to

$$\alpha = \frac{\sqrt{\rho^{\text{dl}}/(2B)}}{\sqrt{\sum_{i=0}^{L-1} \ell_i^2 \left(Q_{\text{G}}(\sqrt{2}\tau_i) - Q_{\text{G}}(\sqrt{2}\tau_{i+1})\right)}}. \tag{40}$$

With this choice of α, we ensure that the average power constraint $\mathbb{E}\left[\|\alpha \mathcal{Q}(\mathbf{Ps}^{\text{dl}}[k])\|^2\right] = \rho^{\text{dl}}$ is satisfied for the case in which the entries of $\mathbf{Ps}^{\text{dl}}$ are modeled as $\mathcal{CN}(0, 1)$ random variables. For the 1-bit case, we compute the correlation of the uplink and downlink quantization distortion $\mathbf{C}_{\mathbf{d}^{\text{ul}}}$ and $\mathbf{C}_{\mathbf{d}^{\text{dl}}}$ using the so called arc-sine law [40] (see [18, Eqs. (34) and (43)]). When $Q > 1$, since no closed-form expressions for $\mathbf{C}_{\mathbf{d}^{\text{ul}}}$ and $\mathbf{C}_{\mathbf{d}^{\text{dl}}}$ are available, we use the diagonal approximation proposed in [18, Sec. IV.C].

Throughout this section, we set $n_{\text{ul}} = n_{\text{dl}} = 500$ and assume that the fronthaul interface can support a rate no larger than 512 bit/s/Hz. For a 50 MHz transmitted-signal bandwidth, this implies a fronthaul rate of about 25.6 Gbit/s, which is in the range of what can be supported with current technologies. This constraint implies that $2BQ \leq 512$ bit/s/Hz. As a consequence, the largest number of BS antennas that is compatible with the use of quantizers with $Q = 1, 2, 3, 4, 5, 6, 7, 8$ bits of resolution is $B = 256, 128, 85, 64, 51, 42, 36, 32$, respectively.

3) CHANNEL ESTIMATION

To estimate the channel, we use the CSI-acquisition algorithm proposed in [13]. In this algorithm, the clustered channel generated according to (34) is estimated in the angle domain. Since in this representation the channel is approximately sparse, the channel estimation problem can be reduced to a quantized compressive-sensing reconstruction problem. The approach followed in [13] is to solve this problem using a method that combines expectation maximization with approximate message passing.

4) LINEAR COMBINER AND PRECODER

The available CSI at the BS is used to construct a linear combiner and a linear precoder. Throughout this section, we will consider the distortion-aware MMSE combiner proposed in [7, Eq. 13] for which

$$\mathbf{w}_u = \left(\rho^{\text{ul}} \left(\sum_{v \neq u} \mathbf{G}^{\text{ul}} \mathbf{A} \hat{\mathbf{h}}_v \left(\mathbf{G}^{\text{ul}} \mathbf{A} \hat{\mathbf{h}}_v\right)^H\right) + N_0 \mathbf{G}^{\text{ul}} \mathbf{A} (\mathbf{G}^{\text{ul}} \mathbf{A})^H + \mathbf{C}_{\mathbf{d}^{\text{ul}}}\right)^{-1} \left(\rho^{\text{ul}} \mathbf{G}^{\text{ul}} \mathbf{A} \hat{\mathbf{h}}_u\right) \tag{41}$$

and a MMSE precoder (which ignores quantization effects)

$$\mathbf{P} = \beta \widehat{\mathbf{H}}^* \left(\widehat{\mathbf{H}}^{\text{T}} \widehat{\mathbf{H}}^* + \frac{UN_0}{\rho^{\text{dl}}} \mathbf{I}_{\text{U}}\right)^{-1} \tag{42}$$

where the normalization factor β is chosen so that $\mathbb{E}\left[\|\mathbf{Ps}^{\mathrm{dl}}[k]\|^2\right] = B$. Some remarks on the performance achievable with the simpler maximum ratio combiner in the uplink and maximum ratio precoder in the downlink are provided in Section IV-F.

B. CHANNEL-ESTIMATION PERFORMANCE

Before analyzing the trade-off between the number of antennas and the resolution of the converters for a given fronthaul constraint in Section IV-F, we start by providing in this section some insights on the choice of the pilot-sequence length for the clustered channel model (34) and the channel-estimation algorithm described in Section IV-A3. We will then discuss in Section IV-C the accuracy of the simple-to-evaluate error-probability approximations provided in (26) and (31), and analyze in Section IV-D the impact of dithering in both uplink and downlink for the case $Q = 1$. We start by investigating the normalized MSE of the channel estimate[6] obtained at the BS as a function of the resolution Q of the data converters for the case in which $U = 8$ users transmit simultaneously pilot sequences of length $n_{\mathrm{p}} = 48$ as described in Section IV-A2. As we shall discuss in Section IV-E, for the error-probability values considered therein, this choice for n_{p} strikes a good balance between accuracy of the channel estimate and pilot overhead. Note that, because of the fronthaul constraint, the number of BS antennas decreases as Q is increased, as described in Section IV-A2. To investigate the impact of quantization on the accuracy of the channel estimates obtainable at the BS, we consider two values of transmit power: a low transmit-power value of 16 dBm, and a high transmit-power value of 24 dBm.[7] The grey lines in Fig 2 illustrate the channel-estimation normalized MSE as a function of Q, for a fixed number of transmit antennas $B \in \{32, 36, 42, 51, 64, 85, 128, 256\}$. The MSE values marked in blue are the ones corresponding to the largest number of antennas that is compatible with the fronthaul constraint, for the corresponding value of Q. We note that for the low transmit-power value, the lowest MSE value among the (Q, B) pairs satisfying the fronthaul constraint is obtained when $Q = 2$ and $B = 128$, whereas, for the high transmit-power value, $Q = 4$ and, hence, $B = 64$ result in the lowest MSE. Increasing the transmit power turns out beneficial in terms of MSE for all Q values except $Q = 1$, for which the MSE deteriorates as the transmit power is increased. This behavior is common in the 1-bit case and has been noticed before for a variety of channel models and channel-estimation algorithms [1], [12], [41], [42], [43]. To shed further light on this phenomenon, we plot in Fig. 3 the channel-estimation normalized MSE for $(Q, B) \in \{(1, 256), (2, 128), (3, 85)\}$. Note that the MSE curve decreases monotonically with ρ^{ul} when $Q = 2, 3$. On the contrary, when $Q = 1$, the MSE curve achieves a global minimum at $\rho^{\mathrm{ul}} \approx 20$ dBm. The reason is as follows: although a single-bit quantizer preserves only the sign of the input signal, with a sufficient amount of noise, amplitude information about the input signal can be recovered via multiple measurements. This well-known phenomenon, usually referred to as stochastic resonance, can be enforced also in the high-SNR regime via the use of nonsubtractive dithering at the receiver, prior to quantization [44]. We will investigate the beneficial effects of dithering in the high-SNR regime for the case $Q = 1$ in Section IV-D. Note that the two transmit-power values chosen in Fig. 2 are to the left and to the right of the transmit-power value minimizing the normalized MSE in Fig. 3, i.e., $\rho^{\mathrm{ul}} = 20$ dBm.

