Asymptotic Analysis of SU-MIMO Channels With Transmitter Noise and Mismatched Joint Decoding

Mikko Vehkaperä, Taneli Riihonen, Maksym Girnyk, Emil Björnson, Mérouane Debbah, Lars Kildehøj Rasmussen, Risto Wichman

Asymptotic Analysis of SU-MIMO Channels With Transmitter Noise and Mismatched Joint Decoding

Mikko Vehkaperä, *Member, IEEE*, Taneli Riihonen, *Member, IEEE*, Maksym A. Girnyk, *Student Member, IEEE*, Emil Björnson, *Member, IEEE*, Mérouane Debbah, *Fellow, IEEE*, Lars Kildehøj Rasmussen, *Senior Member, IEEE*, and Risto Wichman

***Abstract*—Hardware impairments in radio-frequency components of a wireless system cause unavoidable distortions to transmission that are not captured by the conventional linear channel model. In this paper, a "binoisy" single-user multiple-input multiple-output (SU-MIMO) relation is considered where the additional distortions are modeled via an additive noise term at the transmit side. Through this extended SU-MIMO channel model, the effects of transceiver hardware impairments on the achievable rate of multi-antenna point-to-point systems are studied. Channel input distributions encompassing practical discrete modulation schemes, such as, QAM and PSK, as well as Gaussian signaling are covered. In addition, the impact of mismatched detection and decoding when the receiver has insufficient information about the non-idealities is investigated. The numerical results show that for realistic system parameters, the effects of transmit-side noise and mismatched decoding become significant only at high modulation orders.**

***Index Terms*—Multiple antennas, mismatched decoding, ergodic capacity, fading channels, generalized mutual information (GMI), transceiver hardware impairments.**

I. Introduction

MIMO, i.e., multiple-input multiple-output, wireless links are a mature research subject and their theory is already well understood [1]. However, the extensive body of literature on link-level analysis conventionally concerns signal models of the form $\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{n}$ reckoning with an additive thermal-noise term, namely $\boldsymbol{n}$, only at the receiver after the fading channel $\boldsymbol{H}$. In this paper, we investigate single-user MIMO channels and adopt a generalized ("binoisy") input–output relation from [2]–[11]:

$$\boldsymbol{y} = \boldsymbol{H}(\boldsymbol{x} + \boldsymbol{v}) + \boldsymbol{w}, \tag{1}$$

where $\boldsymbol{w}$ is an additive receive-side distortion-plus-noise component. The system model (1) allows including an additive noise term, namely $\boldsymbol{v}$, also at the transmitter, thus making the total effective noise term $\boldsymbol{H}\boldsymbol{v} + \boldsymbol{w}$ colored and correlated with the fading channel. This small but significant complement yields a MIMO link model whose performance analysis is still an open research niche in many respects.

Although we primarily aim at extending the capacity theory of binoisy SU-MIMO channels under fading without committing to any particular application, the signal model (1) originally stems from the practical need for modeling the combined effect of various transceiver hardware impairments which are detailed in [12], [13], and the references therein. However, it is worth acknowledging that the additive noise assumed herein is only a simplified representation of complex nonlinear phenomena occurring due to hardware impairments, especially when considering their joint coupled effects or trying to model residual distortion after compensation. Thus, the binoisy signal model should be regarded as a compromise between facilitating theoretical analysis and resorting to measurements or simulations under more accurate modeling. Yet the central limit theorem further justifies the model by averaging the combined effects of different impairments to additive Gaussian noise when the signal model (1) is understood to represent a single narrowband subcarrier within a wideband system.

Additive receiver hardware impairments can be incorporated into the conventional signal model by increasing the level of the thermal-noise term $\boldsymbol{n}$ by a constant noise figure, e.g., about 3–5 dB, or by scaling it in proportion to the input signal level such that it matches with $\boldsymbol{w}$. On the other hand, regarding the joint effect of transmitter hardware impairments as an additive transmit-side noise term $\boldsymbol{v}$ is analogous to the principles of practical radio conformance testing. In particular, the common transmitter quality indicator is error-vector magnitude (EVM) which reduces the distortion effects to an additive component and measures its level relatively to signal amplitude [14].

Typical target EVM values guarantee that the signal $\boldsymbol{x}$ is at least 20–30 dB above the transmit-side noise $\boldsymbol{v}$. On the other hand, for basic discrete channel inputs such as quadrature phase-shift keying (QPSK), $\boldsymbol{H}\boldsymbol{x}$ is usually at most 10–15 dB above the receive-side noise $\boldsymbol{w}$, after which the communication is not anymore limited by noise but the lack of entropy in the

Manuscript received June 9, 2014; revised October 22, 2014; accepted December 10, 2014. This research was supported in part by the Academy of Finland, the Swedish Research Council and the ERC Starting Grant 305123 MORE. The associate editor coordinating the review of this paper and approving it for publication was V. Raghavan.

M. Vehkaperä, T. Riihonen, and R. Wichman are with the Department of Signal Processing and Acoustics, Aalto University School of Electrical Engineering, 00076 Aalto, Finland (e-mail: mikko.vehkapera@aalto.fi; taneli.riihonen@aalto.fi; risto.wichman@aalto.fi).

M. A. Girnyk and L. K. Rasmussen are with the Department of Communication Theory, School of Electrical Engineering, and the ACCESS Linnaeus Center, KTH Royal Institute of Technology, 100 44 Stockholm, Sweden (e-mail: mgyr@kth.se; lkra@kth.se).

E. Björnson is with Department of Electrical Engineering (ISY), Linköping University, 581 83 Linköping, Sweden (e-mail: emil.bjornson@liu.se).

M. Debbah is with Huawei France R&D Center, 92659 Paris, France, and also with Supelec, 91192 Gif-sur-Yvette, France (e-mail: merouane.debbah@huawei.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCOMM.2014.2385051

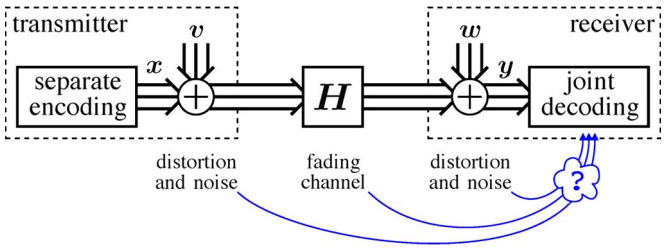

Fig. 1. System model for non-ideal MIMO communications with transmit and receive distortion. The receiver might be misinformed or ignorant of some of the variables in the transmission chain leading to mismatched decoding.

modulation alphabet. This implies that transmitter hardware impairments can be justifiably omitted in the analysis of simple low-rate wireless systems: Either $\boldsymbol{H}\boldsymbol{v}$ is well below the receive-side noise $\boldsymbol{w}$ (say 5–20 dB) or the signal-to-noise ratio (SNR) is set to an uninterestingly high level. However, there has been a trend to improve data rates by using, e.g., quadrature amplitude modulation (QAM) up to 64-QAM at relatively high SNR, in which case the transmit-side noise begins to play a notable role in the link-level performance.

The considered system setup corresponding to (1) is shown in Fig. 1. As for MIMO processing, we focus on regular spatial multiplexing where a conventional transmitter separately encodes and sends an independent stream at each of its antennas without having channel state information or being aware of the transmit-side noise it produces; the receiver jointly decodes the output signals of the MIMO channel knowing its instantaneous realization $\boldsymbol{H}$ and some noise statistics. However, conventional receivers are designed and implemented based on the conventional signal model (where $\boldsymbol{v} = \boldsymbol{0}$) due to which they are prone to lapse into suboptimal *mismatched decoding* by inaccurately postulating the statistics of the actual noise term $\boldsymbol{H}\boldsymbol{v} + \boldsymbol{w}$. Even if off-the-shelf receivers can adapt to colored receiver noise, they may not be able to track the variable statistics of the component $\boldsymbol{H}\boldsymbol{v}$ propagated from the transmitter since it is correlated with the fading channel. Only an advanced receiver would be able to perform *matched decoding* knowing perfectly the noise statistics as if it was designed and implemented explicitly based on the generalized binoisy signal model (1).

A. Related Works

The key reference results for the present study are reported in [2]–[11]. These seminal works originally formulated the research niche around (1) and established the baseline understanding of MIMO communication in the presence of transmit-side noise with numerical simulations and theoretical analysis. The majority of the related works, e.g., [2], [3], [6], [8], concern regular spatial multiplexing using separate encoding like the present paper but also different variations of joint encoding have been creditably investigated, e.g., in [4], [7]. On the other hand, all the studies that we are aware of assume (implicitly) advanced receivers that know the presence of transmit noise, no matter what form of decoding is used.

Especially, the reference results are polarized such that the scope of analytical studies [6], [8] typically differs from that of studies reporting simulations [6], [7], [9] or measurements [4]–[6]. Except for [2], practical discrete modulation schemes, e.g., QAM, have not been previously analytically evaluated in the presence of transmit noise, and simulation-based studies usually concern bit/symbol/packet error rates, not transmission rates which could be more interesting when studying modern adaptive encoding. In contrast, all the analytical capacity studies assume Gaussian signaling and the throughput simulations of [3] with adaptive modulation and coding are their closest counterpart when it comes to experimental work.

If the receiver does not properly account for the additional transmit-side noise in the received signal, conventional mutual information (MI) is not anymore the correct upper bound for coded transmissions. Rather, due to mismatched decoding, one has to employ other metrics, such as *generalized mutual information* (GMI) [15], [16] adopted herein. Another common use for GMI is the analysis of bit-interleaved coded modulation [17], while also transceiver hardware impairments [18] and effects of imperfect channel state information at receiver [19]–[21] are analyzed in terms of GMI. It is important to realize that the GMI framework differs, both conceptually and in technical details, from the approach in [22], [23], where conventional MI of a modified channel model is computed for a decoder that has certain side-information about the variables in the modified system.

In the present paper, MI and GMI are evaluated using the *replica method* [24], [25], originating from the field of statistical physics and introduced to the analysis of wireless systems by [26], [27]. Since then, the replica method has been applied to various problems in communication theory, e.g., MIMO systems [28]–[31]. For some special cases like Gaussian signaling, the replica trick renders exact asymptotic results when the number of antennas grows without bound, while they can be otherwise considered accurate approximations as shown by comparisons to Monte Carlo simulations.

B. Summary of Contributions

In this paper, we investigate two aspects of binoisy MIMO channels that are unexplored in related works despite their fundamental role in understanding the effects of hardware impairments in wireless systems. Firstly, analytical capacity results are limited to Gaussian signaling while practical digital modulation is evaluated only based on simple simulations or measurements. Secondly, the earlier literature focuses on the optimistic case of matched decoding by employing receivers that are actually not available off the shelf but implicitly updated to take account of transmit-side noise.

In particular, this paper contributes to the capacity theory of MIMO communication links by examining the effects of transmit-side noise as follows.

- Analytical GMI expressions are calculated for studying the rate loss of mismatched decoding when using a conventional receiver which is unaware of the transmit-side noise. Especially, it is shown that the performance remains the same irrespective of how well the noise covariance matrix is known if it is a constant.
- The above analysis is further translated into corresponding asymptotic high-SNR limits for Gaussian signaling as a complement for the results of [10], which covers matched decoding and conventional MI.

- The analytical expressions provided for both conventional MI and GMI cover many practical discrete modulation schemes such as variations of PSK and QAM. This resolves the serious problem that evaluating (G)MI with direct Monte Carlo simulations for the present system is computationally infeasible except for cases with small number of antennas and low order modulation sets.

Extending beyond the scope of the paper, the replica analysis of GMI is also a new aspect at large.

C. Outline of the Paper and Its Nomenclature

After the considered system model is specified in the following section, the main analytical content of this paper is divided into two parts: Section III concerns the performance of conventional suboptimal receivers under mismatched decoding, which is analyzed based on GMI; and Section IV studies conventional MI with advanced receivers, which are aware of transmitter noise and, thus, capable of optimal matched decoding. In Section V, the presented theory is illustrated with numerical results, including simulations for double-checking its accuracy, which is finally followed by concluding remarks in the last section. Some general results from literature that are used throughout the paper for derivations are collected in Appendix A for the convenience of the reader. Appendices B contains general description of the replica method and Appendix C sketches the derivation of the main results in Section III.

Notation: Complex Gaussian random variables (RVs) are always assumed to be proper and the density of such $\boldsymbol{x} \in \mathbb{C}^N$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{R}$ is denoted $g(\boldsymbol{x}|\boldsymbol{\mu};\boldsymbol{R})$. For the zero-mean proper Gaussians, we say they are circularly symmetric complex Gaussian (CSCG). For convenience, both discrete and continuous RVs are said to have a probability density function (PDF) that is denoted by p, and we do not separate RVs and their realizations. For postulated PDFs we write q and add tilde on top of the related RVs (most of the time). Given a RV x that has a PDF $p(x)$, we write $x \sim p(x)$ (and $\tilde{x} \sim q(\tilde{x})$ for the postulated case). Statistical expectation is denoted $\mathsf{E}\{\cdot\}$ and, unless stated otherwise, calculated over all randomness in the argument using true or postulated PDFs, depending on which type of RVs are present. Integrals w.r.t. real-valued variables are always over $\mathbb{R}$ (for vectors over the appropriate product space) and we tend to omit the integration limits for notational simplicity. For a complex variable $z = x + \mathrm{j}y$, we denote $\int(\,)\mathrm{d}z = \int(\,)\mathrm{d}x\,\mathrm{d}y$, and similarly for complex vectors. Logarithms are natural logs and denoted ln unless stated otherwise.

II. System Model

Consider the system model depicted in Fig. 1 and the signal model of $\boldsymbol{y} \in \mathbb{C}^N$ written in (1) where $\boldsymbol{H} \in \mathbb{C}^{N\times M}$ is the channel matrix and $\boldsymbol{x} \in \mathbb{C}^M$ the signal of interest. The receive-side distortion plus noise component is divided into two parts, namely $\boldsymbol{w} = \boldsymbol{n} + \boldsymbol{\omega} \in \mathbb{C}^N$ where $\boldsymbol{n}$ is caused by thermal noise and $\boldsymbol{\omega}$ represents hardware impairments arising from the non-ideal behavior of the radio-frequency (RF) transceivers. Similarly, $\boldsymbol{v} = \boldsymbol{m} + \boldsymbol{\nu} \in \mathbb{C}^M$ where $\boldsymbol{m}$ and $\boldsymbol{\nu}$ are related to thermal noise and hardware impairments or distortions, respectively, at the transmit-side. In practice, the effect of $\boldsymbol{m}$ is often negligible compared to $\boldsymbol{\nu}$. In conventional MIMO literature it is common to consider only the thermal noise at the receiver, which translates to assuming $\boldsymbol{\omega} = \boldsymbol{\nu} = \boldsymbol{m} = \boldsymbol{0}$ in our more generic system model.

Let us denote the PDF of the transmit vector $\boldsymbol{x}$ by $p(\boldsymbol{x})$ and assume it factorizes as

$$p(\boldsymbol{x}) = \prod_{m=1}^{M} p(x_m), \tag{2}$$

so that independent streams are transmitted at each transmit antenna. Furthermore, let $p(x_m)$ be a zero-mean distribution with variance $\bar{\gamma}_m$. For later convenience, we let $\boldsymbol{\Gamma}$ be a diagonal matrix whose non-zero elements are given by $\bar{\gamma}_1, \ldots, \bar{\gamma}_M$, that is, $\boldsymbol{\Gamma} = \mathsf{E}\{\boldsymbol{x}\boldsymbol{x}^{\mathsf{H}}\}$. The channel $\boldsymbol{H}$ is assumed to have independent identically distributed (IID) CSCG elements with variance[1] $1/M$. The thermal noise samples at the transceivers are modeled as CSCG random vectors $\boldsymbol{m}$ and $\boldsymbol{n}$ that have independent elements. For simplicity, we assume that any given noise or hardware impairment component is independent of any other RVs in the system. The transmit-and receive-side impairments $\boldsymbol{\nu}$ and $\boldsymbol{\omega}$ are taken to be CSCG random vectors with covariance matrices $\boldsymbol{R}_{\boldsymbol{\nu}}$ and $\boldsymbol{R}_{\boldsymbol{\omega}}$, respectively. The distortion plus noise vectors $\boldsymbol{v}$ and $\boldsymbol{w}$ are thus CSCG random vectors whose covariance matrices we denote $\boldsymbol{R}_{\boldsymbol{v}}$ and $\boldsymbol{R}_{\boldsymbol{w}}$, respectively. Notice that these matrices can be functions of the statistics of some other RVs albeit we suppress the explicit statement of such dependence at this point for notational convenience. The SNR without transmit-side noise is defined as $\mathrm{tr}(\boldsymbol{\Gamma})/\mathrm{tr}(\boldsymbol{R}_{\boldsymbol{w}})$.

The PDF of the received signal, conditioned on $\boldsymbol{x}$, $\boldsymbol{v}$ and $\boldsymbol{H}$, is given by

$$p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{v},\boldsymbol{H}) = g\left(\boldsymbol{y}|\boldsymbol{H}(\boldsymbol{x}+\boldsymbol{v});\boldsymbol{R}_{\boldsymbol{w}}\right), \tag{3}$$

and the receiver is assumed to know $\boldsymbol{H}$ and the true distribution $p(\boldsymbol{x})$ of the channel input. However, the additional transmit-side term $\boldsymbol{v}$ is in general unknown at the receive-side and, thus, the PDF (3) cannot be directly used for detection and decoding. Herein, we consider two different scenarios for the joint decoding operation at the receiver:

1) The receiver knows $\boldsymbol{H}$, the PDFs of the noise plus distortion terms $\boldsymbol{v}$ and $\boldsymbol{w}$ as well as the distribution of the data vector $\boldsymbol{x}$. *Matched joint decoding* is then based on the conditional PDF

$$p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{H}) = \mathsf{E}_{\boldsymbol{v}}\left\{g\left(\boldsymbol{y}|\boldsymbol{H}(\boldsymbol{x}+\boldsymbol{v});\boldsymbol{R}_{\boldsymbol{w}}\right)\right\} \tag{4}$$

$$= g(\boldsymbol{y}|\boldsymbol{H}\boldsymbol{x};\boldsymbol{R}_{\boldsymbol{w}} + \boldsymbol{H}\boldsymbol{R}_{\boldsymbol{v}}\boldsymbol{H}^{\mathsf{H}}), \tag{5}$$

where the second equality follows by first using (62) to calculate the expectation w.r.t. $\boldsymbol{v}$ and simplifying the end result using (63) and (64). Note that the effective noise

[1]Typically the total power emitted from the transmit antennas in MIMO systems is constant; that is, $\mathrm{tr}(\boldsymbol{\Gamma}) = \bar{\gamma}$, where $\bar{\gamma}$ is some fixed power budget that does not depend on M. Hence the elements of $\boldsymbol{\Gamma}$ need to be functions of M to satisfy the transmit power normalization. For the following analysis, however, it is more convenient to treat the elements of $\boldsymbol{\Gamma}$ to be independent of M and let the transmit power normalization be a part of the channel. Clearly, both approaches are mathematically fully equivalent.

covariance matrix in (5) depends now on the instantaneous channel realization $\boldsymbol{H}$.