[6]This quantity is defined as the average of the ratio between the square of the channel estimation error and the square of the norm of the channel.

[7]As we shall clarify shortly, these two values allow us to analyze the impact of dithering for the 1-bit quantization case.

C. ACCURACY OF THE PROPOSED LARGE-BLOCKLENGTH APPROXIMATIONS

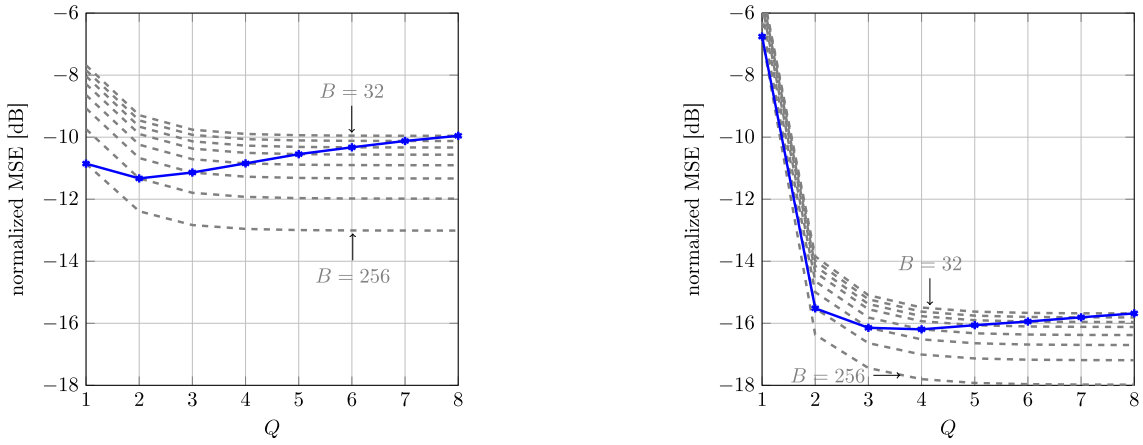
In Section III-D, we proposed two large-blocklength approximations to the RCUs bounds (12) and (14). Focusing on the uplink, we shall now discuss the accuracy of these approximations. Specifically, we compare the error probability bound given by (i) the RCUs bound (12), (ii) the normal approximation (18), (iii) the GMI-based outage probability (26) (computed for $p = n_{\mathrm{p}}/n_{\mathrm{ul}}$), and (iv) the perfect-CSI outage probability (30). In Fig. 4, we present this comparison for the case $Q = 1$, $n_{\mathrm{ul}} = 500$, and $U = 8$. Specifically, we report the error probability as predicted by the different bounds/approximations, versus the number of transmit antennas B when each user transmits at a rate R of 0.5 bit/s/Hz. As in Section IV-B, we consider two values for the transmit power. The reported error probability bounds are optimized over the number of transmitted pilots n_{p}. Furthermore, the bound (12) and the normal approximation (18) are optimized over the parameter s. The optimization over the parameter s and the number of pilot symbols is performed via a grid search.

As shown in the figure, for the range of error probabilities considered here, the predictions obtained using the RCUs bound (12) and the normal approximation (18) match the ones obtained using the GMI-based outage probability (16). On the contrary, the predictions based on the perfect-CSI outage-probability approximation provided in (30) turn out to highly underestimate the error probability. The optimal number of pilot symbols in Fig. 4 is around 48 for higher values of the error probability, and increases to around 56 for error probability values around 10^{-2} and below. A similar result holds for the other values of quantizer resolution Q considered in this section as well as for the downlink.

Based on these results, we shall use the GMI-based outage bounds (26) and (31) as performance metrics when conducting the system-optimization investigations reported in the following two sections.

D. IMPACT OF DITHERING

We now focus on the case $Q = 1$ and investigate the impact of dithering on both uplink and downlink performance. Specifically, we set $B = 256$, $n_{\mathrm{p}} = 48$, and investigate the impact of dithering on the GMI-based outage bounds (26)



(a) $\rho^{\mathrm{ul}} = 16$ dBm (b) $\rho^{\mathrm{ul}} = 24$ dBm

FIGURE 2. Channel estimation normalized MSE as a function of the resolution of the quantizers and the number of BS antennas. The points marked in blue are obtained by considering the largest number of BS antennas that is compatible with a fronthaul constraint of 512 bit/s/Hz.

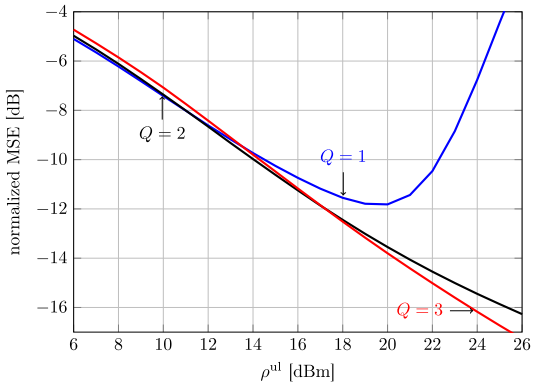


FIGURE 3. Channel-estimation normalized MSE as a function of ρ^{ul} for the (Q, B) pairs $\{(1, 256), (2, 128), (3, 85)\}$ satisfying the fronthaul constraint of 512 bit/s/Hz.

and (31). The other system parameters are as in the previous sections. Dithering is only used in the channel-estimation phase. Indeed, for the parameters considered in this section, dithering in the data-transmission phase does not yield any benefits. The reason is that the residual multiuser interference after linear spatial processing acts as dithering and is sufficient to induce stochastic resonance. We model dithering by assuming that the smallest channel-estimation MSE achievable for the case $Q = 1$, which is achieved by transmitting pilot symbols at a power of around 20 dBm (see Fig. 3) can be maintained for all values of $\rho^{\mathrm{ul}} \geq 20$ dBm. We also assume that the uplink operates at a much lower power than the downlink. Specifically, the channel estimate used to generate the downlink precoder is obtained via a pilot-transmission uplink phase in which the pilot symbols are transmitted at a power level that is 26 dB less than the downlink power ρ^{dl}.

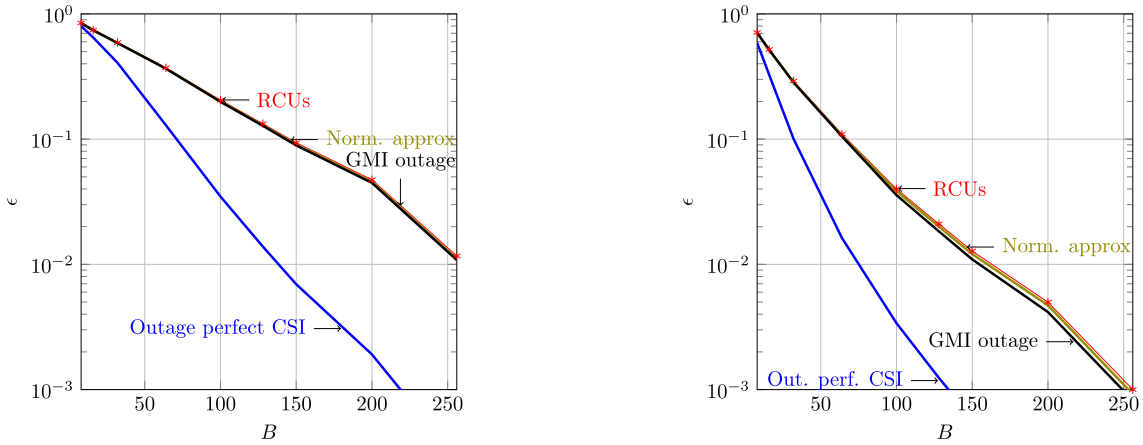
In Fig. 5, we depict the maximum achievable rate compatible with a GMI-based outage probability not exceeding 0.1. As shown in the figure, dithering in the channel estimation phase is beneficial in the uplink. Indeed, without dithering the rate drops rapidly as the transmit power is increased beyond 20 dBm, whereas, with dithering, the achievable rate does not decrease with the transmit power, apart from a small rate reduction at 21 dBm, which is the power level at which dithering is first introduced. Interestingly, using dithering in the channel-estimation phase has no benefits in the downlink, for the range of transmitted-power values considered in the figure. The reason is as follows: although, both in the uplink and in the downlink, the decoder operates according to the scaled nearest-neighbor principle, the scaling parameter in the two setups is different. In the uplink, we use the scaling parameter $\hat{g}_u^{\mathrm{ul}} = \mathbf{w}_u^H \mathbf{G}^{\mathrm{ul}} \mathbf{A} \hat{\mathbf{h}}_u$ whereas in the downlink we use the hardening-bound-inspired scaling parameter $\hat{g}_u^{\mathrm{dl}} = \alpha\, \mathbb{E}\left[\mathbf{h}_u^T \mathbf{G}^{\mathrm{dl}} \mathbf{p}_u\right]$. It turns out that the explicit dependence of $\hat{g}^{\mathrm{ul}}$ on $\hat{\mathbf{h}}_u$ makes this bound sensitive to the lack of stochastic resonance occurring when ρ^{ul} exceeds 20 dBm.

E. OPTIMAL NUMBER OF PILOT SYMBOLS

We consider again as performance metric the maximum rate that is compatible with a GMI-based uplink outage probability not exceeding 0.1, and investigate the optimal number of pilot symbols for different values of Q and B chosen so as to satisfy the fronthaul constraint. In Table 1 and 2, we report, for both $\rho^{\mathrm{ul}} = 16$ dBm and $\rho^{\mathrm{ul}} = 24$ dBm, the optimal number of pilots as well as the rate penalty incurred when setting $n_{\mathrm{p}} = 48$, which is the value we considered in Figs. 2, 3, and 5. As shown in the tables, this rate penalty is negligible for $\rho^{\mathrm{ul}} = 24$ dBm. Indeed, for this transmit-power value the rate curve as a function of the number of pilot symbols is flat around its maximum. The rate penalty is also small for $\rho^{\mathrm{ul}} = 16$ dBm for small Q values, but it increases for larger Q values.