2) The receiver has perfect knowledge of $\boldsymbol{H}$ and the PDF of the data vector $\boldsymbol{x}$. Instead of (4), however, the device uses a postulated channel law

$$q(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{H}) = g(\boldsymbol{y}|\boldsymbol{H}\boldsymbol{x};\tilde{\boldsymbol{R}}), \tag{6}$$

for *mismatched joint decoding* [15], [16]. In contrast to $\boldsymbol{R}_{\boldsymbol{w}}$ in (5), that is a random matrix, the postulated covariance matrix $\tilde{\boldsymbol{R}}$ in (6) is arbitrary but fixed.

If matched joint decoding is employed, the conventional metric for evaluating the (ergodic) achievable rate of the system for given input distribution $p(\boldsymbol{x})$ is the MI between the channel inputs and outputs, namely,

$$I(\boldsymbol{y};\boldsymbol{x}) = \mathsf{E}\left\{\ln p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{H})\right\} - \mathsf{E}\left\{\ln p(\boldsymbol{y}|\boldsymbol{H})\right\}, \tag{7}$$

where $p(\boldsymbol{y}|\boldsymbol{H}) = \mathsf{E}_{\boldsymbol{x}}\{p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{H})\}$ and the expectation is w.r.t. all RVs in the system model, including the channel matrix $\boldsymbol{H}$. From the system design perspective, however, it might be impractical to use (5) due to complexity of implementation, resulting in mismatched decoding. To lower bound the true maximum rate that can be achieved reliably over channel (1) when decoding rule (6) is used at the receiver, we use GMI that is discussed in the next section.

III. Mismatched Joint Decoding: Generalized Mutual Information

A. Definition and Key Properties

Let us assume that the received signal is given by (1) but the receiver uses (6) for decoding. Given $p(\boldsymbol{x})$, the (ergodic) GMI between the channel inputs and outputs is defined as [15], [16]

$$I_{\mathsf{GMI}}(\boldsymbol{y};\boldsymbol{x}) = \sup_{s>0} I^{(s)}_{\mathsf{GMI}}(\boldsymbol{y};\boldsymbol{x}), \tag{8}$$

where, denoting $q^{(s)}(\boldsymbol{y}|\boldsymbol{H}) = \mathsf{E}_{\boldsymbol{x}}\{q(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{H})^s\}$, the s-dependent part reads

$$I^{(s)}_{\mathsf{GMI}}(\boldsymbol{y};\boldsymbol{x}) = \mathsf{E}\left\{\ln q(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{H})^s\right\} - \mathsf{E}\left\{\ln q^{(s)}(\boldsymbol{y}|\boldsymbol{H})\right\}. \tag{9}$$

Notice that since we consider ergodic rates, the expectations in (9) are w.r.t. all RVs in the system model, including the channel matrix $\boldsymbol{H}$.

The GMI in (8) gives the achievable rate for which reliable transmission over the channel (1) is possible given decoding metric (6) and ensemble of codebooks where the code words are independent with IID elements. An important property of GMI is that it always provides a valid lower bound for the maximum rate of the channel, namely, if I is the maximum ergodic rate that can be reliably transmitted over the channel (1) using input distribution $p(\boldsymbol{x})$ and decoding rule (6), then[2] $I \geq I_{\mathsf{GMI}}$. It is important to notice that the task of computing I for an arbitrary channel with an arbitrary decoding rule is in general an open research problem and the very reason why we have to resort to alternative approaches such as GMI.

[2]For the special case of IID codebooks and discrete memoryless channels with mismatched decoding, the lower bound provided by GMI is indeed tight [16]. However, there are examples (see e.g., [15], [16] and references therein) where rates higher than GMI can be obtained through other choice of code distribution. The downside of the techniques employed in the latter case is that they are often limited to finite alphabet channels and are much more cumbersome to use than the GMI, which can be easily applied to very general channel models. For more discussion on GMI, see for example, [15]–[21] and the references therein.

B. Special Case of Gaussian Signaling

We are first interested in evaluating the s-dependent part of the normalized GMI per transmit stream $M^{-1}I^{(s)}_{\mathsf{GMI}}(\boldsymbol{y};\boldsymbol{x})$ for given $s>0$. The optimization over the free parameter s is carried out after the suitable expressions are found. Note that (9) is a valid lower bound on the achievable rate for any $s>0$ and the optimization is carried out to obtain the tightest bound possible. The first term in (9) can be written as

$$\begin{aligned}
&\frac{1}{M}\mathsf{E}\left\{\ln q(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{H})^s\right\} \\
&\quad= -\overbrace{\frac{s}{M}[N\ln\pi + \ln\det\tilde{\boldsymbol{R}}]}^{=c^{(s)}} \\
&\qquad - \frac{s}{M}\mathsf{E}\left\{(\boldsymbol{H}\boldsymbol{v}+\boldsymbol{w})^{\mathsf{H}}\tilde{\boldsymbol{R}}^{-1}(\boldsymbol{H}\boldsymbol{v}+\boldsymbol{w})\right\} \\
&\quad= -c^{(s)} - \frac{s}{M}\left[\mathrm{tr}(\tilde{\boldsymbol{R}}^{-1}\boldsymbol{R}_{\boldsymbol{w}}) + \frac{1}{M}\mathrm{tr}(\tilde{\boldsymbol{R}}^{-1})\mathrm{tr}(\boldsymbol{R}_{\boldsymbol{v}})\right].
\end{aligned} \tag{10}$$

The first equality follows from (6) by the fact that $\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x} = \boldsymbol{H}\boldsymbol{v}+\boldsymbol{w}$ when $\boldsymbol{x}$ is given. The second equality is a consequence of the assumption that the channels and noise vectors are all mutually independent and $\boldsymbol{H}$ has zero-mean IID entries with variance $1/M$. Notice that (10) is independent of $p(\boldsymbol{x})$ and hence valid for all channel inputs. Evaluating the second term in (9) is more complicated but *for the special case of Gaussian inputs* we have the result shown below.

Example 1: For the special case of Gaussian inputs; that is, $p(\boldsymbol{x}) = g(\boldsymbol{x}|\boldsymbol{0};\boldsymbol{\Gamma})$,

$$\begin{aligned}
\frac{1}{M}I^{(s)}_{\mathsf{GMI}}(\boldsymbol{y};\boldsymbol{x}) = \frac{1}{M}\mathsf{E}_{\boldsymbol{H}}\Big\{&\ln\det(\tilde{\boldsymbol{R}} + s\boldsymbol{H}\boldsymbol{\Gamma}\boldsymbol{H}^{\mathsf{H}}) \\
&+ s\,\mathrm{tr}\left[\left(\boldsymbol{R}_{\boldsymbol{w}} + \boldsymbol{H}(\boldsymbol{R}_v+\boldsymbol{\Gamma})\boldsymbol{H}^{\mathsf{H}}\right)\left(\tilde{\boldsymbol{R}} + s\boldsymbol{H}\boldsymbol{\Gamma}\boldsymbol{H}^{\mathsf{H}}\right)^{-1}\right] \\
&- s\,\mathrm{tr}(\tilde{\boldsymbol{R}}^{-1}\boldsymbol{R}_{\boldsymbol{w}}) - \frac{s}{M}\mathrm{tr}(\tilde{\boldsymbol{R}}^{-1})\mathrm{tr}(\boldsymbol{R}_{\boldsymbol{v}}) - \ln\det\tilde{\boldsymbol{R}}\Big\}.
\end{aligned} \tag{11}$$

The result is obtained by first using (62) and then simplifying with (63) and (64). Inserting the RHS of (1) into the obtained expression and taking the expectations w.r.t. the noise terms $\boldsymbol{v}$ and $\boldsymbol{w}$ completes the derivation.

Example 1 shows that for Gaussian signals we only need to average over the channel matrix $\boldsymbol{H}$ to obtain the s-dependent part of GMI. This is doable with Monte Carlo simulation. However, finding the optimal s is time consuming even in this case and a simple analytical expression that does not explicitly depend on the form of the marginals in (2) would be highly desirable. With this in mind, we adopt the following restriction to our system model from the physical characteristics of typical real transmitters for simplifying the analysis.

Assumption 1: The covariance matrix for the transmit-side distortion plus noise term $\boldsymbol{v}$ is diagonal so that we may write $\boldsymbol{R}_{\boldsymbol{v}} = \boldsymbol{R}_{\boldsymbol{m}} + \boldsymbol{R}_{\boldsymbol{\nu}} = \mathrm{diag}(r^{(1)}_{\boldsymbol{v}},\ldots,r^{(M)}_{\boldsymbol{v}})$. Hence, $\boldsymbol{v}$ has

independent (but not necessarily identically distributed) entries drawn according to $p(v_m) = g(v_m|0; r_{\boldsymbol{v}}^{(m)})$.

The physical meaning of this assumption is that hardware impairments at different transmitter branches arise in separate electrical components and there are no mechanisms which generate significant correlation between the elements of the distortion noise vector. Furthermore, it is actually not necessary for the replica analysis but it helps simplify the end result to a form whose numerical evaluation is computationally easy.

C. Analytical Results via the Replica Method

If the goal is to calculate the expectations related to the latter term in (9) analytically and for general input distributions, we need to employ somewhat more advanced analytical tools than the basic probability calculus used in Example 1. As we shall see shortly, employing the replica method provides a formula that is applicable to a variety of input constellations, such as Gaussian or QAM. To begin, let us first denote

$$-\frac{1}{M}\mathsf{E}\ln q^{(s)}(\boldsymbol{y}|\boldsymbol{H}) = c^{(s)} + f(s), \tag{12}$$

where $c^{(s)}$ is defined in (10) and the latter term, equivalent of the so-called *free energy* in statistical mechanics, reads

$$f(s) = -\frac{1}{M}\mathsf{E}\left\{\ln \mathsf{E}_{\tilde{\boldsymbol{x}}}\left\{\mathrm{e}^{-[\boldsymbol{H}(\boldsymbol{x}+\boldsymbol{v}-\tilde{\boldsymbol{x}})+\boldsymbol{w}]^{\mathsf{H}} s\tilde{\boldsymbol{R}}^{-1}[\boldsymbol{H}(\boldsymbol{x}+\boldsymbol{v}-\tilde{\boldsymbol{x}})+\boldsymbol{w}]}\right\}\right\}. \tag{13}$$

Now the inner expectation over the postulated channel input $\tilde{\boldsymbol{x}}$ is w.r.t. a generic PDF (2) and cannot be solved using (62) as before. The outer expectation is w.r.t. the rest of the RVs in the system, namely $\{\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{w}, \boldsymbol{H}\}$. Due to (9) and (10) the expression to be optimized in the GMI formula thus becomes

$$\frac{1}{M} I_{\mathsf{GMI}}^{(s)}(\boldsymbol{y}; \boldsymbol{x}) = f(s) - \frac{s}{M}\left[\mathrm{tr}(\tilde{\boldsymbol{R}}^{-1}\boldsymbol{R}_{\boldsymbol{w}}) + \frac{1}{M}\mathrm{tr}(\tilde{\boldsymbol{R}}^{-1})\mathrm{tr}(\boldsymbol{R}_{\boldsymbol{v}})\right]. \tag{14}$$

Remark 1: By (13) and (14), it is clear that if the receiver assumes that the additive noise in the system is spatially white $\tilde{\boldsymbol{R}} = \tilde{r}\boldsymbol{I}_N$ with some finite sample variance $\tilde{r}$, the GMI remains the same for all $\tilde{r} > 0$ since the optimization over $s > 0$ in (8) can be replaced by an optimization over a new variable $\tilde{s} = s/\tilde{r} > 0$. Thus, if the receiver uses $\tilde{\boldsymbol{R}} = \tilde{r}\boldsymbol{I}_N$ for decoding, the GMI is the same for all $\tilde{r} > 0$ when the transmit and receive covariance matrices $\boldsymbol{R}_{\boldsymbol{v}}$ and $\boldsymbol{R}_{\boldsymbol{w}}$ are fixed.

The main obstacle in evaluating (14) is clearly $f(s)$. This term happens to be, however, of a form that can be tackled by the replica method, as outlined in Appendix B. The following result is derived in Appendix C under the assumption of the so-called *replica symmetric* (RS) ansatz (see Step 3 in Appendix B) when the system approaches the *large system limit* (LSL), that is, $M, N \to \infty$ with finite and fixed ratio $\alpha = M/N > 0$. The limit notation is omitted below and the results should therefore be interpreted as approximations for systems that have finite dimensions.

Proposition 1: Let $m = 1, \ldots, M$ and denote

$$\chi_m = x_m + v_m, \tag{15}$$
$$\tilde{\chi}_m = \tilde{x}_m, \tag{16}$$

where $x_m, \tilde{x}_m \sim p(x_m)$ and $v_m \sim g(v_m|0; r_{\boldsymbol{v}}^{(m)})$ are independent for all m by assumption. Let

$$p(z_m|\chi_m) = g(z_m|\chi_m; \eta^{-1}), \tag{17}$$
$$q(z_m|\tilde{\chi}_m) = g(z_m|\tilde{\chi}_m; \xi^{-1}), \tag{18}$$

be the PDF of an output z_m of an additive white Gaussian noise (AWGN) channel whose input is either (15) or (16), respectively, and corrupted by additive noise with variance η^{-1} or ξ^{-1}, respectively. The parameters η, ξ satisfy

$$\eta = \frac{1}{\alpha}\frac{\left[\frac{1}{N}\mathrm{tr}(\tilde{\boldsymbol{\Omega}}^{-1})\right]^2}{\frac{1}{N}\mathrm{tr}(\tilde{\boldsymbol{\Omega}}^{-1}\boldsymbol{\Omega}\tilde{\boldsymbol{\Omega}}^{-1})}, \tag{19}$$

$$\xi = \frac{1}{\alpha N}\mathrm{tr}(\tilde{\boldsymbol{\Omega}}^{-1}), \tag{20}$$

for the given matrices

$$\boldsymbol{\Omega} = \boldsymbol{R}_{\boldsymbol{w}} + \varepsilon\boldsymbol{I}_N, \tag{21}$$
$$\tilde{\boldsymbol{\Omega}} = s^{-1}\tilde{\boldsymbol{R}} + \tilde{\varepsilon}\boldsymbol{I}_N, \tag{22}$$

and variables

$$\varepsilon = \frac{1}{M}\sum_{m=1}^{M}\mathsf{E}\left\{|v_m + x_m - \langle\tilde{x}_m\rangle_q|^2\right\}, \tag{23}$$

$$\tilde{\varepsilon} = \frac{1}{M}\sum_{m=1}^{M}\mathsf{E}\left\{|\tilde{x}_m - \langle\tilde{x}_m\rangle_q|^2\right\}. \tag{24}$$

The notation $\langle\tilde{x}_m\rangle_q$ above refers to a decoupled posterior mean estimator

$$\langle\tilde{x}_m\rangle_q = \frac{\mathsf{E}_{\tilde{x}_m}\{\tilde{x}_m q(z_m|\tilde{x}_m)\}}{q(z_m)}, \tag{25}$$

where $q(z_m) = \mathsf{E}_{\tilde{\chi}_m}\{q(z_m|\tilde{\chi}_m)\}$. If we also write $p(z_m) = \mathsf{E}_{\chi_m}\{p(z_m|\chi_m)\}$, the free energy $f(s)$ defined in (14) is given under the assumption of the RS ansatz by

$$\begin{aligned} f_{\mathsf{RS}}(s) = {} & \frac{1}{\alpha N}\left[\ln\det\tilde{\boldsymbol{\Omega}} + \mathrm{tr}(\tilde{\boldsymbol{\Omega}}^{-1}\boldsymbol{\Omega}) - \ln\det(s^{-1}\tilde{\boldsymbol{R}})\right] \\ & - \left(\ln\frac{\pi}{\xi} + \frac{\xi}{\eta} + \frac{1}{M}\sum_{m=1}^{M}\int p(z_m)\ln q(z_m)\mathrm{d}z_m\right) \\ & - \xi\varepsilon + \frac{\xi(\xi-\eta)}{\eta}\tilde{\varepsilon}. \end{aligned} \tag{26}$$

If multiple solutions to the coupled fixed point (19)–(24) are found, the one minimizing (26) should be chosen.

Proof: An outline of the derivation is given in Appendix C.

The above result extends some previous works such as [26], [27] in the direction of correlated noise at the receiver and additive transmit-side impairments. It is thus clear that the original GMI term (9) of the MIMO system that suffers from transceiver hardware impairments has an interpretation in terms of an

equivalent *decoupled*[3] scalar system. This decoupled channel has only additive distortions but unlike in the conventional case of replica analysis [26], [27], the transmit-side has its own noise term. It should be remarked, however, that the implicit assumption here is that $f_{\mathsf{RS}}(s) = f(s)$; that is, the system is not replica symmetry breaking (RSB). We leave the RSB case as a possible future work and check the validity of the solution with selected numerical simulations.

For simplicity of presentation, we consider next a few practical special cases of Proposition 1 where the transmit power is the same for all antennas and the noise and distortions at the transmit-side are spatially uncorrelated, namely, $\boldsymbol{\Gamma} = \bar{\gamma}\boldsymbol{I}_M$ and $\boldsymbol{R}_{\boldsymbol{v}} = r_{\boldsymbol{v}}\boldsymbol{I}_M$. The receiver postulates spatially white noise $\tilde{\boldsymbol{R}} = \tilde{r}\boldsymbol{I}_N$ with some variance $\tilde{r} > 0$. This allows us to write

$$\frac{1}{M} I_{\mathsf{GMI}}(\boldsymbol{y};\boldsymbol{x}) = \sup_{\tilde{s}>0} \left\{ f(\tilde{s}) - \alpha^{-1}\tilde{s}\left[N^{-1}\mathrm{tr}(\boldsymbol{R}_{\boldsymbol{w}}) + r_{\boldsymbol{v}}\right] \right\}, \tag{27}$$

where $f(\tilde{s})$ is given by (13) with $s\tilde{\boldsymbol{R}}^{-1}$ replaced by $\tilde{s}\boldsymbol{I}_N$. Furthermore, in this case all variables are identically distributed for all $m = 1, 2, \ldots, M$ so we may omit the subscripts related to m in the following. We still need to fix the input distribution (2) to obtain the parameters (23) and (24). For this, we give two concrete examples: 1) Gaussian signaling; and 2) discrete channel inputs, such as, QAM.

Example 2: Let the channel inputs (2) be IID Gaussian, namely, $p(\boldsymbol{x}) = g(\boldsymbol{x}|\boldsymbol{0}; \bar{\gamma}\boldsymbol{I}_M)$ so that $p(\tilde{\chi}_m) = p(x_m) = g(x|0;\bar{\gamma})$ and $p(\chi_m) = g(\chi_m|0;\bar{\gamma} + r_{\boldsymbol{v}})$ in Proposition 1. The parameter ξ can then be obtained explicitly as

$$\xi = \frac{\bar{\gamma}\tilde{s}(1-\alpha) - \alpha + \sqrt{4\alpha\bar{\gamma}\tilde{s} + \left[\bar{\gamma}\tilde{s}(1-\alpha) - \alpha\right]^2}}{2\alpha\bar{\gamma}}, \tag{28}$$

while η and ε are obtained by solving the coupled fixed point equations

$$\eta = \frac{1}{\alpha\left[N^{-1}\mathrm{tr}(\boldsymbol{R}_{\boldsymbol{w}}) + \varepsilon\right]}, \tag{29}$$

$$\varepsilon = \frac{\eta r_{\boldsymbol{v}} + \bar{\gamma}(\eta + \xi^2\bar{\gamma})}{\eta(1+\xi\bar{\gamma})^2} = \frac{\bar{\gamma} + r_{\boldsymbol{v}}}{(1+\xi\bar{\gamma})^2} + \frac{1}{\eta(1+1/\xi\bar{\gamma})^2}. \tag{30}$$

Additional algebra shows that for IID Gaussian inputs, the free energy (26) reduces to

$$f_{\mathsf{RS}}(\tilde{s}) = \frac{1}{\alpha}\left(\frac{\xi}{\eta} + \ln\tilde{s} + \ln\frac{1}{\alpha\xi}\right) - \xi\varepsilon + \ln(1+\xi\bar{\gamma}) + \frac{\xi r_{\boldsymbol{v}}}{1+\xi\bar{\gamma}}. \tag{31}$$

Note that the expression for parameter $\tilde{\varepsilon}$ in (24) is not explicitly given here but it is implicitly a part of (28) due to relations (20) and (22).

The computational formula for obtaining the GMI with the above example is detailed in Table I. Notice that there are two non-trivial steps in the algorithm: 1) the optimization over $\tilde{s} > 0$; and 2) the problem of solving a system of two nonlinear equations with two unknowns. The first difficulty is not specific to the current study and is present in any work that considers GMI as means to analyze mismatched decoding. The computational complexity of the second problem is negligible compared to the original task of taking an expectation over the channel matrices in (11). Indeed, a typical solution for η and ε is obtained after some tens of iterations of an iterative substitution method.

TABLE I
HOW TO OBTAIN GMI FOR GAUSSIAN SIGNALING FROM EXAMPLE

1) Choose the parameters that define the MIMO system of interest, namely, antenna ratio $\alpha = M/N$, transmit- and receive-side distortion plus noise covariance matrices $\boldsymbol{R}_{\boldsymbol{v}} = r_{\boldsymbol{v}}\boldsymbol{I}_M$ and $\boldsymbol{R}_{\boldsymbol{w}}$, respectively, and the average transmit power per antenna $\bar{\gamma}$. Let also the optimization parameter $\tilde{s} > 0$ be given.
2) Plug the values of $\{\alpha, \bar{\gamma}, \tilde{s}\}$ to (28) and obtain ξ.
3) Insert ξ along with the rest of the necessary parameters in (29) and (30), and solve η numerically, e.g., using an iterative substitution method.
4) Use the solutions of ξ and η in (31) to obtain the free energy.
5) Optimize (27) over $\tilde{s} > 0$.

For the high-SNR case where $\bar{\gamma} \to \infty$ for a fixed covariance matrix $\boldsymbol{R}_{\boldsymbol{w}}$, the result in Example 2 can be further simplified as shown in Example 3 below.

Example 3: Let us consider the case of Gaussian signaling as given in Example 2 in the limit $\bar{\gamma} \to \infty$. We assume for simplicity (see, e.g., [10]) that $\boldsymbol{R}_{\boldsymbol{w}} = r_{\boldsymbol{w}}\boldsymbol{I}_N$ and $r_{\boldsymbol{v}} = \bar{\gamma}\kappa^2$ where $\kappa > 0$ and $r_{\boldsymbol{w}} > 0$ are fixed and finite parameters. At high-SNR, there are two possibilities for the parameter $\tilde{s} = s/\tilde{r}$ in the GMI: 1) the optimal value of $\tilde{s}$ is a strictly positive constant; and 2) the value of $\tilde{s}$ goes to zero when $\bar{\gamma} \to \infty$. For the first case, $M^{-1}I^{(s)}_{\mathsf{GMI}}(\boldsymbol{y};\boldsymbol{x}) \to -\infty$ so to obtain a consistent solution for the fixed point equations, the parameter $\tilde{s}$ has to be inversely proportional to $\bar{\gamma}$, i.e., $\tilde{s} = s_{\bar{\gamma}}/\bar{\gamma}$ where $s_{\bar{\gamma}}$ is a strictly positive finite constant. Then $\xi \to 0$ as $\bar{\gamma} \to \infty$, and the normalized GMI reduces to

$$\frac{1}{M} I^{\infty}_{\mathsf{GMI}}(\boldsymbol{y};\boldsymbol{x}) = \sup_{s_{\bar{\gamma}}>0} \left\{ \frac{1}{\alpha}\ln\left(\frac{s_{\bar{\gamma}}}{\alpha\xi_{\bar{\gamma}}}\right) + \ln(1+\xi_{\bar{\gamma}}) + \frac{\kappa^2\xi_{\bar{\gamma}}}{1+\xi_{\bar{\gamma}}} - \frac{s_{\bar{\gamma}}\kappa^2}{\alpha} \right\}, \tag{32}$$

in the limit $\bar{\gamma} \to \infty$. The auxiliary parameter $\xi_{\bar{\gamma}} \triangleq \xi\bar{\gamma} > 0$ is given by

$$\xi_{\bar{\gamma}} = \frac{s_{\bar{\gamma}}(1-\alpha) - \alpha + \sqrt{4\alpha s_{\bar{\gamma}} + \left[s_{\bar{\gamma}}(1-\alpha) - \alpha\right]^2}}{2\alpha}. \tag{33}$$

Compared to the finite-SNR case in Example 2, the GMI is now directly given by (32) and there are no fixed-point equations that need to be solved.

The next example provides explicit formulas for the computation of GMI given finite discrete constellations, such as, PSK or QAM.

Example 4: Let $\mathcal{A}$ be a discrete modulation alphabet with fixed and finite cardinality $|\mathcal{A}|$ and consider the GMI (27). Let the channel inputs x_m be drawn independently and uniformly

[3]This decoupling property is ubiquitous in replica analysis (see for example [26], [27]) as well as in random matrix theory (see [32], [33] and references therein), and is one of the key reasons why the asymptotic methods provide computationally feasible solutions for complex problems.

from $\mathcal{A}$. The parameters of the decoupled channel model in Proposition 1 can be obtained by first solving ξ and $\tilde{\varepsilon}$ from

$$\xi = \frac{\tilde{s}}{\alpha(1+\tilde{s}\tilde{\varepsilon})}, \tag{34}$$

$$\tilde{\varepsilon} = \bar{\gamma} - \int q(z)\,|\langle \tilde{x}\rangle_q|^2\,\mathrm{d}z, \tag{35}$$

using the following definitions for the decoupled estimator and the postulated channel probability

$$\langle \tilde{x}\rangle_q = \frac{1}{q(z)|\mathcal{A}|}\sum_{\tilde{x}\in\mathcal{A}} \tilde{x}\, g(z|\tilde{x};\xi^{-1}), \tag{36}$$

$$q(z) = \frac{1}{|\mathcal{A}|}\sum_{x\in\mathcal{A}} g(z|x;\xi^{-1}), \tag{37}$$

respectively. Note that this implies solving two parameters from two nonlinear equations and can be done, for example, by using an iterative substitution method. After obtaining the solutions for ξ (and $\tilde{\varepsilon}$), the rest of the parameters can be obtained by solving the two coupled equations

$$\eta = \frac{1}{\alpha\left[N^{-1}\mathrm{tr}(\boldsymbol{R}_{\boldsymbol{w}})+\varepsilon\right]}, \tag{38}$$

$$\varepsilon = \mathsf{E}\left\{\left|v + x - \langle \tilde{x}\rangle_q\right|^2\right\}, \tag{39}$$

for η and ε, where the expectation is w.r.t. the true joint probability of $\{x, v, z\}$. Finally, the free energy reads

$$f_{\mathsf{RS}}(\tilde{s}) = \frac{1}{\alpha}\left(\frac{\xi}{\eta} + \ln\tilde{s} + \ln\frac{1}{\alpha\xi}\right) - \xi\varepsilon + \frac{\xi(\xi-\eta)}{\eta}\tilde{\varepsilon} - \left(\frac{\xi}{\eta} + \ln\frac{\pi}{\xi} + \int p(z)\ln q(z)\mathrm{d}z\right), \tag{40}$$

where we denoted

$$p(z) = \frac{1}{|\mathcal{A}|}\sum_{x\in\mathcal{A}} g(z|x;\eta^{-1}+r_{\boldsymbol{v}}), \tag{41}$$

for the decoupled PDF of the received signal.

Notice that the form of η in Example 4 is the same as in Example 2, but the parameter ε has now a different structure. Compared to the Gaussian case, the equivalent result for IID discrete channel inputs looks in general more cumbersome. First of all, we need to solve now two sets of equations instead of just one. They both contain terms that involve $|\mathcal{A}|$ summations and there are also two expectations left to evaluate, one in (35) and another in (39). However, both expectations involve only scalar variables. This is in stark contrast to the original problem that involved computing $|\mathcal{A}|^M$ summations for every channel and noise/distortion realization and taking expectation over the channel and noise that are multidimensional integrals. This makes direct Monte Carlo computation of the GMI for discrete signaling in practice infeasible for large constellations and numbers of antennas.

IV. Matched Joint Decoding

A. Definition and the Special Case of Gaussian Signaling

Let us now consider the case of matched decoding where the correct channel transition probability (5) is utilized at the receiver. The first entropy term in (7) reads

$$\mathsf{E}\left\{\ln p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{H})\right\} = -\mathsf{E}_{\boldsymbol{H}}\left\{\ln\det(\boldsymbol{R}_{\boldsymbol{w}} + \boldsymbol{H}\boldsymbol{R}_{\boldsymbol{v}}\boldsymbol{H}^{\mathsf{H}})\right\} - c, \tag{42}$$

where $c = N\ln(\mathrm{e}\pi)$. It should be remarked that there is still an expectation left w.r.t. the channel realizations $\boldsymbol{H}$ in (42). This could be evaluated, for example, using Monte Carlo methods or random matrix theory [32], [33]. For the special case of Gaussian inputs, the identities in Appendix A allow us to partially calculate also the latter entropy term in (7), providing the following result that is useful for Monte Carlo simulations.

Example 5: Let $p(\boldsymbol{x}) = g(\boldsymbol{x}|\boldsymbol{0};\boldsymbol{\Gamma})$.Then,

$$\frac{1}{M}I(\boldsymbol{y};\boldsymbol{x}) = \frac{1}{M}\mathsf{E}_{\boldsymbol{H}}\left\{\ln\det\left(\boldsymbol{R}_{\boldsymbol{w}} + \boldsymbol{H}(\boldsymbol{\Gamma}+\boldsymbol{R}_{\boldsymbol{v}})\boldsymbol{H}^{\mathsf{H}}\right)\right\} - \frac{1}{M}\mathsf{E}_{\boldsymbol{H}}\left\{\ln\det(\boldsymbol{R}_{\boldsymbol{w}} + \boldsymbol{H}\boldsymbol{R}_{\boldsymbol{v}}\boldsymbol{H}^{\mathsf{H}})\right\}, \tag{43}$$

is the normalized ergodic MI for matched decoding.

The above expression is relatively easy to compute also by brute-force Monte Carlo methods since there is only an expectation over the fading. Unfortunately, to the best of our knowledge, the latter entropy term in (7) is mathematically intractable for rigorous methods like random matrix theory when $p(\boldsymbol{x})$ is an arbitrary distribution that satisfies (2). For example, given discrete inputs as in Example 4, calculating $\mathsf{E}\{\ln p(\boldsymbol{y}|\boldsymbol{H})\}$ and combining it with (42) reduces the MI to (44), shown at the bottom of the page. This form is computationally very complex and can be evaluated using Monte Carlo methods only for small number of antennas and simple constellations. To obtain a result for general input distribution $p(\boldsymbol{x})$ that has lower computational complexity, we resort to the replica method (see Appendix B). As before, the results that follow have been written in a simplified form where the assumption of LSL is suppressed for notational simplicity.

B. Analytical Results via the Replica Method

Proposition 2: Let us write for notational convenience

$$\chi_m = x_m + v_m, \qquad m = 1,\ldots,M, \tag{45}$$

$$I(\boldsymbol{y};\boldsymbol{x}) = M\ln|\mathcal{A}| - N - \frac{1}{|\mathcal{A}|}\sum_{\boldsymbol{x}\in\mathcal{A}^M}\mathsf{E}_{\boldsymbol{v},\boldsymbol{w},\boldsymbol{H}}\left\{\ln\left(\sum_{\tilde{\boldsymbol{x}}\in\mathcal{A}^M}\mathrm{e}^{-[\boldsymbol{H}(\boldsymbol{x}-\tilde{\boldsymbol{x}}+\boldsymbol{v})+\boldsymbol{w}]^{\mathsf{H}}(\boldsymbol{R}_{\boldsymbol{w}}+\boldsymbol{H}\boldsymbol{R}_{\boldsymbol{v}}\boldsymbol{H})^{-1}[\boldsymbol{H}(\boldsymbol{x}-\tilde{\boldsymbol{x}}+\boldsymbol{v})+\boldsymbol{w}]}\right)\right\} \tag{44}$$

where $x_m \sim p(x_m)$ and $v_m \sim g(v_m|0; r_{\boldsymbol{v}}^{(m)})$ are independent for all m. Let

$$p(z_m|\chi_m) = g(z_m|\chi_m; \eta^{-1}), \tag{46}$$

be a conditional PDF of an AWGN channel whose input is (45) and noise variance is η^{-1}. The conditional mean estimator of χ_m received over this channel reads

$$\langle \chi_m \rangle = \frac{\mathsf{E}_{\chi_m}\left\{\chi_m p(z_m|\chi_m)\right\}}{\mathsf{E}_{\chi_m}\left\{p(z_m|\chi_m)\right\}}, \tag{47}$$

where the parameter η is given, along with another parameter ε, as the solution to the coupled fixed point equations

$$\eta = \frac{1}{\alpha N}\mathrm{tr}\left[(\boldsymbol{R}_{\boldsymbol{w}} + \varepsilon \boldsymbol{I}_N)^{-1}\right], \tag{48}$$

$$\varepsilon = \frac{1}{M}\sum_{m=1}^{M}\left[\bar{\gamma}_m + r_{\boldsymbol{v}}^{(m)} - \mathsf{E}\left|\langle \chi_m \rangle\right|^2\right]. \tag{49}$$

If we also define a second set of parameters η' and ε' that are solutions to the coupled fixed point equations

$$\eta' = \frac{1}{\alpha N}\mathrm{tr}\left[(\boldsymbol{R}_{\boldsymbol{w}} + \varepsilon' \boldsymbol{I}_N)^{-1}\right], \tag{50}$$

$$\varepsilon' = \frac{1}{M}\sum_{m=1}^{M}\frac{r_{\boldsymbol{v}}^{(m)}}{1+\eta' r_{\boldsymbol{v}}^{(m)}}, \tag{51}$$

the per-stream MI is finally given by

$$\frac{1}{M}I(\boldsymbol{y};\boldsymbol{x}) = \frac{\ln\det(\boldsymbol{R}_{\boldsymbol{w}} + \varepsilon \boldsymbol{I}_N) - \ln\det(\boldsymbol{R}_{\boldsymbol{w}} + \varepsilon' \boldsymbol{I}_N)}{\alpha N}$$
$$-(\eta\varepsilon - \eta'\varepsilon') + \frac{1}{M}\sum_{m=1}^{M}\left[I(z_m;\chi_m) - \ln\left(1+\eta' r_{\boldsymbol{v}}^{(m)}\right)\right], \tag{52}$$

where

$$I(z_m;\chi_m) = -1 - \ln\frac{\pi}{\eta} - \int p(z_m)\ln p(z_m)\mathrm{d}z_m, \tag{53}$$

is the MI of the Gaussian channel defined by (45) and (46).

Proof: The result can be obtained using Appendix B for two separate MIMO channels. For the first one, we replace everywhere $\boldsymbol{x}_a \to \boldsymbol{x}_a + \boldsymbol{v}_a,\ a = 0, 1, \ldots, u$ and an application of the RM provides the equations (45)–(49). The formulas (50)–(53), on the other hand, are obtained by substituting $\boldsymbol{x}_a \to \boldsymbol{v}_a,\ a = 0, 1, \ldots, u$ in Appendix B.

Just like Proposition 1 in Section III, Proposition 2 is valid for any input distribution that satisfies (2). The solutions to the coupled (48) and (49) as well as (50) and (51) can be obtained numerically, e.g., using an iterative substitution method.