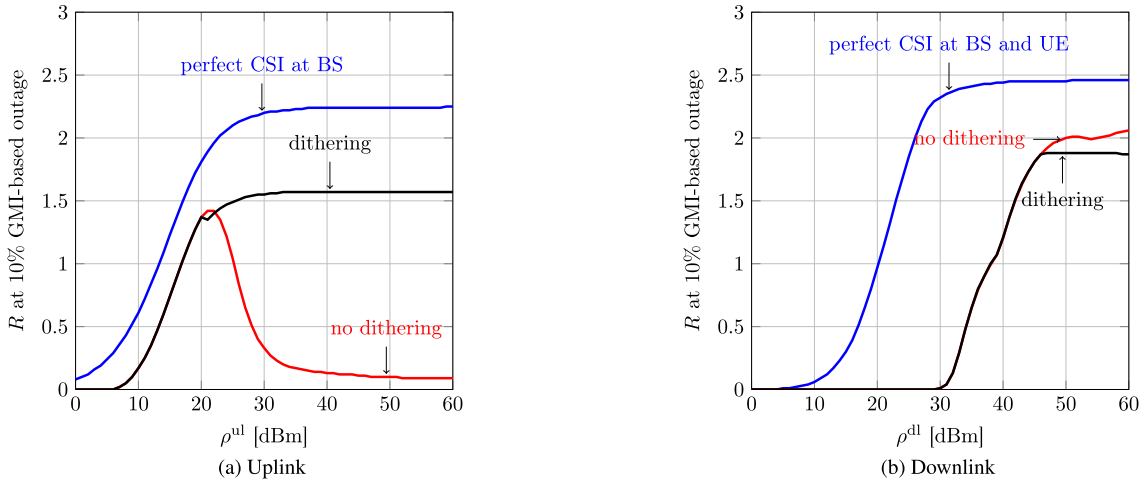
F. NUMBER OF ANTENNAS VS. DATA-CONVERTER RESOLUTION

Finally, we investigate how one should select the number of antennas and the resolution of the quantizers to maximize



(a) $\rho^{\mathrm{ul}} = 16$ dBm (b) $\rho^{\mathrm{ul}} = 24$ dBm

FIGURE 4. Comparison between the proposed bounds and approximations on the uplink packet-error probability achievable for the case $Q = 1$, $U = 8$, $n_{\mathrm{ul}} = 500$, and $R = 0.5$ bit/s/Hz. The error probability curves are optimized over the number of transmitted pilots.



(a) Uplink (b) Downlink

FIGURE 5. Impact of dithering during the channel-estimation phase on the uplink and downlink performance.

TABLE 1. Optimal value of n_{p} and rate penalty when n_{p} is set to 48: $\rho^{\mathrm{ul}} = 16$ dBm.

(Q, B)	$(1, 256)$	$(2, 128)$	$(3, 85)$	$(4, 64)$	$(5, 51)$	$(6, 42)$	$(7, 38)$	$(8, 32)$
$n_{\mathrm{p}}^{\star}$	64	60	60	60	128	128	128	128
Rate pen. [%]	4.34	4.41	4.83	4.83	5.88	14.70	17.85	20.83

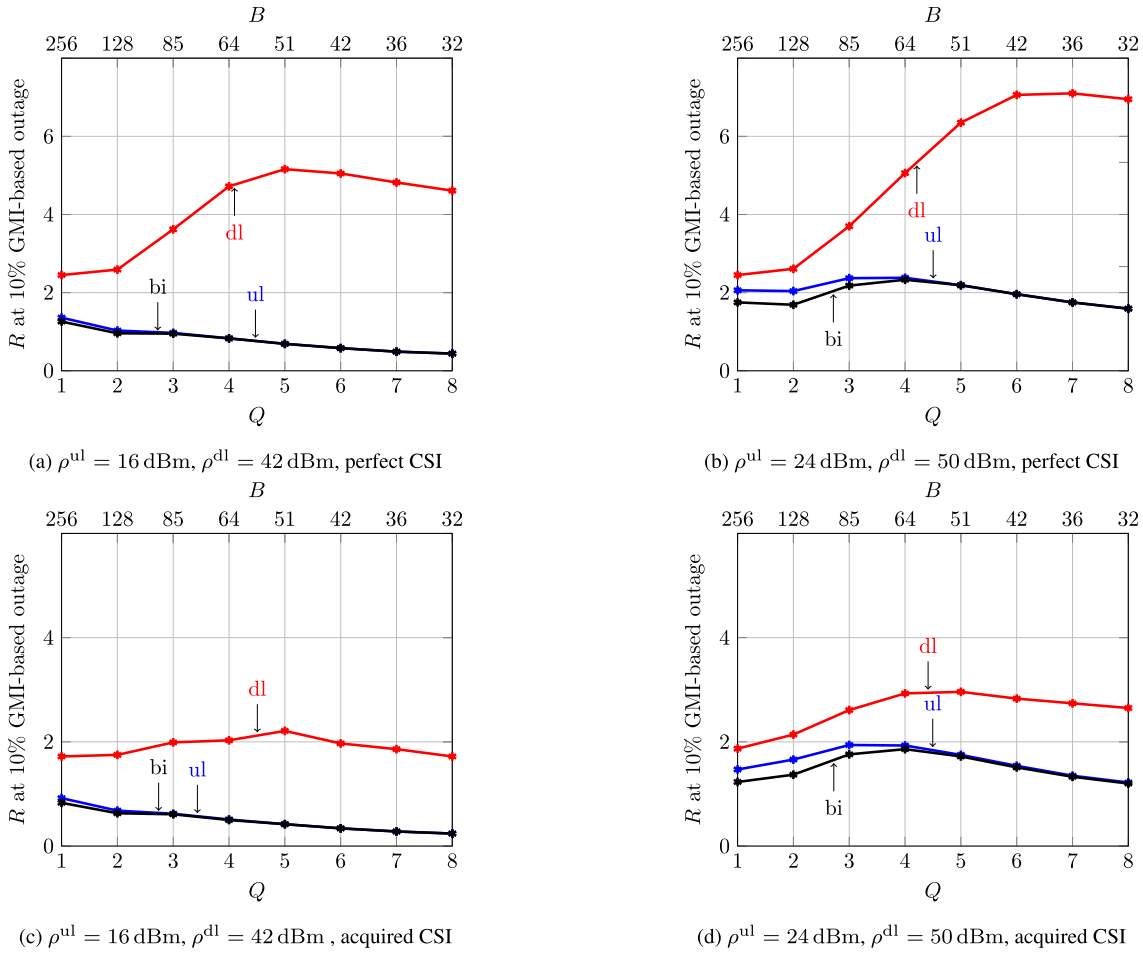
the uplink and downlink rates given a GMI-based outage constraint of 0.1, hence addressing the central question that motivated our investigation. In Fig. 6, we report the uplink and the downlink rates for the pair $\rho^{\mathrm{ul}} = 16\,\mathrm{dBm}$, $\rho^{\mathrm{dl}} = 42\,\mathrm{dBm}$, as well as for the pair $\rho^{\mathrm{ul}} = 24\,\mathrm{dBm}$, $\rho^{\mathrm{dl}} = 50\,\mathrm{dBm}$, as a function of the resolution Q of the quantizers, for a fronthaul rate of 512 bit/s/Hz. Motivated by machine-type communications where an uplink data-collection phase is followed by the transmission of a control command on the downlink, we also report the bi-directional rate, which we define as the largest rate R for which the bi-directional outage probability