For concreteness, we again give examples for Gaussian and discrete signaling when the noise plus distortion is spatially white $\boldsymbol{R}_{\boldsymbol{v}} = r_{\boldsymbol{v}}\boldsymbol{I}_M$ and transmit power is uniformly allocated $\boldsymbol{\Gamma} = \bar{\gamma}\boldsymbol{I}_M$. This makes the channels $m = 1, 2, \ldots, M$ identically distributed so we omit the subscript m in the following.

Example 6: Let $\boldsymbol{R}_{\boldsymbol{v}} = r_{\boldsymbol{v}}\boldsymbol{I}_M$ and consider the special case of Gaussian inputs $p(\boldsymbol{x}) = g(\boldsymbol{x}|\boldsymbol{0}; \bar{\gamma}\boldsymbol{I}_M)$. Then

$$I(z;\chi) = \ln\left[1 + \eta(\bar{\gamma} + r_{\boldsymbol{v}})\right], \tag{54}$$

$$\varepsilon = \frac{\bar{\gamma} + r_{\boldsymbol{v}}}{1 + \eta(\bar{\gamma} + r_{\boldsymbol{v}})}, \tag{55}$$

and the rest of the parameters are given in Proposition 2.

We next consider the high-SNR case $\bar{\gamma} \to \infty$ as in Example 3 and compare it to the result obtained in [10] using completely different mathematical methods.

Example 7: For the case $\boldsymbol{R}_{\boldsymbol{w}} = r_{\boldsymbol{w}}\boldsymbol{I}$, $\boldsymbol{R}_{\boldsymbol{v}} = \kappa^2\bar{\gamma}\boldsymbol{I}$ (see, e.g., [10]) we find that if $\alpha \leq 1$ then $\bar{\gamma} \to \infty$ yields $\eta = \eta'$ and $\varepsilon = \varepsilon'$. The high SNR limit is therefore

$$\frac{1}{M}I^{\infty}(\boldsymbol{y};\boldsymbol{x}) = \log\left(\frac{1+\kappa^2}{\kappa^2}\right), \qquad \alpha \leq 1. \tag{56}$$

For the case $\alpha > 1$, both η and η' tend to zero at high SNR while ε and ε' grow without bound. This is not yet sufficient to solve (52). However, combining this with the relations $\eta'\varepsilon' = \eta\varepsilon$ and $\varepsilon' = \varepsilon\frac{\kappa^2}{1+\kappa^2}$, that hold in the limit $\bar{\gamma} \to \infty$ for $\alpha > 1$, provides the second part of the high SNR result

$$\frac{1}{M}I^{\infty}(\boldsymbol{y};\boldsymbol{x}) = \frac{1}{\alpha}\log\left(\frac{1+\kappa^2}{\kappa^2}\right), \qquad \alpha > 1. \tag{57}$$

The asymptotic mutual information expressions in (56) and (57) coincide exactly with the results obtained previously in [10], as expected.

Example 8: If the channel inputs are from a discrete alphabet $\mathcal{A}$ as in Example 4, the parameter ε in (49) is obtained using

$$\langle \chi \rangle = \frac{1}{p(z)}\sum_{x\in\mathcal{A}}\left[\frac{1}{|\mathcal{A}|}g(z|x;\eta^{-1}+r_{\boldsymbol{v}})\left(\frac{x+\eta r_{\boldsymbol{v}} z}{1+\eta r_{\boldsymbol{v}}}\right)\right], \tag{58}$$

$$\mathsf{E}\left|\langle\chi\rangle\right|^2 = \int p(z)\mathsf{E}\left\{\left|\langle\chi\rangle\right|^2\right\}\mathrm{d}z, \tag{59}$$

in Proposition 2. Here $p(z)$ is given by (41) and $\langle\chi\rangle$ denotes the conditional mean estimator of (45) from the observations (46).The related MI term reads by definition

$$I(z;\chi) = \ln\left(\frac{\eta}{\mathrm{e}\pi}\right) - \int p(z)\ln p(z)\,\mathrm{d}z. \tag{60}$$

Both (49) and (60) need, in general, to be solved numerically.

V. Numerical Examples

In the following, assume for simplicity that $\boldsymbol{\Gamma} = \bar{\gamma}\boldsymbol{I}$, $\boldsymbol{R}_{\boldsymbol{w}} = \boldsymbol{I}$ and $\boldsymbol{R}_{\boldsymbol{v}} = \kappa^2\bar{\gamma}\boldsymbol{I}$, where $\kappa = 10^{\mathsf{EVM}/20}$ and EVM denotes the EVM of the transmitter in decibels. The SNR without transmit-side noise is therefore simply $\bar{\gamma}$, or in decibels, $\bar{\gamma}_{\mathsf{dB}} = 10\log_{10}(\bar{\gamma})$. Furthermore, all cases assume a symmetric antenna setup $\alpha = M/N = 1$ for simplicity.

The first numerical experiment plotted in Fig. 2 examines the accuracy of the asymptotic analytical results when applied to finite-sized systems. The EVM is fixed to a rather pessimistic value $\mathsf{EVM} = -10$ dB to highlight the differences between the ideal and imperfect hardware configurations. The normalized

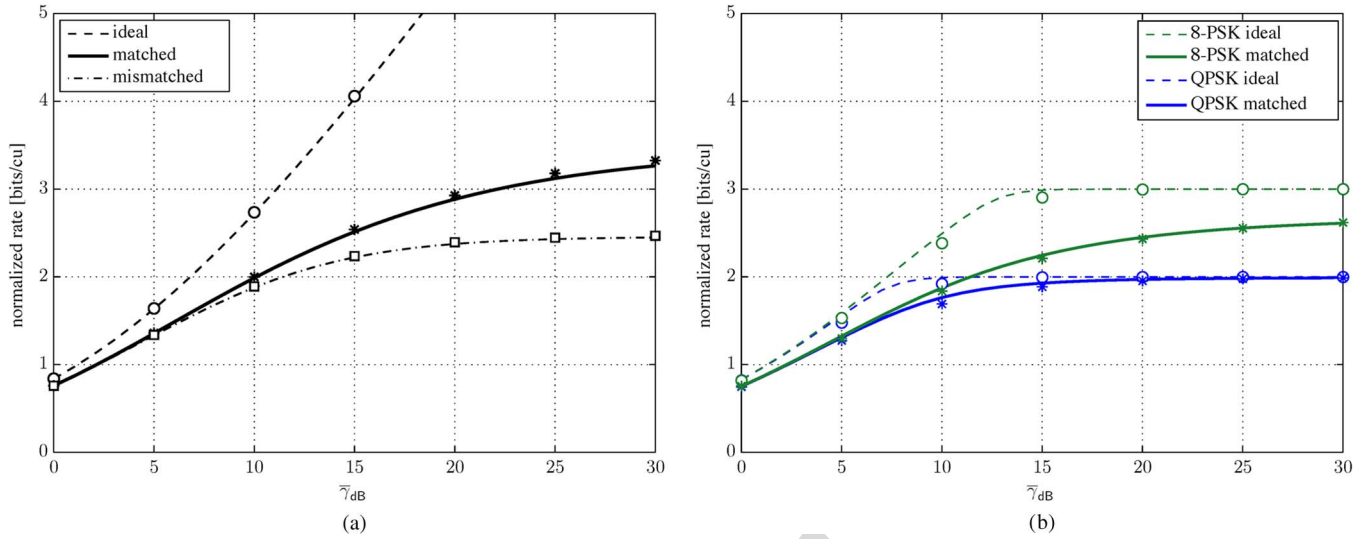

Fig. 2. Normalized rate $M^{-1}I(\boldsymbol{y};\boldsymbol{x})$ in bits per channel use (cu) vs. SNR for MIMO transmission. Lines for replica results and markers for Monte Carlo simulations for $M = N = 4$ antenna configuration. Selected cases of ideal hardware $\mathsf{EVM} = -\infty$ dB and hardware impairments ($\mathsf{EVM} = -10$ dB) with matched and mismatched decoding are plotted. (a) Gaussian signaling. (b) Discrete signaling.

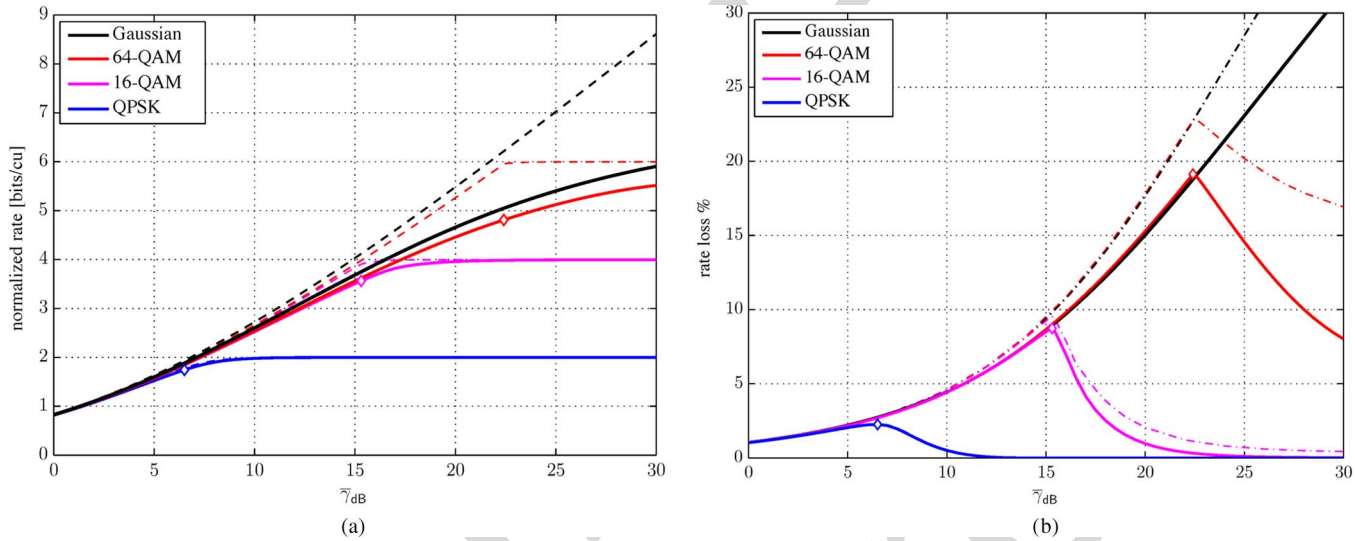

Fig. 3. Performance of a MIMO system with $M = N$ antennas and given ideal ($\mathsf{EVM} = -\infty$ dB) or non-ideal hardware ($\mathsf{EVM} = -20$ dB) for different signaling methods. Markers depict the points where discrete constellations and matched decoding with hardware impairments experience the maximum rate losses compared to the ideal cases. (a) Normalized rate $M^{-1}I(\boldsymbol{y};\boldsymbol{x})$ given ideal hardware (dashed lines) or non-ideal hardware and matched decoding (solid lines). (b) Rate loss percentage compared to ideal hardware for matched (solid lines) and mismatched (dash-dotted lines) decoding.

rate is shown using the asymptotic replica analysis (lines) and Monte Carlo simulations (markers) for a finite-size symmetric antenna setup with $M = N = 4$. In the case of Gaussian signaling, plotted in Fig. 2(a), the analytical approximations for the normalized rate $M^{-1}I(\boldsymbol{y};\boldsymbol{x})$ given by Examples 2 and 5 are quite good when compared to the finite size simulations based on Examples 1 and 5. For discrete signaling depicted in Fig. 2(b) we have plotted only the case of matched decoding due to the computational complexity of Monte Carlo simulations in the mismatched case. The gap between asymptotic result presented in Example 8 and Monte Carlo averaging of (44) is similar to the Gaussian case for both constellations. Fig. 2 shows that the analytical approximation given by the replica method is reasonably good already at $M = N = 4$, even though formally the limit $M, N \to \infty$ is required by the analysis. Note that Monte Carlo simulation of (44) has exponential computational complexity and the system size cannot be increased much higher than $M = 4$. Therefore, the rest of the examples are generated using only the analytical results given in the previous sections.

Fig. 3 illustrates the performance of an $M = N$ MIMO system for a more realistic EVM value $\mathsf{EVM} = -20$ dB. For the case of matched decoding we used Examples 6 and 8, while Examples 2 and 4 were used to obtain the curves representing mismatched decoding. In Fig. 3(a), the normalized rate $M^{-1}I(\boldsymbol{y};\boldsymbol{x})$ is depicted as a function of SNR $\bar{\gamma}$ in decibels. For clarity of presentation, we have plotted only the ideal case and the case of non-ideal hardware with matched decoding. The Gaussian curves (black lines) here are the same as the simulation curves in [10, Fig. 2] given the parameter value

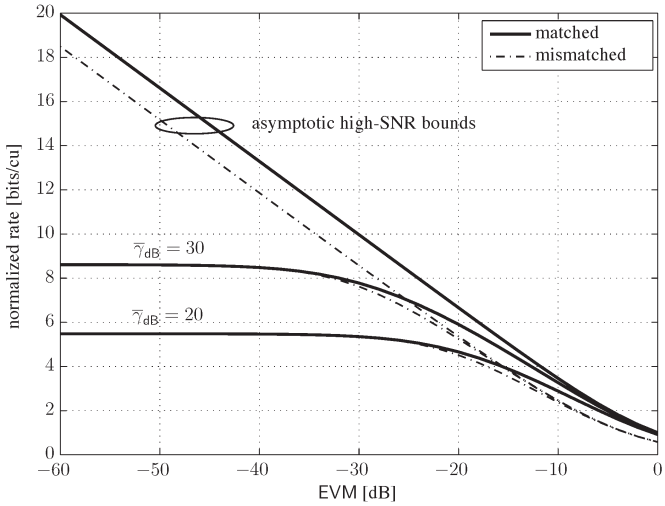

Fig. 4. Normalized rate $M^{-1}I(\boldsymbol{y};\boldsymbol{x})$ in bits per channel use vs. EVM in decibels for MIMO transmission with Gaussian signaling. Solid lines for matched decoding and dash-dotted lines for mismatched decoding.

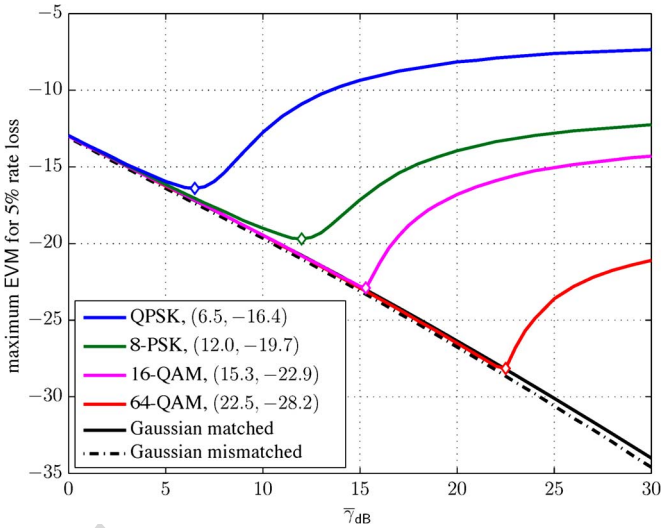

Fig. 5. Maximum allowed EVM in decibels for matched decoding so that the system experiences at most 5% loss in rate compared to the case with ideal hardware ($\mathsf{EVM} = -\infty$ dB). Markers depict the worst case EVM requirement for the discrete constellations and parenthesis in the legend provide the respective values as $(\bar{\gamma}_{\mathrm{dB}}, \mathsf{EVM})$. All discrete cases correspond to matched joint decoding at the receiver.

$\kappa = 0.1$. Apart from 64-QAM and Gaussian signaling, the figure seems to imply that lower order constellations exhaust the source entropy before the transmit-side noise has any significant effect for this choice of EVM. To see more clearly the effect of transmit noise, Fig. 3(b) shows the rate loss (in percentage) for the case with transmit noise $\mathsf{EVM} = -20$ dB when compared to the ideal case $\mathsf{EVM} = -\infty$ dB. The solid lines represent again matched decoding while dash-dotted lines are for mismatched decoding. As expected, mismatched decoding reduces the achievable rate when compared to matched decoding, but the effect is relatively minor when compared to the total rate loss caused by the presence of transmit noise itself. The markers depict the points where maximum relative rate loss is experienced for matched decoding. The same markers are also plotted in Fig. 3(a) for comparison.

In Fig. 4 we have plotted the asymptotic high-SNR results given in Examples 3 and 7. Note that given a finite value of EVM, the normalized rates for matched and mismatched decoding have a gap in this case. For more realistic, but still quite high SNR values of 20 dB and 30 dB, the two decoding strategies converge to the same value roughly when $\bar{\gamma}_{\mathrm{dB}} < -\mathsf{EVM}$. The apparent discrepancy is explained by recalling that the asymptotic cases assume $\bar{\gamma} \to \infty$ for a fixed and nonzero EVM and, thus, as a finite SNR approximation implies $\bar{\gamma} \gg 1/\kappa^2$. As may be observed from the lower right corner of the figure, the SNR values 20 dB and 30 dB have also a similar behavior near $\bar{\gamma} \gg 1/\kappa^2$. Thus, the high-SNR result is consistent with the finite-SNR cases.

It is important to guarantee certain performance when designing a system. The maximum EVM that leads to at most 5% rate loss (as compared to having ideal hardware) for a fixed input distribution and different given SNRs is plotted in Fig. 5. For Gaussian signaling we have plotted both the matched and mismatched cases while discrete cases assume matched joint decoding for simplicity. As expected, the EVM requirement for Gaussian signaling is a monotonically decreasing, but not linear, function of SNR. A simple linear approximation that provides a lower bound for the case of Gaussian signaling with matched decoding is given by

$$\mathsf{EVM} = -0.7 \cdot \bar{\gamma}_{\mathrm{dB}} - 13, \tag{61}$$

in decibels for the depicted region. This can be used as a simple rule-of-thumb for worst-case maximum allowed EVM in the system, although we recommend that EVM target values obtained in this way are always rounded down to 1–5 dB precision to include extra safety margin. For discrete constellations, the EVM requirement first follows the Gaussian case but then starts to get looser at higher SNRs. This is expected, as can be observed from Fig. 3(a), since the maximum achievable rate for a discrete constellation saturates at a certain SNR when the input distribution runs out of entropy. After this point, the rate loss can be held fixed for increasing SNR by increasing the transmit-side noise variance, or EVM, accordingly.