$$\mathbb{P}\left[\min\left\{(1-p)\left(\sup_{s>0} I_{u,s}^{\mathrm{ul}}\right), \left(\sup_{s>0} I_{u,s}^{\mathrm{dl}}\right)\right\} < R\right] \qquad (43)$$

does not exceed 0.1. The uplink rates are optimized over the choice of n_{p}; the channel estimates obtained using the resulting number of pilots is used to determine the downlink precoder. For the case $Q = 1$, dithering in the channel-estimation phase is introduced whenever beneficial. We see from the figure that, in the perfect-CSI case (Figs. 6a and 6b), the system is uplink-limited and the bi-directional rate curve

TABLE 2. Optimal value of n_{p} and rate penalty when n_{p} is set to 48: $\rho^{\mathrm{ul}} = 24$ dBm; dithering is used for the pair $(1, 256)$.

(Q, B)	$(1, 256)$	$(2, 128)$	$(3, 85)$	$(4, 64)$	$(5, 51)$	$(6, 42)$	$(7, 38)$	$(8, 32)$
$n_{\mathrm{p}}^{\star}$	48	36	36	40	36	40	40	48
Rate pen. [%]	0	0	0	0	0	0	0	0



(a) $\rho^{\mathrm{ul}} = 16$ dBm, $\rho^{\mathrm{dl}} = 42$ dBm, perfect CSI

(b) $\rho^{\mathrm{ul}} = 24$ dBm, $\rho^{\mathrm{dl}} = 50$ dBm, perfect CSI

(c) $\rho^{\mathrm{ul}} = 16$ dBm, $\rho^{\mathrm{dl}} = 42$ dBm , acquired CSI

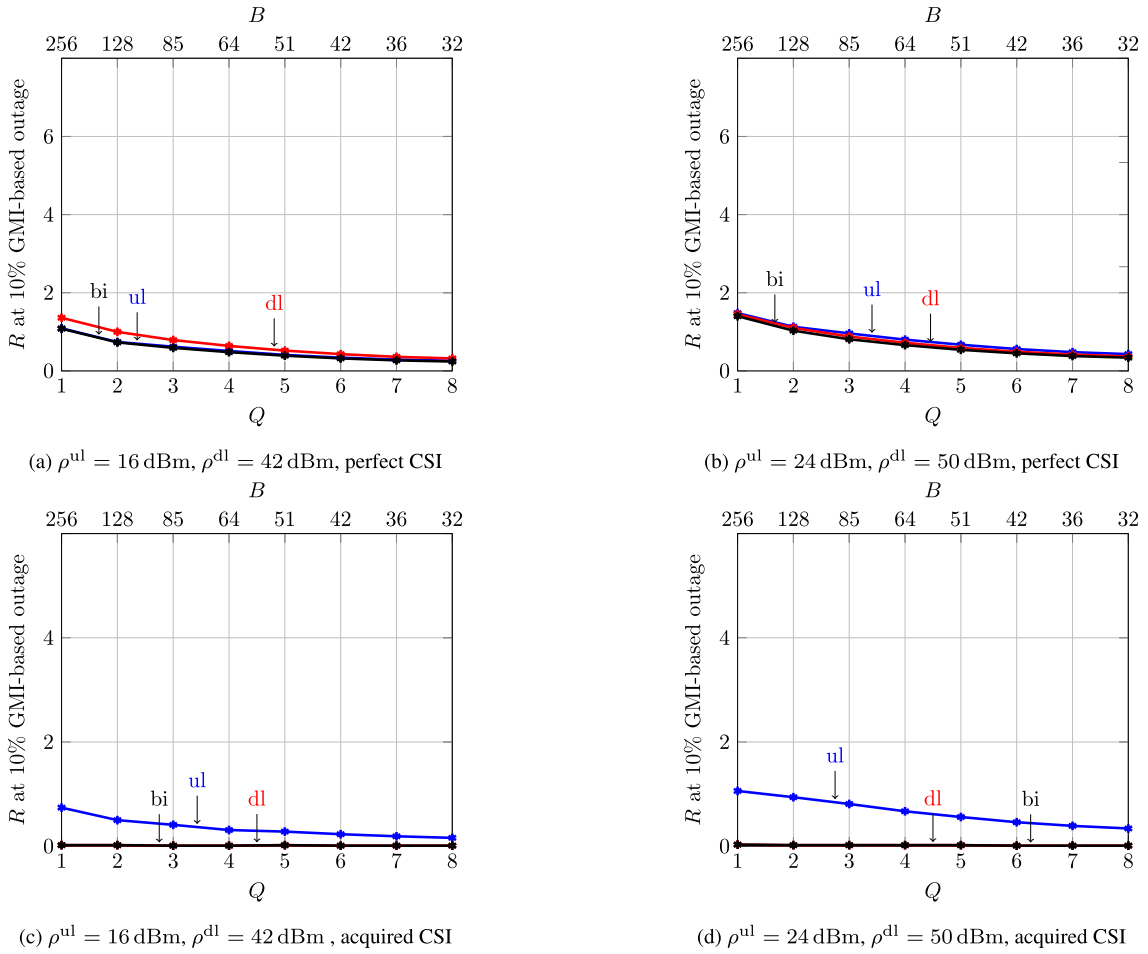
(d) $\rho^{\mathrm{ul}} = 24$ dBm, $\rho^{\mathrm{dl}} = 50$ dBm, acquired CSI

FIGURE 6. Achievable rate at 10% GMI-based outage probability as a function of the quantizer resolution Q, for a fronthaul constraint of 512 bit/s/Hz.

follows closely the uplink-rate curve. In the uplink, for both transmit power values considered in the figure, the rate is maximized when $Q = 1$, which yields a BS with $B = 256$ antennas. Indeed, since the system is power-limited in the uplink, the array gain resulting from the deployment of additional antennas, offsets the increased quantization noise resulting from the choice of 1-bit converters. On the contrary, the choice $Q = 1$ is suboptimal in the downlink, where, instead, the rate is maximized when $Q = 5$ and $Q = 6$, respectively. Here, the reduction in quantization noise and, hence, also in multiuser interference (recall that we use a quantization-unaware linear precoder) resulting from these choices of Q, which yield $B = 51$ and $B = 42$, respectively, offsets the reduction in array gain.