VI. Conclusion and Future Work

Considering a 'binoisy' channel model, we have derived asymptotic expressions for the achievable rate of SU-MIMO systems suffering from transceiver hardware impairments. For matched decoding, where the receiver is designed and implemented explicitly based on the generalized system model, expressions for the ergodic mutual information between the channel inputs and outputs have been given. In addition, a simplified receiver that neglected the hardware imperfections and performed mismatched detection and decoding has been studied via generalized mutual information. The mathematical expressions provided in the paper cover practical discrete modulation schemes, such as, quadrature amplitude modulation, as well as Gaussian signaling. The numerical results showed that for realistic system parameters, the effects of transmit-side noise and mismatched decoding become significant only at high modulation orders. Furthermore, the effect of mismatched

decoding was found to be relatively minor compared to the total rate loss caused by the presence of transmit noise itself. The results were also used to identify the maximum EVM values that allows for certain system operation.

A. Future Work

For the ease of exposition, the present paper considered the analysis of a relatively simple SU-MIMO system where the channel had IID Gaussian elements. An extension of the present replica analysis to correlated or Ricean fading channels can be done by following, for example, the analysis in [29] and [30], respectively. Establishing the effects of transmit-side noise for the cases of correlated and non-Rayleigh fading channels is an important avenue for future work.

As a further extension, it is important to investigate whether similar phenomena as observed in the present paper are present also for more complicated signal models with discrete channel inputs. Such systems already analyzed in the ideal setting with the replica method include, for example, multiuser MIMO and base station collaboration [34], channels with interference and precoding [35] and K-hop relay channels [36]. Combining the ideas from the present paper and [34]–[36] would provide a possible approach to solving such cases.

Appendix A
Useful Results

Here we collect useful results that are used often in the paper. All matrix operations below are implicitly assumed to be well-defined. The *Gaussian integration* formula for vector $\boldsymbol{x} \in \mathbb{C}^N$ is given by (see, e.g., [37, Appendix I])

$$\frac{1}{\pi^N} \int \mathrm{e}^{-\boldsymbol{x}^{\mathsf{H}} \boldsymbol{M} \boldsymbol{x} + 2\Re\{\boldsymbol{b}^{\mathsf{H}} \boldsymbol{x}\}} \mathrm{d}\boldsymbol{x} = \frac{1}{\det(\boldsymbol{M})} \mathrm{e}^{\boldsymbol{b}^{\mathsf{H}} \boldsymbol{M}^{-1} \boldsymbol{b}}, \tag{62}$$

and used in Sections II–IV and Appendix C. Similarly, the *matrix inversion lemma* [38]

$$(\boldsymbol{W}^{-1} + \boldsymbol{U}\boldsymbol{T}^{-1}\boldsymbol{V}^{\mathsf{H}})^{-1} = \boldsymbol{W} - \boldsymbol{W}\boldsymbol{U}(\boldsymbol{T} + \boldsymbol{V}^{\mathsf{H}}\boldsymbol{W}\boldsymbol{U})^{-1}\boldsymbol{V}^{\mathsf{H}}\boldsymbol{W}, \tag{63}$$

and the related *determinant identity*

$$\det(\boldsymbol{W}^{-1} + \boldsymbol{U}\boldsymbol{T}^{-1}\boldsymbol{V}^{\mathsf{H}}) = \det(\boldsymbol{T} + \boldsymbol{V}^{\mathsf{H}}\boldsymbol{W}\boldsymbol{U}) \det(\boldsymbol{W}^{-1}) \det(\boldsymbol{T}^{-1}), \tag{64}$$

are employed several times in the paper.

Appendix B
Replica Method

Consider a function Z that maps RVs to positive real numbers[4] and define two sets of RVs, $V \in \mathcal{V}$ and $X \in \mathcal{X}$, with joint probability $P_{V,X}$. Assume for convenience that $P_{V,X}$ can be described in terms of a joint PDF $p(V, X)$ and denote the marginal PDFs of X and V $p_X(X)$ and $p_V(V)$, respectively. Then, both in statistical mechanics and communication theory, we often encounter a formula

$$\begin{aligned} f &= -\frac{1}{M} \mathsf{E}_V \left\{ \ln \mathsf{E}_X \left\{ Z(V, X) \right\} \right\} \\ &= -\frac{1}{M} \int_{\mathcal{V}} p_V(V) \ln Z(V) \mathrm{d}V, \end{aligned} \tag{65}$$

where $Z(V) = \int_{\mathcal{X}} p_X(X) Z(V, X) \mathrm{d}X$. In physics jargon, the variables V are said to be *quenched* and the quantity (65) is the average *free energy density* of a system whose *partition function* is $Z(V)$. Two concrete examples of (65) are:

1) Let $Z(V, X) = g(\boldsymbol{y} | \boldsymbol{H}\boldsymbol{x}; \boldsymbol{R}_{\boldsymbol{w}})$ be the conditional PDF of the observation in an ideal MIMO channel with $V = \{\boldsymbol{y}, \boldsymbol{H}\}$ and $X = \{\boldsymbol{x}\}$, where $\boldsymbol{x}$ has IID elements from a discrete modulation set $\mathcal{A}$, such as PSK or QAM. Then (65) represents a normalized version of the second term in (7), namely, the (normalized) total *entropy* of the received signal $\boldsymbol{y}$ given a realization of $\boldsymbol{H}$ and averaged over all possible realizations of $\boldsymbol{H}$.
2) Let $Z(V, X) = \mathrm{e}^{\beta \boldsymbol{\sigma}^{\mathsf{H}} \boldsymbol{J} \boldsymbol{\sigma}}$, where $\beta > 0$ denotes the inverse temperature, $V = \boldsymbol{J} \in \mathbb{R}^{M \times M}$ a coupling matrix and $X = \boldsymbol{\sigma} \in \{\pm 1\}^M$ a spin configuration. If $p_V(V)$ is a uniform probability over $\boldsymbol{\sigma}$ and $\boldsymbol{J}$ has, e.g., IID Gaussian elements, then (65) is the average *free energy density* of a mean-field Ising spin glass in the absence of external field (up to trivial constants).

In both cases, f captures important properties of the system at hand and obtaining a computable formula for (65) would be of great interest. This seems infeasible though since the number of terms in the expectation is exponential in M.

A. Outline of the Replica Method

One method for solving (65) is the *replica method* (RM) from equilibrium statistical mechanics. While the RM is extremely versatile, it unfortunately lacks mathematical rigor in some parts (see, e.g., [24]–[26]). However, due to its success both in physics and engineering, it is generally agreed to be at least a valuable starting point for analysis of problems that seem otherwise too difficult to handle. A cursory overview of literature about the RM inside a specific field or topic may paint the picture that the RM is a fixed set of mathematical methods which can be applied to any suitable problem at hand. This is not entirely accurate and conceptually the RM can be seen more like a systematic way of turning a very difficult problem into a more manageable one than a set of specific tools that actually solve the problem. Indeed, the mathematical methods that are used at different stages of the RM can often be chosen from a variety of choices, although it is very common to have some form of *large deviations theory* as part of the analysis (see Step 2 below). Thus, instead of trying to be entirely general, we describe next (one form of) the steps taken in the RM in the context of the first example above.

[4] In the following we refrain differentiating random variables and their realizations for notational convenience. Also, Z and, as a result, f can depend on some parameters (non-random variables) that are not explicitly stated.

Step 1 (Replica Trick): Consider (65) and write it as

$$\begin{aligned} f &= -\frac{1}{M}\lim_{u\to 0^+}\frac{\partial}{\partial u}\ln \mathsf{E}_V\left\{[Z(V)]^u\right\} \\ &= -\frac{1}{M}\lim_{u\to 0^+}\frac{\partial}{\partial u}\ln \mathsf{E}_V\left\{\left(\sum_{\boldsymbol{x}\in\mathcal{A}^M} p_X(\boldsymbol{x})Z(V,\boldsymbol{x})\right)^u\right\} \\ &= -\frac{1}{M}\lim_{u\to 0^+}\frac{\partial}{\partial u}\ln \Xi(u), \end{aligned} \tag{66}$$

where $u\in\mathbb{R}$ and we denoted $\Xi(u)=\mathsf{E}_V\{[Z(V)]^u\}$. Then, *assume* that we can treat u as an integer when we take the expectation, namely,

$$\begin{aligned} \Xi(u) &= \mathsf{E}_V\left\{\prod_{a=1}^{u}\sum_{\boldsymbol{x}_a\in\mathcal{A}^N} p_X(\boldsymbol{x}_a)Z(V,\boldsymbol{x}_a)\right\} \\ &= \frac{1}{\pi^{uN}(\det \boldsymbol{R}_{\boldsymbol{w}})^u} \\ &\quad\times \mathsf{E}_V\left\{\sum_{\{\boldsymbol{x}_a\}}\prod_{a=1}^{u}\left[\mathrm{e}^{-(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x}_a)^{\mathsf{H}}\boldsymbol{R}_{\boldsymbol{w}}^{-1}(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x}_a)}p_X(\boldsymbol{x}_a)\right]\right\}, \end{aligned} \tag{67}$$

where the summation in the last expression is over the set $\{\boldsymbol{x}_a\}_{a=1}^u$. After taking the expectations, if we manage to write (67) in a form that does not explicitly force u to be an integer, *invoke analytical continuity* to extend u to real numbers.

The step above is at the very heart of the RM. It is important to realize that the equalities in (66) are provably true if differentiation under the integral sign is permitted and $u\in\mathbb{R}$. The part lacking rigorous mathematical justification is (67), especially when combined with the next two steps. Somewhat surprisingly, however, the end results of RM can sometimes be proved to be exact. Examples of such cases are: MIMO channel with Gaussian inputs, random energy model (REM) and Sherrington-Kirkpatrick model of spin glasses (see, e.g., [24]–[27], and references therein).

Step 2 (Large System Limit): Let the system approach the LSL, that is, the dimensions of the channel matrix $\boldsymbol{H}$ grow without bound at a finite and fixed ratio $\alpha=M/N>0$. Furthermore, assume that the limits w.r.t. u and M commute, so that we can first calculate the expectations in (67) in the LSL and then let $u\to 0$, as in (66).

The LSL assumption is natural in equilibrium statistical mechanics (e.g., the second example above), where the systems contain usually very large numbers of interacting particles M. In communication theory, the equivalent would be, e.g., a MIMO systems with large antenna arrays or a CDMA with large number of simultaneous users. It is in fact quite common to write the LSL assumption directly as a part of the replica trick in (66). The steps are separated here since the replica trick could also be used for finite sized systems. Due to mathematical difficulty of such cases, however, both steps are usually found together. The assumption of commuting limits is typically postulated *a priori* and rigorous justification of this step is beyond the scope of the present paper.

Let us denote the true transmitted vector $\boldsymbol{x}_0$, so that $\boldsymbol{y}=\boldsymbol{H}\boldsymbol{x}_0+\boldsymbol{w}$ is the generating model for the observation $\boldsymbol{y}$ and we can equivalently write $V=\{\boldsymbol{w},\boldsymbol{x}_0,\boldsymbol{H}\}$. Returning then to Step 1, we note that although the *replicated* vectors $\{\boldsymbol{x}_a\}_{a=1}^u$ act as IID RVs drawn according to p_X in (67) when conditioned on V, they can be correlated if not conditioned on V. We examine this through the empirical correlations between the vectors in the set $X_{u+1}=\{\boldsymbol{x}_a\}_{a=0}^u$ using *overlap matrix* $\boldsymbol{Q}\in\mathbb{C}^{(u+1)\times(u+1)}$, whose (a,b)th element[5] is given by $Q_{a,b}=M^{-1}\boldsymbol{x}_b^{\mathsf{H}}\boldsymbol{x}_a$. Then, the structure that is imposed on $\boldsymbol{Q}$ divides the replica analysis into two rough categories as described below.

Step 3 (Replica Symmetry): The *RS ansatz* or *RS assumption* means that the indexes $a=1,\ldots,u$ are permutation symmetric and $\boldsymbol{Q}$ can be written in terms of four parameters, for example, $Q_{0,0}=p$, $Q_{0,a}=m, a\geq 1$, $Q_{a,a}=q, a\geq 1$, and $Q_{a,b}=q, a\neq b\geq 1$. Note that $\boldsymbol{Q}=\boldsymbol{Q}^{\mathsf{H}}$ by construction. If $\boldsymbol{Q}$ is not of the RS form, it is said to have replica symmetry breaking (RSB) structure whose analysis is much more involved [24], [25].

The importance of the RS assumption will become clear when we present a rough sketch of the analysis of an ideal MIMO channel. We also note that the overlap matrix given in Step 3 allows the "zeroth" index to be treated separately to take into account the possibility that either $\boldsymbol{x}_0$ has different distribution than $\boldsymbol{x}_a$ when $a\geq 1$, or the decoder uses mismatched statistics, i.e., $Z(V,\boldsymbol{x}_a)$ does not match the probability law of the observation $\boldsymbol{y}=\boldsymbol{H}\boldsymbol{x}_0+\boldsymbol{w}$ as in Appendix C. For the simplified case considered below, however, we have $p=Q$ and $m=q$ since the indexes $a=0,1,\ldots,u$ can be treated on equal footing and two parameters are sufficient to define the RS form of $\boldsymbol{Q}$.

Next we give a brief and informal example of replica analysis for an ideal MIMO channel. The reader may be surprised to find out that most of the discussion below deals with details about how to obtain the necessary formulas when we follow the three stages above and not about those stages per se.

B. Average Over the Channel and Noise

The starting point of our replica calculation is (67), where we use the generating model of $\boldsymbol{y}$ to write in the exponential $\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x}_a=\boldsymbol{w}-\boldsymbol{H}(\boldsymbol{x}_0-\boldsymbol{x}_a)$. The first task is then to compute the expectation w.r.t. $\boldsymbol{w}$ and $\boldsymbol{H}$ for a fixed set $X_{u+1}=\{\boldsymbol{x}_a\}_{a=0}^u$ that satisfies the correlations of the RS overlap matrix $\boldsymbol{Q}$. Note that we cannot assume anymore that the vectors in X_{u+1} are independent since we changed the order of expectations in (67) and the average over X_{u+1} is carried out (later) without conditioning on $\boldsymbol{w}$ and $\boldsymbol{H}$. With this in mind, it follows that given X_{u+1}, the set $\{\boldsymbol{H}\boldsymbol{x}_a\}$ consists of CSCG RVs with correlations $\mathsf{E}_{\boldsymbol{H}}\{(\boldsymbol{H}\boldsymbol{x}_a)(\boldsymbol{H}\boldsymbol{x}_b)^{\mathsf{H}}\}=M^{-1}\boldsymbol{x}_b^{\mathsf{H}}\boldsymbol{x}_a\boldsymbol{I}_N=Q_{a,b}\boldsymbol{I}_N$ that are deterministic in the LSL. Thus, we can replace $\{\boldsymbol{H}(\boldsymbol{x}_0-\boldsymbol{x}_a)\}_{a=1}^u$ by a set of CSCG RVs $\{\boldsymbol{\Delta}_a\}_{a=1}^u$ and use Gaussian integration (62) to average over both $\boldsymbol{w}$ and $\{\boldsymbol{\Delta}_a\}_{a=1}^u$ to obtain (for more details, see Appendix C-B).

$$\Xi(u) = \int \mathrm{e}^{NG^{(u)}(\boldsymbol{Q})}\mu(\boldsymbol{Q})\mathrm{d}\boldsymbol{Q}, \tag{68}$$

$$\begin{aligned} G^{(u)}(\boldsymbol{Q}) = &-u\ln\det\left[\boldsymbol{R}_{\boldsymbol{w}}+(Q-q)\boldsymbol{I}_N\right] \\ &-u\ln\pi-\ln(u+1), \end{aligned} \tag{69}$$

[5]The row/column indexes of $\boldsymbol{Q}$ are $0,1,\ldots,u$ so that the correlations are measured also w.r.t. the true transmitted vector $\boldsymbol{x}_0$. Furthermore, due to (2), the empirical correlations can be expected to converge to the true ones in the LSL postulated in Step 2.

where $\boldsymbol{Q}$ should be understood to be in its RS parametrized form and $\mu(\boldsymbol{Q})$ is the PDF of the overlap matrix $\boldsymbol{Q}$.

Remark 2: Firstly, note that due to the RS assumption (Step 3), the function (69) is of a form that does not restrict u to be an integer, as desired. This is one of the reasons why we need to express matrix $\boldsymbol{Q}$ in a parametrized way instead of using it "as-is." Secondly, there is some universality in this derivation and the form (68) is a typical result of replica analysis. In some cases, however, different techniques are needed. One example is non-IID "mixing matrix" that requires direct matrix integration [39], [40].