The picture changes when one considers the case of pilot-assisted transmission and accounts for the inaccurate channel estimate available at the BS. Indeed, as shown in Fig. 6c and 6d, imperfect CSI yields a significant reduction in the downlink rates, although the system still remains uplink limited. Similarly to the perfect-CSI case, for both values of transmitted power considered in the figures, the downlink rate is maximized when $Q = 5$. In the uplink, however, the value $Q = 1$ is optimal only for the pair $\rho^{\mathrm{ul}} = 16$ dBm and $\rho^{\mathrm{dl}} = 42$ dBm, whereas for the pair $\rho^{\mathrm{ul}} = 24$ dBm and $\rho^{\mathrm{dl}} = 50$ dBm, the uplink rate is maximizes when $Q = 3$ and the bi-directional rate when $Q = 4$.

Finally, we report in Fig. 7 the performance achievable using maximum-ratio combiner in the uplink and maximum-ratio beamformer in the downlink. As shown in the figure, the downlink performance reduces significantly so that no positive rate can be achieve in the bidirectional-transmission case, for the acquired-CSI scenario.



(a) $\rho^{\mathrm{ul}} = 16\,\mathrm{dBm}$, $\rho^{\mathrm{dl}} = 42\,\mathrm{dBm}$, perfect CSI

(b) $\rho^{\mathrm{ul}} = 24\,\mathrm{dBm}$, $\rho^{\mathrm{dl}} = 50\,\mathrm{dBm}$, perfect CSI

(c) $\rho^{\mathrm{ul}} = 16\,\mathrm{dBm}$, $\rho^{\mathrm{dl}} = 42\,\mathrm{dBm}$, acquired CSI

(d) $\rho^{\mathrm{ul}} = 24\,\mathrm{dBm}$, $\rho^{\mathrm{dl}} = 50\,\mathrm{dBm}$, acquired CSI

FIGURE 7. Achievable rate at 10% GMI-based outage probability as a function of the quantizer resolution Q, for a fronthaul constraint of 512 bit/s/Hz; Maximum ratio combiner and maximum ratio precoder.

V. CONCLUSION

We have considered the problem of designing a multiuser massive MIMO architecture where the BS is equipped with low-precision converters and a fronthaul constraint limits the amount of data that can be exchanged between the RRH and the BBU. Furthermore, we have assumed that the communication link is used to exchange short packets over a quasi-static fading channels that is not known a priori to the BS and the UEs and is estimated via uplink pilots. Our main contribution is a general framework for the characterization of the error probability in this setup, which relies on the RCUs bound from finite-blocklength information theory, a scaled nearest-neighbor decoder, and the use of Bussgang decomposition. We present both finite-blocklength bounds, and asymptotic approximations based on the GMI, which turn out to be accurate for moderate error-probability targets (see Fig. 4).

Using our bounds, we have conducted a number of experiments that shed light on the optimal design of the considered system. In particular, we have shown that when the lack of accuracy in the acquired CSI is accounted for, architectural solutions involving large antenna arrays connected to 1-bit to 4-bit converters, depending on the transmit-power values, are preferable.

Although presented for the quasi-static setup, our analysis can be extended to account for variations of the fading process within each codeword. When the fading evolves according to a block-memoryless model, such a generalization can be performed following the steps detailed for the unquantized case in [45].

ACKNOWLEDGMENT

An earlier version of this paper was presented at the Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, USA, December 2019 [DOI: 10.1109/IEEECONF44664.2019.9048859].

REFERENCES

[1] Y. Ettefagh, S. Jacobsson, A. Hu, G. Durisi, and C. Studer, "All-digital massive MIMO uplink and downlink rates under a fronthaul constraint," in *Proc. 53rd Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, USA, Nov. 2019, pp. 416–420.

[2] L. Lu, G. Ye Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive MIMO: Benefits and challenges," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 742–758, Oct. 2014.

[3] O. T. Demir, E. Björnson, and L. Sanguinetti, *Foundations of User Centric Cell-Free Massive MIMO* (Foundations and Trends in Signal Processing), vol. 14, nos. 3–4. Boston, MA, USA: Now, 2020.

[4] S. Park, O. Simeone, O. Sahin, and S. S. Shitz, "Fronthaul compression for cloud radio access networks: Signal processing advances inspired by network information theory," *IEEE Signal Process. Mag.*, vol. 31, no. 6, pp. 69–79, Nov. 2014.

[5] S. Jacobsson, G. Durisi, M. Coldrey, U. Gustavsson, and C. Studer, "Throughput analysis of massive MIMO uplink with low-resolution ADCs," *IEEE Trans. Wireless Commun.*, vol. 16, no. 6, pp. 4038–4051, Jun. 2017.

[6] C. Mollén, J. Choi, E. G. Larsson, and R. W. Heath, "Uplink performance of wideband massive MIMO with one-bit ADCs," *IEEE Trans. Wireless Commun.*, vol. 16, no. 1, pp. 87–100, Jan. 2017.

[7] E. Björnson, L. Sanguinetti, and J. Hoydis, "Hardware distortion correlation has negligible impact on UL massive MIMO spectral efficiency," *IEEE Trans. Commun.*, vol. 67, no. 2, pp. 1085–1098, Feb. 2019.