C. Distribution of the Overlap Matrix and Large Deviations

The second major step in the analysis is to find an explicit formula for $\mu(\boldsymbol{Q})$, i.e., for the probability weight of the set $\{\boldsymbol{x}_a\}_{a=0}^{u}$ that satisfies $Q_{a,b} = M^{-1}\boldsymbol{x}_b^{\mathsf{H}}\boldsymbol{x}_a$. The form of (68) suggest that we should try to represent $\mu(\boldsymbol{Q})$ as an exponential whose argument is linear in N (or M) so that we can employ Laplace's method or the method of steepest descent to evaluate the integral w.r.t. $\boldsymbol{Q}$. If $\boldsymbol{x}_a \in \mathbb{R}^M$, due to (2), the elements of $\boldsymbol{x}_a$ are IID for all $a = 0, 1, \ldots, u$ and μ follows the large deviation principle [25], [41]. Informally this implies[6] $\mu(\boldsymbol{Q}) \asymp \mathrm{e}^{-Mc^{(u)}(\boldsymbol{Q})}$, where the *rate function*

$$c^{(u)}(\boldsymbol{Q}) = \sup_{\tilde{\boldsymbol{Q}}} \left\{ \mathrm{tr}(\boldsymbol{Q}\tilde{\boldsymbol{Q}}) - \lim_{M\to\infty} \frac{1}{M} \ln \phi^{(u)}(\tilde{\boldsymbol{Q}}) \right\}, \quad (70)$$

describes the exponential behavior of the probability,

$$\phi^{(u)}(\tilde{\boldsymbol{Q}}) = \mathsf{E}_{X_{u+1}} \left\{ \exp\left(\sum_{a,b=0}^{u} \tilde{Q}_{a,b} \boldsymbol{x}_b^{\mathsf{H}} \boldsymbol{x}_a \right) \right\}, \quad (71)$$

is the moment generating function (MGF) associated with $\mu(Q)$ and the supremum is over all $(u+1)\times(u+1)$ matrices $\tilde{\boldsymbol{Q}}$ that have the same RS form as Q, that is, $\tilde{Q}_{0,0} = \tilde{p}$, $\tilde{Q}_{0,a} = \tilde{m}, a \geq 1$, $\tilde{Q}_{a,a} = \tilde{Q}, a \geq 1$, and $\tilde{Q}_{a,b} = \tilde{q}, a \neq b \geq 1$. Thus, we can assess (68) in the LSL up to the leading order by using the exponential form of μ and Laplace's method, namely,

$$\begin{aligned} \Xi(u) &\asymp \int \mathrm{e}^{M\alpha^{-1}G^{(u)}(\boldsymbol{Q})} \mathrm{e}^{-Mc^{(u)}(\boldsymbol{Q})} \mathrm{d}\boldsymbol{Q} \\ &= \int \exp\left(N \left[\alpha^{-1} G^{(u)}(\boldsymbol{Q}) - c^{(u)}(\boldsymbol{Q}) \right] \right) \mathrm{d}\boldsymbol{Q} \quad (72) \\ &\asymp \exp\left(M \sup_{\boldsymbol{Q},\tilde{\boldsymbol{Q}}} \left\{ T^{(u)}(\boldsymbol{Q}, \tilde{\boldsymbol{Q}}) \right\} \right), \quad (73) \end{aligned}$$

where we denoted for notational convenience

$$T^{(u)}(\boldsymbol{Q}, \tilde{\boldsymbol{Q}}) = \frac{1}{\alpha} G^{(u)}(\boldsymbol{Q}) - \mathrm{tr}(\boldsymbol{Q}\tilde{\boldsymbol{Q}}) + \lim_{M\to\infty} \frac{1}{M} \ln \phi^{(u)}(\tilde{\boldsymbol{Q}}). \quad (74)$$

For complex vectors $\{\boldsymbol{x}_a\}$, the end result is essentially the same and the solution to the supremum is found among the critical points of the argument (see e.g., [29], [31], [39]). The large deviations analysis also guarantees that $\tilde{\boldsymbol{Q}}$ is in general a real symmetric matrix and if $(\boldsymbol{Q}^*, \tilde{\boldsymbol{Q}}^*)$ is the solution of the optimization problem in (73) then $T^{(u)}(\boldsymbol{Q}^*, \tilde{\boldsymbol{Q}}^*) \in \mathbb{R}$, as expected since f is in our case real.

However, in RM there is some ambiguity as to whether the correct point in the saddle-point approximation (73) minimizes or maximizes the exponential when we let $u \to 0$ [24], [25]. Thus, in RM, we seek in practice the critical points and (66) is thus of the form

$$f = -\lim_{u\to 0^+} \frac{\partial}{\partial u} \operatorname*{extr}_{\boldsymbol{Q},\tilde{\boldsymbol{Q}}} \left\{ T^{(u)}(\boldsymbol{Q}, \tilde{\boldsymbol{Q}}) \right\}, \quad (75)$$

where $\mathrm{extr}_X\{h(X)\}$ denotes finding the critical points of a function $h(X)$.

D. Decoupled MGF and Critical Points

The second part of replica analysis where the RS assumption plays an important role (for the first one, see Remark 2) is when we try to solve (71) and find the critical points of $T^{(u)}(\boldsymbol{Q}, \tilde{\boldsymbol{Q}})$. For the simplified setup in this section where $\boldsymbol{Q}$ and $\tilde{\boldsymbol{Q}}$ are represented with parameter $\{Q, q\}$ and $\{\tilde{Q}, \tilde{q}\}$, respectively, the MGF can be expressed as (see, e.g., [27] for details)

$$\phi^{(u)}(\tilde{\boldsymbol{Q}}) = \prod_{m=1}^{M} \left[\left(\frac{\tilde{q}}{\pi} \right)^{-u} \int \left[\mathsf{E}_{x_m} g(z_m | x_m; \tilde{q}^{-1}) \right]^{u+1} \mathrm{d}z_m \right], \quad (76)$$

where z_m are just dummy variables. On the other hand, finding the critical points involves taking eight partial derivatives for the RS case in *Step 3* (for the simplified case here, four is enough). Then, one should pick the solution that satisfies the conditions at the critical point while providing the global extremum of (66).In the case considered here, we can actually get rid of two parameters since $p = M^{-1}\mathsf{E}\|\boldsymbol{x}\|^2$ and $\tilde{p} = 0$ always at the critical point. Note that if we did not parametrize $\boldsymbol{Q}$, the critical points would be described by $u(u+1)$ equations and $\Xi(u)$ would depend explicitly on the fact that u is an integer. This is one of the reasons why even the full-RSB solution (see [24], [25]) uses a round-about way of presenting $\boldsymbol{Q}$ instead of using it "as-is."

Finally, we remark that it is quite common (see, e.g., [27]) to represent the end result in terms of new variables. For example, if we have equal transmit powers for each antennas $\bar{\gamma} = \bar{\gamma}_m$ in the simplified case considered here, then the parameters $\eta = \tilde{q}$ and $\varepsilon = Q - q = \bar{\gamma} - q$ fully describe the RS matrices $\boldsymbol{Q}$ and $\tilde{\boldsymbol{Q}}$ at the critical point. The former variable is inverse noise variance of a decoupled Gaussian channel

$$z = x + w, \qquad p(w) = g(w|0; \eta^{-1}), \quad (77)$$

and the latter variable ε is the MMSE of this channel when the inputs are drawn according to $p_X(x)$. The rest of RM is straightforward, albeit tedious algebra to arrive at (66).

[6] We use notation $a_M \asymp b_M$ to denote "equality up to the leading exponential order," that is $\lim_{M\to\infty} M^{-1} \ln(a_M / b_M) = 0$.

APPENDIX C
REPLICA ANALYSIS FOR MISMATCHED CASE

The analysis herein follows the main steps of RM as listed in Appendix B. Reader who is not familiar with the RM is encouraged to use discussion there as a guide to the derivations below.

A. Replica Trick

Let us consider the function $f(s)$ (free-energy) defined in (13). We then postulate that it can be expressed in the LSL using the standard replica trick (cf. Appendix B)

$$f(s) = -\lim_{M\to\infty}\frac{1}{M}\lim_{u\to 0}\frac{\partial}{\partial u}\ln \Xi^{(u,M)}(s), \tag{78}$$

where we defined for later convenience

$$\Xi^{(u,M)}(s) = \mathsf{E}\left\{\prod_{a=1}^{u} \mathrm{e}^{-[\boldsymbol{w}+\boldsymbol{H}(\boldsymbol{\chi}_0-\boldsymbol{\chi}_a)]^{\mathsf{H}}\boldsymbol{\Sigma}^{-1}[\boldsymbol{w}+\boldsymbol{H}(\boldsymbol{\chi}_0-\boldsymbol{\chi}_a)]}\right\}, \tag{79}$$

and denoted[7] $\boldsymbol{\Sigma} = s^{-1}\tilde{\boldsymbol{R}}$ along with $\boldsymbol{\chi}_0 = \boldsymbol{x}_0 + \boldsymbol{v}_0$ and $\boldsymbol{\chi}_a = \boldsymbol{x}_a, a = 1,\ldots,u$. Here $\boldsymbol{x}_0$ is the original transmit vector in (1) and $\{\boldsymbol{x}_a\}_{a=1}^{u}$ are replicated data vectors, which are IID drawn according to $p(\boldsymbol{x})$ when conditioned on $\{\boldsymbol{x}_0, \boldsymbol{v}_0, \boldsymbol{w}, \boldsymbol{H}\}$. On the other hand, $\boldsymbol{v}_0$ represents the noise plus distortion component at the transmit-sidethat is CSCG with covariance matrix $\boldsymbol{R}_{\boldsymbol{v}}$. Starting with (79), the goal is then to obtain a functional expression for $\Xi^{(u,M)}(s)$ in the LSL that does not enforce u to be an integer and then use (78) to obtain the desired quantity. In the following, explicit limit notations are often omitted for notational convenience.

B. Average Over the Channel and Noise

To proceed with the evaluation of (79), we first make the RS assumption

$$p = M^{-1}\|\boldsymbol{\chi}_0\|^2, \tag{80}$$
$$m = M^{-1}\boldsymbol{\chi}_0^{\mathsf{H}}\boldsymbol{\chi}_a, \qquad a = 1,\ldots,u, \tag{81}$$
$$Q = M^{-1}\|\boldsymbol{\chi}_a\|^2, \qquad a = 1,\ldots,u, \tag{82}$$
$$q = M^{-1}\boldsymbol{\chi}_a^{\mathsf{H}}\boldsymbol{\chi}_b, \qquad a \neq b \in \{1,\ldots,u\}. \tag{83}$$

and remind the reader that if we average first over $\boldsymbol{H}$, the empirical correlations between $\{\boldsymbol{x}_a\}_{a=0}^{u}$ are not zero in general as discussed in Appendix B. Thus, noticing that

$$\mathsf{E}_{\boldsymbol{H}}\left\{[\boldsymbol{H}(\boldsymbol{\chi}_0-\boldsymbol{\chi}_a)]\,[\boldsymbol{H}(\boldsymbol{\chi}_0-\boldsymbol{\chi}_b)]^{\mathsf{H}}\right\} = \begin{cases}[p-(m+m^*)+Q]\,\boldsymbol{I}_N, & a=b,\\ [p-(m+m^*)+q]\,\boldsymbol{I}_N, & a\neq b,\end{cases} \tag{84}$$

we may replace $\{\boldsymbol{H}(\boldsymbol{\chi}_0-\boldsymbol{\chi}_a)\}_{a=1}^{u}$ in (79) in the LSL by CSCG vectors $\{\boldsymbol{\Delta}_a\}_{a=1}^{u}$ that are constructed as

$$\boldsymbol{\Delta}_a = \boldsymbol{d}_a\sqrt{Q-q} + \boldsymbol{t}\sqrt{p-(m+m^*)+q} \tag{85}$$
$$= \boldsymbol{d}_a\sqrt{A} + \boldsymbol{t}\sqrt{B}, \tag{86}$$

where $\{\boldsymbol{t}, \{\boldsymbol{d}_a\}_{a=1}^{u}\}$ are IID standard complex Gaussian RVs independent of $\boldsymbol{w}$. Plugging (86) into $\Xi^{(u,M)}(s)$ and recalling that $\boldsymbol{\Sigma}$ is a fixed predefined matrix gives

$$\Xi^{(u,M)}(s) = \frac{1}{\det(\boldsymbol{R}_{\boldsymbol{w}})}\mathsf{E}\int\frac{\mathrm{d}\boldsymbol{w}}{\pi^N}\mathrm{e}^{-\boldsymbol{w}^{\mathsf{H}}\left(\boldsymbol{R}_{\boldsymbol{w}}^{-1}+u\boldsymbol{\Sigma}^{-1}\right)\boldsymbol{w}} \times\int\frac{\mathrm{d}\boldsymbol{t}}{\pi^N}\mathrm{e}^{-\boldsymbol{t}^{\mathsf{H}}(\boldsymbol{I}+uB\boldsymbol{\Sigma}^{-1})\boldsymbol{t}-2\Re\left\{\boldsymbol{w}^{\mathsf{H}}(u\sqrt{B}\boldsymbol{\Sigma}^{-1})\boldsymbol{t}\right\}} \times\left[\int\mathrm{e}^{-\boldsymbol{d}^{\mathsf{H}}(\boldsymbol{I}+A\boldsymbol{\Sigma}^{-1})\boldsymbol{d}+2\Re\left\{\left[-\sqrt{A}\boldsymbol{\Sigma}^{-1}(\boldsymbol{w}+\sqrt{B}\boldsymbol{t})\right]^{\mathsf{H}}\boldsymbol{d}\right\}}\frac{\mathrm{d}\boldsymbol{d}}{\pi^N}\right]^{u}. \tag{87}$$

Next, Gaussian integration (62) is applied on the integral w.r.t. $\boldsymbol{d}$. Using also (63) we arrive at

$$\Xi^{(u,M)}(s) = \mathsf{E}\int\frac{\mathrm{d}\boldsymbol{w}}{\pi^N}\frac{\mathrm{e}^{-\boldsymbol{w}^{\mathsf{H}}\left(\boldsymbol{R}_{\boldsymbol{w}}^{-1}+u(A\boldsymbol{I}_N+\boldsymbol{\Sigma})^{-1}\right)\boldsymbol{w}}}{\left[\det(\boldsymbol{I}+A\boldsymbol{\Sigma}^{-1})\right]^{u}\det(\boldsymbol{R}_{\boldsymbol{w}})} \times\int\mathrm{e}^{-\boldsymbol{t}^{\mathsf{H}}\left[\boldsymbol{I}_N+uB(A\boldsymbol{I}_N+\boldsymbol{\Sigma})^{-1}\right]\boldsymbol{t}+2\Re\left\{\left[-u\sqrt{B}(A\boldsymbol{I}_N+\boldsymbol{\Sigma})^{-1}\boldsymbol{w}\right]^{\mathsf{H}}\boldsymbol{t}\right\}}\frac{\mathrm{d}\boldsymbol{t}}{\pi^N}. \tag{88}$$

Application of (62) and (63) again for the integral w.r.t. $\boldsymbol{t}$ provides

$$\begin{aligned}\Xi^{(u,M)}(s) &= \mathsf{E}\left\{\frac{\left[\det(\boldsymbol{I}+A\boldsymbol{\Sigma}^{-1})\right]^{-u}}{\det\left[\boldsymbol{I}_N+uB(A\boldsymbol{I}_N+\boldsymbol{\Sigma})^{-1}\right]\det(\boldsymbol{R}_{\boldsymbol{w}})} \times\int\mathrm{e}^{-\boldsymbol{w}^{\mathsf{H}}\left(\boldsymbol{R}_{\boldsymbol{w}}^{-1}+u[(A+uB)\boldsymbol{I}_N+\boldsymbol{\Sigma}]^{-1}\right)\boldsymbol{w}}\frac{\mathrm{d}\boldsymbol{w}}{\pi^N}\right\}\\ &= \mathsf{E}\left\{\frac{\left(\det\left[\boldsymbol{I}_N+u\boldsymbol{R}_{\boldsymbol{w}}\left((A+uB)\boldsymbol{I}_N+\boldsymbol{\Sigma}\right)^{-1}\right]\right)^{-1}}{\det\left[\boldsymbol{I}_N+uB(A\boldsymbol{I}_N+\boldsymbol{\Sigma})^{-1}\right]\left[\det(\boldsymbol{I}+A\boldsymbol{\Sigma}^{-1})\right]^{u}}\right\},\end{aligned} \tag{89}$$

where the second line is also obtained through Gaussian integration. The above holds for any $\boldsymbol{R}_{\boldsymbol{w}}$ and $\boldsymbol{\Sigma}$ that are Hermitian and invertible. The determinants in (89) can be further simplified using (64), so that recalling $\boldsymbol{\Sigma} = s^{-1}\tilde{\boldsymbol{R}}$ and defining two auxiliary matrices

$$\boldsymbol{\Omega}(p,m,q) = \boldsymbol{R}_{\boldsymbol{w}} + (p-(m+m^*)+q)\,\boldsymbol{I}_N, \tag{90}$$
$$\tilde{\boldsymbol{\Omega}}(Q,q) = s^{-1}\tilde{\boldsymbol{R}} + (Q-q)\boldsymbol{I}_N, \tag{91}$$

that are both Hermitian, we finally have

$$\Xi^{(u,M)}(s) = \det(s^{-1}\tilde{\boldsymbol{R}})^{u}\mathsf{E}\left\{\mathrm{e}^{G^{(u)}(p,m,q,Q)}\right\}, \tag{92}$$

$$G^{(u)}(p,m,q,Q) = (1-u)\ln\det\tilde{\boldsymbol{\Omega}}(Q,q) - \ln\det\left[\tilde{\boldsymbol{\Omega}}(Q,q)+u\boldsymbol{\Omega}(p,m,q)\right], \tag{93}$$

[7] We remind the reader that for the case of mismatched decoding, the postulated covariance matrix $\tilde{\boldsymbol{R}}$ is fixed by definition so that $\boldsymbol{\Sigma} = s^{-1}\tilde{\boldsymbol{R}}$ is also a fixed predefined matrix. This is in contrast to the case of matched decoding (5), where the effective covariance matrix $\boldsymbol{R}_{\boldsymbol{w}} + \boldsymbol{H}\boldsymbol{R}_{\boldsymbol{v}}\boldsymbol{H}$ is random and depends directly on the channel matrix $\boldsymbol{H}$.

Using the differentiation rule $\frac{\partial}{\partial x}\ln\det\boldsymbol{A}=\mathrm{tr}(\boldsymbol{A}^{-1}\frac{\partial \boldsymbol{A}}{\partial x})$, where the partial derivative should be understood as an elementwise operation on $\boldsymbol{A}$, we also obtain for later use the equalities

$$\frac{\partial}{\partial p}G^{(u)}(\boldsymbol{Q}) = -u\mathrm{tr}\left((\tilde{\boldsymbol{\Omega}}+u\boldsymbol{\Omega})^{-1}\right), \tag{94}$$

$$\frac{\partial}{\partial m}G^{(u)}(\boldsymbol{Q}) = \frac{\partial}{\partial m^*}G^{(u)}(\boldsymbol{Q}) = u\mathrm{tr}\left((\tilde{\boldsymbol{\Omega}}+u\boldsymbol{\Omega})^{-1}\right), \tag{95}$$

$$\frac{\partial}{\partial q}G^{(u)}(\boldsymbol{Q}) = u(u-1)\mathrm{tr}\left(\tilde{\boldsymbol{\Omega}}^{-1}\boldsymbol{\Omega}(\tilde{\boldsymbol{\Omega}}+u\boldsymbol{\Omega})^{-1}\right), \tag{96}$$

$$\frac{\partial}{\partial Q}G^{(u)}(\boldsymbol{Q}) = u\mathrm{tr}\left(\tilde{\boldsymbol{\Omega}}^{-1}\boldsymbol{\Omega}(\tilde{\boldsymbol{\Omega}}+u\boldsymbol{\Omega})^{-1}\right) - u\mathrm{tr}(\tilde{\boldsymbol{\Omega}}^{-1}), \tag{97}$$

where the dependencies to $\{p,m,q,Q\}$ were omitted on the RHSs of the equations for notational implicity.

C. Distribution of the Overlap Matrix and Large Deviations

Let us now write the general form of empirical correlations between $\{\boldsymbol{\Delta}_a\}$ as

$$\begin{aligned}\frac{1}{M}\mathsf{E}_{\boldsymbol{H}}\left\{\boldsymbol{\Delta}_b^{\mathsf{H}}\boldsymbol{\Delta}_a\right\} &= \left(\frac{\|\boldsymbol{\chi}_0\|^2}{M}-\frac{\boldsymbol{\chi}_b^{\mathsf{H}}\boldsymbol{\chi}_0}{M}-\frac{\boldsymbol{\chi}_0^{\mathsf{H}}\boldsymbol{\chi}_a}{M}+\frac{\boldsymbol{\chi}_b^{\mathsf{H}}\boldsymbol{\chi}_a}{M}\right)\\ &= (Q_{0,0}-Q_{0,b}-Q_{a,0}+Q_{a,b}),\end{aligned} \tag{98}$$

where $Q_{a,b}$ are the elements of the overlap matrix $\boldsymbol{Q}\in\mathbb{C}^{(u+1)\times(u+1)}$ and have the obvious definitions. We then need to find a suitable formula for the rate function (70). By the RS assumption,

$$\mathrm{tr}(\boldsymbol{Q}\tilde{\boldsymbol{Q}}) = p\tilde{p}+u\tilde{m}(m+m^*)+uQ\tilde{Q}+u(u-1)q\tilde{q}, \tag{99}$$

since $\tilde{\boldsymbol{Q}}$ is real symmetric and we may write (78) as in (100), shown at the bottom of the page, where the per-antenna rate function reads

$$\phi_m^{(u)}(\tilde{\boldsymbol{Q}}) = \mathsf{E}_{\{\chi_{a,m}\}}\left\{\exp\left[\sum_{a=0}^{u}\sum_{b=0}^{u}\tilde{Q}_{a,b}\chi_{b,m}^*\chi_{a,m}\right]\right\}, \tag{101}$$

and $\chi_a=[\chi_{a,1}\cdots\chi_{a,M}]^\top$.

D. Decoupled MGF and Critical Points

The first set of equations for the critical point arises from the equality

$$\frac{\partial}{\partial x}\mathrm{tr}(\boldsymbol{Q}\tilde{\boldsymbol{Q}}) = \frac{1}{M}\frac{\partial}{\partial x}G^{(u)}(\boldsymbol{Q}), \tag{102}$$

for $x\in\{p,m,q,Q\}$. The partial derivatives on the LHS are trivial due to (99) and the RHSs we already obtained in (94)–(97). If we drop the explicit dependence of $\boldsymbol{\Omega}$ and $\tilde{\boldsymbol{\Omega}}$ on $\{p,m,q,Q\}$ for notational simplicity, the RS conjugate parameters satisfy

$$\tilde{p} = -u\frac{1}{M}\mathrm{tr}\left[(\tilde{\boldsymbol{\Omega}}+u\boldsymbol{\Omega})^{-1}\right] = -u\tilde{m}, \tag{103}$$

$$\tilde{m} = \frac{1}{M}\mathrm{tr}\left[(\tilde{\boldsymbol{\Omega}}+u\boldsymbol{\Omega})^{-1}\right], \tag{104}$$

$$\tilde{q} = \frac{1}{M}\mathrm{tr}\left[\tilde{\boldsymbol{\Omega}}^{-1}\boldsymbol{\Omega}(\tilde{\boldsymbol{\Omega}}+u\boldsymbol{\Omega})^{-1}\right], \tag{105}$$

$$\tilde{Q} = \frac{1}{M}\mathrm{tr}\left[\tilde{\boldsymbol{\Omega}}^{-1}\boldsymbol{\Omega}(\tilde{\boldsymbol{\Omega}}+u\boldsymbol{\Omega})^{-1}\right] - \frac{1}{M}\mathrm{tr}\left(\tilde{\boldsymbol{\Omega}}^{-1}\right). \tag{106}$$

Note that the above implies that in the limit $u\to 0$, we have $\tilde{p}\to 0$, and $\tilde{m}\to -(\tilde{Q}-\tilde{q})$, so that the relevant critical point can be written by using two instead of four "tilde-parameters."

The next task is to obtain an explicit expression for the per-component moment generating function (MGF) in (101) that does not require u to be an integer. Since this part is closely similar to the analysis carried out, e.g., in [27] we omit the details of the derivations. Following the notation of [27], we let $\xi=\tilde{m}$ and $\eta=\tilde{m}^2/\tilde{q}$ which is sufficient to describe $\tilde{\boldsymbol{Q}}$ here. Then, if we denote $\chi_m=x_m+v_m$ and $\tilde{\chi}_m=\tilde{x}_m$, the scalar MGF (101) can be written as

$$\begin{aligned}\phi_m^{(u)}(\tilde{\boldsymbol{Q}}) = \left(\frac{\pi}{\xi}\right)^u \mathsf{E}\Bigg\{\int \mathrm{d}z_m \mathrm{e}^{u\xi\left(|z_m|^2-|\chi_m|^2\right)}\\ \times p(z_m|\chi_m)\left[\mathsf{E}_{\tilde{\chi}_m}q(z_m|\tilde{\chi}_m)\right]^u\Bigg\},\end{aligned} \tag{107}$$

where $p(z_m|\chi_m)=g(z_m|\chi_m;,\eta^{-1})$ and $q(z_m|\tilde{\chi}_m)=g(z_m|\tilde{\chi}_m;\xi^{-1})$. As a consequence of the above, u does not need to be an integer anymore and the limit $u\to 0$ is well defined. From the partial derivatives of $\{\tilde{p},\tilde{m},\tilde{q},\tilde{Q}\}$ we obtain the second set of conditions at the critical point

$$p = \lim_{M\to\infty}\frac{1}{M}\sum_{m=1}^{M}\mathsf{E}|x_m+v_m|^2, \tag{108}$$

$$Q = \lim_{M\to\infty}\frac{1}{M}\sum_{m=1}^{M}\mathsf{E}\left\langle|\tilde{x}_m|^2\right\rangle_q, \tag{109}$$

$$m = \lim_{M\to\infty}\frac{1}{M}\sum_{m=1}^{M}\mathsf{E}(x_m+v_m)\left\langle\tilde{x}_m^*\right\rangle_q, \tag{110}$$

$$q = \lim_{M\to\infty}\frac{1}{M}\sum_{m=1}^{M}\mathsf{E}\left\langle\tilde{x}_m^*\right\rangle_q\left\langle\tilde{x}_m^*\right\rangle_q, \tag{111}$$

where $x_m,\tilde{x}_m\sim p(x_m)$, $v_m\sim g(v_m|0;r_{\boldsymbol{v}}^m)$,

$$\left\langle f(\tilde{x}_m)\right\rangle_q = \mathsf{E}_{\tilde{x}_m}f(\tilde{x}_m)\frac{q(z_m|\tilde{x}_m)}{q(z_m)}, \tag{112}$$

$$\begin{aligned}f_{\mathsf{RS}} = -\lim_{M\to\infty}\frac{1}{M}\ln\det(\boldsymbol{\Sigma}) - \underset{\boldsymbol{Q},\tilde{\boldsymbol{Q}}}{\mathrm{extr}}\lim_{M\to\infty}\Bigg\{\frac{1}{M}\lim_{u\to 0}\frac{\partial}{\partial u}G^{(u)}(\boldsymbol{Q})\\ -\lim_{u\to 0}\frac{\partial}{\partial u}\left[p\tilde{p}+u(m\tilde{m}^*+\tilde{m}m^*)+uQ\tilde{Q}+u(u-1)q\tilde{q}\right]+\frac{1}{M}\sum_{m=1}^{M}\lim_{u\to 0}\frac{\partial}{\partial u}\ln\phi_m^{(u)}(\tilde{\boldsymbol{Q}})\Bigg\}\end{aligned} \tag{100}$$

and $q(z_m) = \mathsf{E}_{\tilde{x}_m} q(z_m|\tilde{x}_m)$. The interpretation is that (112) represents the conditional mean estimator for postulated channel $q(z_m|\tilde{\chi}_m)$ when the true channel is given by $p(z_m|\chi_m)$. Then the true $\varepsilon = p - (m + m^*) + q$, and postulated $\tilde{\varepsilon} = Q - q$ MMSE reduce to (23) and (24), respectively. Finally, computing the partial derivatives w.r.t. u in (100) and taking the limit $u \to 0$ provides after some algebra the free energy (26).

REFERENCES

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

[2] T. C. W. Schenk, P. F. M. Smulders, and E. R. Fledderus, "Performance of MIMO OFDM systems in fading channels with additive TX and RX impairments," in *Proc. 1st Annu. IEEE BENELUX/DSP Valley Signal Process. Symp.*, Apr. 2005, pp. 41–44.

[3] B. Göransson, S. Grant, E. Larsson, and Z. Feng, "Effect of transmitter and receiver impairments on the performance of MIMO in HSDPA," in *Proc. 9th IEEE Workshop Signal Process. Adv. Wireless Commun.*, Jul. 2008, pp. 496–500.

[4] H. Suzuki, T. V. Anh Tran, I. B. Collings, G. Daniels, and M. Hedley, "Transmitter noise effect on the performance of a MIMO-OFDM hardware implementation achieving improved coverage," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 6, pp. 867–876, Aug. 2008.

[5] H. Suzuki, I. B. Collings, M. Hedley, and G. Daniels, "Practical performance of MIMO-OFDM-LDPC with low complexity double iterative receiver," in *Proc. 20th IEEE Int. Symp. Pers., Indoor, Mobile Radio Commun.*, Sep. 2009, pp. 2469–2473.

[6] C. Studer, M. Wenk, and A. Burg, "MIMO transmission with residual transmit-RF impairments," in *Proc. Int. ITG Workshop Smart Antennas*, Feb. 2010, pp. 189–196.

[7] J. González-Coma, P. M. Castro, and L. Castedo, "Impact of transmit impairments on multiuser MIMO non-linear transceivers," in *Proc. Int. ITG Workshop Smart Antennas*, Feb. 2011, pp. 1–8.

[8] C. Studer, M. Wenk, and A. Burg, "System-level implications of residual transmit-RF impairments in MIMO systems," in *Proc. 5th Eur. Conf. Antennas Propag.*, Apr. 2011, pp. 2686–2689.

[9] J. González-Coma, P. M. Castro, and L. Castedo, "Transmit impairments influence on the performance of MIMO receivers and precoders," in *Proc. 11th Eur. Wireless Conf.*, Apr. 2011, pp. 1–8.

[10] E. Björnson, P. Zetterberg, M. Bengtsson, and B. Ottersten, "Capacity limits and multiplexing gains of MIMO channels with transceiver impairments," *IEEE Commun. Lett.*, vol. 17, no. 1, pp. 91–94, Jan. 2013.

[11] X. Zhang, M. Matthaiou, E. Björnson, M. Coldrey, and M. Debbah, "On the MIMO capacity with residual transceiver hardware impairments," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2014, pp. 5299–5305.

[12] G. Fettweis *et al.*, "Dirty RF: A new paradigm," *Int. J. Wireless Inf. Netw.*, vol. 14, no. 2, pp. 133–148, Jun. 2007.

[13] T. Schenk, *RF Imperfections in High-Rate Wireless Systems: Impact and Digital Compensation*. Berlin, Germany: Springer-Verlag, 2008.

[14] F. Gregorio, J. Cousseau, S. Werner, T. Riihonen, and R. Wichman, "EVM analysis for broadband OFDM direct-conversion transmitters," *IEEE Trans. Veh. Technol.*, vol. 62, no. 7, pp. 3443–3451, Sep. 2013.

[15] N. Merhav, G. Kaplan, A. Lapidoth, and S. Shamai, "On information rates for mismatched decoders," *IEEE Trans. Inf. Theory*, vol. 40, no. 6, pp. 1953–1967, Nov. 1994.

[16] A. Ganti, A. Lapidoth, and I. Telatar, "Mismatched decoding revisited: General alphabets, channels with memory, the wide-band limit," *IEEE Trans. Inf. Theory*, vol. 40, no. 6, pp. 1953–1967, Nov. 2000.

[17] A. Guillén i Fàbregas, A. Martinez, and G. Caire, "Bit-interleaved coded modulation," *Found. Trends Commun. Inf. Theory*, vol. 5, no. 1/2, pp. 1–153, Jan. 2008.

[18] W. Zhang, "A general framework for transmission with transceiver distortion and some applications," *IEEE Trans. Commun.*, vol. 60, no. 2, pp. 384–399, Feb. 2012.

[19] 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.

[20] H. Weingarten, Y. Steinberg, and S. Shamai, "Gaussian codes and weighted nearest neighbor decoding in fading multiple-antenna channels," *IEEE Trans. Inf. Theory*, vol. 50, no. 8, pp. 1665–1686, Aug. 2004.

[21] A. T. Asyhari and A. Guillén i Fàbregas, "Nearest neighbor decoding in MIMO block-fading channels with imperfect CSIR," *IEEE Trans. Inf. Theory*, vol. 58, no. 3, pp. 1483–1517, Mar. 2012.

[22] M. Médard, "The effect upon channel capacity in wireless communications of perfect and imperfect knowledge of the channel," *IEEE Trans. Inf. Theory*, vol. 46, no. 3, pp. 933–946, May 2000.

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

[24] H. Nishimori, *Statistical Physics of Spin Glasses and Information Processing*. New York, NY, USA: Oxford Univ. Press, 2001.

[25] M. Mézard and A. Montanari, *Information, Physics, Computation*. New York, NY, USA: Oxford Univ. Press, 2009.

[26] T. Tanaka, "A statistical-mechanics approach to large-system analysis of CDMA multiuser detectors," *IEEE Trans. Inf. Theory*, vol. 48, no. 11, pp. 2888–2910, Nov. 2002.

[27] D. Guo and S. Verdú, "Randomly spread CDMA: Asymptotics via statistical physics," *IEEE Trans. Inf. Theory*, vol. 51, no. 6, pp. 1983–2010, Jun. 2005.

[28] R. R. Müller, "Channel capacity and minimum probability of error in large dual antenna array systems with binary modulation," *IEEE Trans. Signal Process.*, vol. 51, no. 11, pp. 2821–2828, Nov. 2003.

[29] C.-K. Wen, K.-K. Wong, and J.-C. Chen, "Spatially correlated MIMO multiple-access systems with macrodiversity: Asymptotic analysis via statistical physics," *IEEE Trans. Commun.*, vol. 55, no. 3, pp. 477–488, Mar. 2007.

[30] C.-K. Wen, K.-K. Wong, and J.-C. Chen, "Asymptotic mutual information for Rician MIMO-MA channels with arbitrary inputs: A replica analysis," *IEEE Trans. Commun.*, vol. 58, no. 10, pp. 2782–2788, Oct. 2010.

[31] K. Takeuchi, R. R. Müller, M. Vehkaperä, and T. Tanaka, "On an achievable rate of large Rayleigh block-fading MIMO channels with no CSI," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6517–6541, Oct. 2013.

[32] A. M. Tulino and S. Verdú, "Random matrix theory and wireless communications," *Found. Trends Commun. Inf. Theory*, vol. 1, no. 1, pp. 1–182, Jan. 2004.

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

[34] C.-K. Wen and K.-K. Wong, "On the sum-rate of uplink MIMO cellular systems with amplify-and-forward relaying and collaborative base stations," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 9, pp. 1409–1424, Dec. 2010.

[35] M. A. Girnyk, M. Vehkaperä, and L. K. Rasmussen, "Large-system analysis of correlated MIMO multiple access channels with arbitrary signaling in the presence of interference," *IEEE Trans. Wireless Commun.*, vol. 13, no. 4, pp. 2060–2073, Apr. 2014.

[36] M. A. Girnyk, M. Vehkaperä, and L. K. Rasmussen, "Asymptotic performance analysis of a K-hop amplify-and-forward relay MIMO channel," arXiv:1410.5716.

[37] A. L. Moustakas and S. H. Simon, "On the outage capacity of correlated multiple-path MIMO channels," *IEEE Trans. Inf. Theory*, vol. 53, no. 11, pp. 3887–3903, Nov. 2007.

[38] D. S. Bernstein, *Matrix Mathematics: Theory, Facts, and Formulas*. Princeton, NJ, USA: Princeton Univ. Press, 2009.

[39] R. R. Müller, D. Guo, and A. Moustakas, "Vector precoding for wireless MIMO systems and its replica analysis," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 3, pp. 530–540, Apr. 2008.

[40] M. Vehkaperä, Y. Kabashima, and S. Chatterjee, "Analysis of regularized LS reconstruction and random matrix ensembles in compressed sensing," arXiv:1312.0256.

[41] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*. New York, NY, USA: Springer-Verlag, 1998.

Mikko Vehkaperä (S'03–M'10) received the M.Sc. and Lic.Sc. degrees from University of Oulu, Oulu, Finland, in 2004 and 2006, respectively, and the Ph.D. degree from Norwegian University of Science and Technology (NTNU), Trondheim, Norway, in 2010.

Between 2010 and 2013, he was a Postdoctoral Researcher at School of Electrical Engineering, and the ACCESS Linnaeus Center, KTH Royal Institute of Technology, Stockholm, Sweden. Since 2013, he has been an Academy of Finland Postdoctoral Researcher at Aalto University School of Electrical Engineering. He held visiting appointments at Massachusetts Institute of Technology (MIT), USA, Kyoto University and Tokyo Institute of Technology, Japan, and University of Erlangen-Nuremberg, Germany. His research interests are in the field of wireless communications and signal processing.

Dr. Vehkaperä was a co-recipient for the Best Student Paper award at IEEE International Conference on Networks (ICON2011) for the paper "Delay Constrained Throughput Analysis of CDMA Using Stochastic Network Calculus."