[8] S. Jacobsson, G. Durisi, M. Coldrey, T. Goldstein, and C. Studer, "Quantized precoding for massive MU-MIMO," *IEEE Trans. Commun.*, vol. 65, no. 11, pp. 4670–4684, Nov. 2017.

[9] Y. Li, C. Tao, A. Lee Swindlehurst, A. Mezghani, and L. Liu, "Downlink achievable rate analysis in massive MIMO systems with one-bit DACs," *IEEE Commun. Lett.*, vol. 21, no. 7, pp. 1669–1672, Jul. 2017.

[10] N. Liang and W. Zhang, "Mixed-ADC massive MIMO," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 983–997, Apr. 2016.

[11] A. K. Saxena, I. Fijalkow, and A. L. Swindlehurst, "Analysis of one-bit quantized precoding for the multiuser massive MIMO downlink," *IEEE Trans. Signal Process.*, vol. 65, no. 17, pp. 4624–4634, Sep. 2017.

[12] Y. Li, C. Tao, G. Seco-Granados, A. Mezghani, A. L. Swindlehurst, and L. Liu, "Channel estimation and performance analysis of one-bit massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 65, no. 15, pp. 4075–4089, Aug. 2017.

[13] J. Mo, P. Schniter, and R. W. Heath, "Channel estimation in broadband millimeter wave MIMO systems with few-bit ADCs," *IEEE Trans. Signal Process.*, vol. 66, no. 5, pp. 1141–1154, Mar. 2018.

[14] C. Studer and G. Durisi, "Quantized massive MU-MIMO-OFDM uplink," *IEEE Trans. Commun.*, vol. 64, no. 6, pp. 2387–2399, Jun. 2016.

[15] L. V. Nguyen, A. L. Swindlehurst, and D. H. N. Nguyen, "SVM-based channel estimation and data detection for one-bit massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 69, pp. 2086–2099, 2021.

[16] S. S. Thoota and C. R. Murthy, "Massive MIMO-OFDM systems with low resolution ADCs: Cramér–Rao bound, sparse channel estimation, and soft symbol decoding," *IEEE Trans. Signal Process.*, vol. 70, pp. 4835–4850, 2022.

[17] L. V. Nguyen, D. H. N. Nguyen, and A. L. Swindlehurst, "Deep learning for estimation and pilot signal design in few-bit massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 22, no. 1, pp. 379–392, Jan. 2023.

[18] S. Jacobsson, G. Durisi, M. Coldrey, and C. Studer, "Linear precoding with low-resolution DACs for massive MU-MIMO-OFDM downlink," *IEEE Trans. Wireless Commun.*, vol. 18, no. 3, pp. 1595–1609, Mar. 2019.

[19] Y. Khorsandmanesh, E. Björnson, and J. Jaldén, "Quantization-aware precoding for MU-MIMO with limited-capacity fronthaul," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Singapore, May 2022, pp. 5378–5382.

[20] A. Nedelcu, F. Steiner, and G. Kramer, "Low-resolution precoding for multi-antenna downlink channels and OFDM," *Entropy*, vol. 24, no. 4, p. 504, Apr. 2022.

[21] O. Castañeda, S. Jacobsson, G. Durisi, T. Goldstein, and C. Studer, "Finite-alphabet MMSE equalization for all-digital massive MU-MIMO mmWave communication," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 9, pp. 2128–2141, Sep. 2020.

[22] M. Bashar, K. Cumanan, A. G. Burr, H. Q. Ngo, E. G. Larsson, and P. Xiao, "Energy efficiency of the cell-free massive MIMO uplink with optimal uniform quantization," *IEEE Trans. Green Commun. Netw.*, vol. 3, no. 4, pp. 971–987, Dec. 2019.

[23] Y. Zhang, M. Zhou, X. Qiao, H. Cao, and L. Yang, "On the performance of cell-free massive MIMO with low-resolution ADCs," *IEEE Access*, vol. 7, pp. 117968–117977, 2019.

[24] X. Hu, C. Zhong, X. Chen, W. Xu, H. Lin, and Z. Zhang, "Cell-free massive MIMO systems with low resolution ADCs," *IEEE Trans. Commun.*, vol. 67, no. 10, pp. 6844–6857, Oct. 2019.

[25] G. Femenias and F. Riera-Palou, "Fronthaul-constrained cell-free massive MIMO with low resolution ADCs," *IEEE Access*, vol. 8, pp. 116195–116215, 2020.

[26] M. Bashar, P. Xiao, R. Tafazolli, K. Cumanan, A. G. Burr, and E. Björnson, "Limited-fronthaul cell-free massive MIMO with local MMSE receiver under Rician fading and phase shifts," *IEEE Wireless Commun. Lett.*, vol. 10, no. 9, pp. 1934–1938, Sep. 2021.

[27] J. Östman, A. Lancho, G. Durisi, and L. Sanguinetti, "URLLC with massive MIMO: Analysis and design at finite blocklength," *IEEE Trans. Wireless Commun.*, vol. 20, no. 10, pp. 6387–6401, Oct. 2021.

[28] E. Björnson, J. Hoydis, and L. Sanguinetti, *Massive MIMO Networks: Spectral, Energy, and Hardware Efficiency* (Foundations and Trends in Signal Processing), vol. 11, nos. 3–4. Boston, MA, USA: Now, 2017.

[29] Y. Polyanskiy, H. Vincent 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.

[30] A. Martinez and A. G. I Fàbregas, "Saddlepoint approximation of random-coding bounds," in *Proc. Inf. Theory Appl. Workshop*, San Diego, CA, USA, Feb. 2011, pp. 1–6.

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

[32] J. J. Bussgang, "Crosscorrelation functions of amplitude-distorted Gaussian signals," Res. Lab. Elec., Cambridge, MA, Tech. Rep., 216, Mar. 1952.

[33] A. Ganti, A. Lapidoth, and I. E. Telatar, "Mismatched decoding revisited: General alphabets, channels with memory, and the wideband limit," *IEEE Trans. Inf. Theory*, vol. 46, no. 7, pp. 2315–2328, Nov. 2000.

[34] S. Jacobsson, U. Gustavsson, G. Durisi, and C. Studer, "Massive MU-MIMO-OFDM uplink with hardware impairments: Modeling and analysis," in *Proc. 52nd Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove CA, USA, Oct. 2018, pp. 1829–1835.

[35] O. T. Demir and E. Bjornson, "The Bussgang decomposition of nonlinear systems: Basic theory and MIMO extensions [lecture notes]," *IEEE Signal Process. Mag.*, vol. 38, no. 1, pp. 131–136, Jan. 2021.

[36] W. Feller, *An Introduction to Probability Theory and its Applications*, 2nd ed. New York, NY, USA: Wiley, 1971.

[37] O. El Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, Mar. 2014.

[38] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, "Channel estimation and hybrid precoding for millimeter wave cellular systems," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 831–846, Oct. 2014.

[39] R. W. Heath, N. González-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 436–453, Apr. 2016.

[40] J. H. Van Vleck and D. Middleton, "The spectrum of clipped noise," *Proc. IEEE*, vol. 54, no. 1, pp. 2–19, Jan. 1966.

[41] J. Mo, P. Schniter, N. G. Prelcic, and R. W. Heath, "Channel estimation in millimeter wave MIMO systems with one-bit quantization," in *Proc. 48th Asilomar Conf. Signals, Syst. Comput.*, Pacific Grove, CA, USA, Nov. 2014, pp. 957–961.

[42] N. J. Myers, K. N. Tran, and R. W. Heath, "Low-rank MMWAVE MIMO channel estimation in one-bit receivers," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 5005–5009.

[43] R. Zhou, H. Du, and D. Zhang, "Millimeter wave MIMO channel estimation with one-bit receivers," *IEEE Commun. Lett.*, vol. 26, no. 1, pp. 158–162, Jan. 2022.

[44] O. Dabeer and A. Karnik, "Signal parameter estimation using 1-bit dithered quantization," *IEEE Trans. Inf. Theory*, vol. 52, no. 12, pp. 5389–5405, Dec. 2006.

[45] A. O. Kislal, A. Lancho, G. Durisi, and E. G. Ström, "Efficient evaluation of the error probability for pilot-assisted URLLC with massive MIMO," *IEEE J. Sel. Areas Commun.*, early access, May 29, 2023, doi: 10.1109/JSAC.2023.3280972.

YASAMAN ETTEFAGH (Member, IEEE) received the B.S. degree in electrical engineering from the Iran University of Science and Technology, Tehran, Iran, in 2012, and the M.S. degree in electrical engineering from the Amirkabir University of Technology (Tehran Polytechnic), Tehran, in 2014. She is currently pursuing the Ph.D. degree with the Electrical Engineering Department, Chalmers University of Technology, Gothenburg, Sweden. She was a RF Optimization Engineer with Ericsson, from 2015 to 2017. Her research interests include wireless communication and signal processing.

SINA REZAEI AGHDAM (Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from the Amirkabir University of Technology (Tehran Polytechnic), in 2011 and 2013, respectively, and the Ph.D. degree from Bilkent University, in 2018. His Ph.D. thesis was awarded as the best dissertation of the year by IEEE Turkey. Between 2018 and 2022, he was a Researcher with the Department of Electrical Engineering, Chalmers University of Technology, Gothenburg, Sweden. He is currently with Incluso AB, working as a Researcher and providing consultancy services in the telecommunications industry. His research interests include wireless communication and signal processing, with a special emphasis on hardware impairment modeling and compensation.

GIUSEPPE DURISI (Senior Member, IEEE) received the Laurea (summa cum laude) and Ph.D. degrees from Politecnico di Torino, Italy, in 2001 and 2006, respectively. From 2002 to 2006, he was with Istituto Superiore Mario Boella, Turin, Italy. From 2006 to 2010, he was a Postdoctoral Researcher with ETH Zurich, Zürich, Switzerland. In 2010, he joined the Chalmers University of Technology, Gothenburg, Sweden, where he is currently a Professor with the Communication Systems Group. His research interests include communication and information theory and machine learning. He was a recipient of the 2013 IEEE ComSoc Best Young Researcher Award for the Europe, Middle East, and Africa region. He is the coauthor of a paper that won the Student Paper Award at the 2012 International Symposium on Information Theory and a paper that won the 2013 IEEE Sweden VT-COM-IT Joint Chapter Best Student Conference Paper Award. From 2011 to 2014, he served as a Publications Editor for the IEEE TRANSACTIONS ON INFORMATION THEORY. From 2015 to 2021, he served as an Associate Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS.

SVEN JACOBSSON received the Ph.D. degree in electrical engineering from the Chalmers University of Technology, Gothenburg, Sweden, in 2019. In 2015, he joined Ericsson Research, where he is currently an experienced Researcher. His research interests include advanced antenna systems and hardware-constrained communications.

MIKAEL COLDREY received the M.Sc. degree in applied physics and electrical engineering from Linköping University, Linköping, Sweden, in 2000, and the Ph.D. degree in electrical engineering from the Chalmers University of Technology, Gothenburg, Sweden, in 2006. He joined Ericsson Research, in 2006, where he is currently a Principal Researcher. He has been working with 4G research and for several years with 5G research. Since 2012, he has been an Adjunct Associate Professor with the Chalmers University of Technology. His research interests include advanced antenna systems, channels, models, algorithms, and millimeter-wave communications for both radio access and wireless backhaul systems.

CHRISTOPH STUDER (Senior Member, IEEE) received the M.S. and Ph.D. degrees in electrical engineering from ETH Zürich, Switzerland, in 2005 and 2009, respectively. Between 2009 and 2013, he held a postdoctoral researcher positions with ETH Zürich and Rice University, Houston, TX, USA. In 2014, he joined Cornell University, Ithaca, NY, USA, as an Assistant Professor. From 2019 to 2020, he was an Associate Professor with Cornell University and Cornell Tech, New York City. In 2020, he joined ETH Zürich, where he is currently an Associate Professor with the Department of Information Technology and Electrical Engineering. His research interests include the design of digital integrated circuits, wireless communications, digital signal processing, numerical optimization, and machine learning.