Taneli Riihonen (S'06–M'14) received the M.Sc. and D.Sc. degrees in communications and electrical engineering (both with distinction) from Helsinki University of Technology, Espoo, Finland, in February 2006 and from Aalto University, Helsinki, Finland in August 2014, respectively. During summer 2005, he was a Seasonal Trainee at Nokia Research Center, Helsinki, Finland and, since Fall 2005, he has held various research positions at the Department of Signal Processing and Acoustics, Aalto University School of Electrical Engineering, Helsinki, Finland. He is currently a Visiting Associate Research Scientist at Columbia University in the City of New York, USA. He is serving as an Editor of IEEE COMMUNICATIONS LETTERS since October 2014. His research activity is focused on physical-layer OFDM(A), multiantenna, relaying and full-duplex wireless techniques.

Maksym A. Girnyk (S'09) received the B.Sc. and M.Sc. degrees from the KPI National Technical University of Ukraine, Kyiv, Ukraine in 2005 and 2007, respectively, the M.Sc. degree from Supélec, Paris, France in 2008, as well as the Lic.Eng. and Ph.D. degrees from the KTH Royal Institute of Technology, Stockholm, Sweden in 2012 and 2014, respectively. He was a Visiting Researcher at the Alcatel-Lucent Chair on Flexible Radio at Supélec, Paris, France in 2013 and Aalto University, Aalto, Finland in 2014. His research interests include wireless communications, information theory, random matrix theory, and statistical physics.

Emil Björnson (S'07–M'12) received the M.S. degree in engineering mathematics from Lund University, Lund, Sweden, in 2007. He received the Ph.D. degree in telecommunications from the Department of Signal Processing at KTH Royal Institute of Technology, Stockholm, Sweden, in 2011. From 2012 to 2014, he was a joint postdoc at the Alcatel-Lucent Chair on Flexible Radio, Supélec, Paris, France, and at KTH Royal Institute of Technology. He is currently an Assistant Professor at the Division of Communication Systems at Linköping University, Sweden.

His research interests include multi-antenna cellular communications, massive MIMO techniques, radio resource allocation, green energy efficient systems, and network topology design. He is the first author of the book "Optimal Resource Allocation in Coordinated Multi-Cell Systems" from 2013. He is also dedicated to reproducible research and has made a large amount of simulation code publically available.

Dr. Björnson has received 4 best paper awards (as first author or coauthor) for novel research on optmization and design of multi-cell multi-antenna communications: WCNC 2014, SAM 2014, CAMSAP 2011, and WCSP 2009.

Mérouane Debbah (S'01–AM'03–M'04–SM'08–F'15) received the M.Sc. and Ph.D. degrees from the Ecole Normale Supérieure de Cachan, France. He worked for Motorola Labs (Saclay, France) from 1999 to 2002, and the Vienna Research Center for Telecommunications (Vienna, Austria) until 2003. From 2003 to 2007, he joined the Mobile Communications Department of the Institut Eurecom (Sophia Antipolis, France) as an Assistant Professor. Since 2007, he is a Full Professor at Supelec (Gif-sur-Yvette, France). From 2007 to 2014, he was Director of the Alcatel-Lucent Chair on Flexible Radio. Since 2014, he is Vice-President of the Huawei France R&D Center and Director of the Mathematical and Algorithmic Sciences Lab. His research interests are in information theory, signal processing, and wireless communications. He is an Associate Editor in Chief of the journal, *Random Matrix: Theory and Applications* and was an associate and senior area editor for IEEE TRANSACTIONS ON SIGNAL PROCESSING, respectively in 2011–2013 and 2013–2014. Dr. Debbah is a recipient of the ERC grant MORE (Advanced Mathematical Tools for Complex Network Engineering). He is a WWRF Fellow, and a member of the academic senate of Paris-Saclay. He is the recipient of the Mario Boella Award in 2005, the 2007 IEEE GLOBECOM Best Paper Award, the Wi-Opt 2009 Best Paper Award, the 2010 Newcom++ Best Paper, the WUN CogCom Best Paper 2012 and 2013 Award, the 2014 WCNC Best Paper Award as well as the Valuetools 2007, Valuetools 2008, CrownCom2009, Valuetools 2012 and SAM 2014 best student paper awards. In 2011, he received the IEEE Glavieux Prize Award and in 2012, the Qualcomm Innovation Prize Award.

Lars Kildehøj Rasmussen (S'92–M'93–SM'01) received the M.Eng. degree in 1989 from the Technical University of Denmark, Lyngby, Denmark, and the Ph.D. degree from Georgia Institute of Technology, Atlanta, Georgia, USA, in 1993. He is now a Professor in the Department of Communication Theory at the School of Electrical Engineering, and the ACCESS Linnaeus Center, at the KTH Royal Institute of Technology, Stockholm, Sweden.

He has prior experience from the Institute for Telecommunications Research at the University of South Australia (Adelaide, Australia), the Center for Wireless Communications at National University of Singapore (Singapore), Chalmers University of Technology (Gothenburg, Sweden), and University of Pretoria (Pretoria, South Africa). He is a Co-Founder of Cohda Wireless Pty., Ltd. (http://www.cohdawireless.com/); a Leading Developer of Safe Vehicle and Connected Vehicle design solutions. His research interests include transmission strategies and coding schemes for wireless communications, communications and control, and vehicular communication systems.

He is a member of the IEEE Information Theory Society, Communications Society, and Vehicular Technology Society. He served as Chairman for the Australian Chapter of the IEEE Information Theory Society 2004–2005, and has been a board member of the IEEE Sweden Section Vehicular Technology, Communications, and Information Theory Joint Societies Chapter since 2010. He is an Associate Editor for IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, and *Physical Communications*, as well as a Former Associate Editor of IEEE TRANSACTIONS ON COMMUNICATIONS, 2002–2013.

Risto Wichman received the M.Sc. and D.Sc. (Tech.) degrees in digital signal processing from Tampere University of Technology, Tampere, Finland, in 1990 and 1995, respectively. From 1995 to 2001, he worked at Nokia Research Center as a Senior Research Engineer. In 2002, he joined Department of Signal Processing and Acoustics, Aalto University School of Electrical Engineering, Aalto, Finland, where he is a Full Professor since 2008. His research interests include signal processing techniques for wireless communication systems.

Asymptotic Analysis of SU-MIMO Channels With Transmitter Noise and Mismatched Joint Decoding

Mikko Vehkaperä, *Member, IEEE*, Taneli Riihonen, *Member, IEEE*, Maksym A. Girnyk, *Student Member, IEEE*, Emil Björnson, *Member, IEEE*, Mérouane Debbah, *Fellow, IEEE*, Lars Kildehøj Rasmussen, *Senior Member, IEEE*, and Risto Wichman

***Abstract*—Hardware impairments in radio-frequency components of a wireless system cause unavoidable distortions to transmission that are not captured by the conventional linear channel model. In this paper, a "binoisy" single-user multiple-input multiple-output (SU-MIMO) relation is considered where the additional distortions are modeled via an additive noise term at the transmit side. Through this extended SU-MIMO channel model, the effects of transceiver hardware impairments on the achievable rate of multi-antenna point-to-point systems are studied. Channel input distributions encompassing practical discrete modulation schemes, such as, QAM and PSK, as well as Gaussian signaling are covered. In addition, the impact of mismatched detection and decoding when the receiver has insufficient information about the non-idealities is investigated. The numerical results show that for realistic system parameters, the effects of transmit-side noise and mismatched decoding become significant only at high modulation orders.**

***Index Terms*—Multiple antennas, mismatched decoding, ergodic capacity, fading channels, generalized mutual information (GMI), transceiver hardware impairments.**

I. Introduction

MIMO, i.e., multiple-input multiple-output, wireless links are a mature research subject and their theory is already well understood [1]. However, the extensive body of literature on link-level analysis conventionally concerns signal models of the form $\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{n}$ reckoning with an additive thermal-noise term, namely $\boldsymbol{n}$, only at the receiver after the fading channel $\boldsymbol{H}$. In this paper, we investigate single-user MIMO channels and adopt a generalized ("binoisy") input–output relation from [2]–[11]:

$$\boldsymbol{y} = \boldsymbol{H}(\boldsymbol{x} + \boldsymbol{v}) + \boldsymbol{w}, \tag{1}$$

where $\boldsymbol{w}$ is an additive receive-side distortion-plus-noise component. The system model (1) allows including an additive noise term, namely $\boldsymbol{v}$, also at the transmitter, thus making the total effective noise term $\boldsymbol{H}\boldsymbol{v} + \boldsymbol{w}$ colored and correlated with the fading channel. This small but significant complement yields a MIMO link model whose performance analysis is still an open research niche in many respects.

Although we primarily aim at extending the capacity theory of binoisy SU-MIMO channels under fading without committing to any particular application, the signal model (1) originally stems from the practical need for modeling the combined effect of various transceiver hardware impairments which are detailed in [12], [13], and the references therein. However, it is worth acknowledging that the additive noise assumed herein is only a simplified representation of complex nonlinear phenomena occurring due to hardware impairments, especially when considering their joint coupled effects or trying to model residual distortion after compensation. Thus, the binoisy signal model should be regarded as a compromise between facilitating theoretical analysis and resorting to measurements or simulations under more accurate modeling. Yet the central limit theorem further justifies the model by averaging the combined effects of different impairments to additive Gaussian noise when the signal model (1) is understood to represent a single narrowband subcarrier within a wideband system.

Additive receiver hardware impairments can be incorporated into the conventional signal model by increasing the level of the thermal-noise term $\boldsymbol{n}$ by a constant noise figure, e.g., about 3–5 dB, or by scaling it in proportion to the input signal level such that it matches with $\boldsymbol{w}$. On the other hand, regarding the joint effect of transmitter hardware impairments as an additive transmit-side noise term $\boldsymbol{v}$ is analogous to the principles of practical radio conformance testing. In particular, the common transmitter quality indicator is error-vector magnitude (EVM) which reduces the distortion effects to an additive component and measures its level relatively to signal amplitude [14].

Typical target EVM values guarantee that the signal $\boldsymbol{x}$ is at least 20–30 dB above the transmit-side noise $\boldsymbol{v}$. On the other hand, for basic discrete channel inputs such as quadrature phase-shift keying (QPSK), $\boldsymbol{H}\boldsymbol{x}$ is usually at most 10–15 dB above the receive-side noise $\boldsymbol{w}$, after which the communication is not anymore limited by noise but the lack of entropy in the

Manuscript received June 9, 2014; revised October 22, 2014; accepted December 10, 2014. This research was supported in part by the Academy of Finland, the Swedish Research Council and the ERC Starting Grant 305123 MORE. The associate editor coordinating the review of this paper and approving it for publication was V. Raghavan.

M. Vehkaperä, T. Riihonen, and R. Wichman are with the Department of Signal Processing and Acoustics, Aalto University School of Electrical Engineering, 00076 Aalto, Finland (e-mail: mikko.vehkapera@aalto.fi; taneli.riihonen@aalto.fi; risto.wichman@aalto.fi).

M. A. Girnyk and L. K. Rasmussen are with the Department of Communication Theory, School of Electrical Engineering, and the ACCESS Linnaeus Center, KTH Royal Institute of Technology, 100 44 Stockholm, Sweden (e-mail: mgyr@kth.se; lkra@kth.se).

E. Björnson is with Department of Electrical Engineering (ISY), Linköping University, 581 83 Linköping, Sweden (e-mail: emil.bjornson@liu.se).

M. Debbah is with Huawei France R&D Center, 92659 Paris, France, and also with Supelec, 91192 Gif-sur-Yvette, France (e-mail: merouane.debbah@huawei.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCOMM.2014.2385051

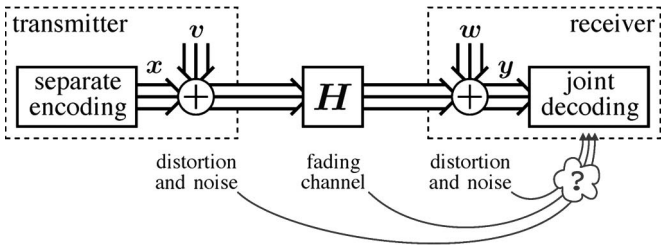

Fig. 1. System model for non-ideal MIMO communications with transmit and receive distortion. The receiver might be misinformed or ignorant of some of the variables in the transmission chain leading to mismatched decoding.

modulation alphabet. This implies that transmitter hardware impairments can be justifiably omitted in the analysis of simple low-rate wireless systems: Either $\boldsymbol{H}\boldsymbol{v}$ is well below the receive-side noise $\boldsymbol{w}$ (say 5–20 dB) or the signal-to-noise ratio (SNR) is set to an uninterestingly high level. However, there has been a trend to improve data rates by using, e.g., quadrature amplitude modulation (QAM) up to 64-QAM at relatively high SNR, in which case the transmit-side noise begins to play a notable role in the link-level performance.

The considered system setup corresponding to (1) is shown in Fig. 1. As for MIMO processing, we focus on regular spatial multiplexing where a conventional transmitter separately encodes and sends an independent stream at each of its antennas without having channel state information or being aware of the transmit-side noise it produces; the receiver jointly decodes the output signals of the MIMO channel knowing its instantaneous realization $\boldsymbol{H}$ and some noise statistics. However, conventional receivers are designed and implemented based on the conventional signal model (where $\boldsymbol{v} = \boldsymbol{0}$) due to which they are prone to lapse into suboptimal *mismatched decoding* by inaccurately postulating the statistics of the actual noise term $\boldsymbol{H}\boldsymbol{v} + \boldsymbol{w}$. Even if off-the-shelf receivers can adapt to colored receiver noise, they may not be able to track the variable statistics of the component $\boldsymbol{H}\boldsymbol{v}$ propagated from the transmitter since it is correlated with the fading channel. Only an advanced receiver would be able to perform *matched decoding* knowing perfectly the noise statistics as if it was designed and implemented explicitly based on the generalized binoisy signal model (1).

A. Related Works

The key reference results for the present study are reported in [2]–[11]. These seminal works originally formulated the research niche around (1) and established the baseline understanding of MIMO communication in the presence of transmit-side noise with numerical simulations and theoretical analysis. The majority of the related works, e.g., [2], [3], [6], [8], concern regular spatial multiplexing using separate encoding like the present paper but also different variations of joint encoding have been creditably investigated, e.g., in [4], [7]. On the other hand, all the studies that we are aware of assume (implicitly) advanced receivers that know the presence of transmit noise, no matter what form of decoding is used.

Especially, the reference results are polarized such that the scope of analytical studies [6], [8] typically differs from that of studies reporting simulations [6], [7], [9] or measurements [4]–[6]. Except for [2], practical discrete modulation schemes, e.g., QAM, have not been previously analytically evaluated in the presence of transmit noise, and simulation-based studies usually concern bit/symbol/packet error rates, not transmission rates which could be more interesting when studying modern adaptive encoding. In contrast, all the analytical capacity studies assume Gaussian signaling and the throughput simulations of [3] with adaptive modulation and coding are their closest counterpart when it comes to experimental work.

If the receiver does not properly account for the additional transmit-side noise in the received signal, conventional mutual information (MI) is not anymore the correct upper bound for coded transmissions. Rather, due to mismatched decoding, one has to employ other metrics, such as *generalized mutual information* (GMI) [15], [16] adopted herein. Another common use for GMI is the analysis of bit-interleaved coded modulation [17], while also transceiver hardware impairments [18] and effects of imperfect channel state information at receiver [19]–[21] are analyzed in terms of GMI. It is important to realize that the GMI framework differs, both conceptually and in technical details, from the approach in [22], [23], where conventional MI of a modified channel model is computed for a decoder that has certain side-information about the variables in the modified system.

In the present paper, MI and GMI are evaluated using the *replica method* [24], [25], originating from the field of statistical physics and introduced to the analysis of wireless systems by [26], [27]. Since then, the replica method has been applied to various problems in communication theory, e.g., MIMO systems [28]–[31]. For some special cases like Gaussian signaling, the replica trick renders exact asymptotic results when the number of antennas grows without bound, while they can be otherwise considered accurate approximations as shown by comparisons to Monte Carlo simulations.

B. Summary of Contributions

In this paper, we investigate two aspects of binoisy MIMO channels that are unexplored in related works despite their fundamental role in understanding the effects of hardware impairments in wireless systems. Firstly, analytical capacity results are limited to Gaussian signaling while practical digital modulation is evaluated only based on simple simulations or measurements. Secondly, the earlier literature focuses on the optimistic case of matched decoding by employing receivers that are actually not available off the shelf but implicitly updated to take account of transmit-side noise.

In particular, this paper contributes to the capacity theory of MIMO communication links by examining the effects of transmit-side noise as follows.

- Analytical GMI expressions are calculated for studying the rate loss of mismatched decoding when using a conventional receiver which is unaware of the transmit-side noise. Especially, it is shown that the performance remains the same irrespective of how well the noise covariance matrix is known if it is a constant.
- The above analysis is further translated into corresponding asymptotic high-SNR limits for Gaussian signaling as a complement for the results of [10], which covers matched decoding and conventional MI.

- The analytical expressions provided for both conventional MI and GMI cover many practical discrete modulation schemes such as variations of PSK and QAM. This resolves the serious problem that evaluating (G)MI with direct Monte Carlo simulations for the present system is computationally infeasible except for cases with small number of antennas and low order modulation sets.

Extending beyond the scope of the paper, the replica analysis of GMI is also a new aspect at large.

C. Outline of the Paper and Its Nomenclature

After the considered system model is specified in the following section, the main analytical content of this paper is divided into two parts: Section III concerns the performance of conventional suboptimal receivers under mismatched decoding, which is analyzed based on GMI; and Section IV studies conventional MI with advanced receivers, which are aware of transmitter noise and, thus, capable of optimal matched decoding. In Section V, the presented theory is illustrated with numerical results, including simulations for double-checking its accuracy, which is finally followed by concluding remarks in the last section. Some general results from literature that are used throughout the paper for derivations are collected in Appendix A for the convenience of the reader. Appendices B contains general description of the replica method and Appendix C sketches the derivation of the main results in Section III.

Notation: Complex Gaussian random variables (RVs) are always assumed to be proper and the density of such $\boldsymbol{x} \in \mathbb{C}^N$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{R}$ is denoted $g(\boldsymbol{x}|\boldsymbol{\mu};\boldsymbol{R})$. For the zero-mean proper Gaussians, we say they are circularly symmetric complex Gaussian (CSCG). For convenience, both discrete and continuous RVs are said to have a probability density function (PDF) that is denoted by p, and we do not separate RVs and their realizations. For postulated PDFs we write q and add tilde on top of the related RVs (most of the time). Given a RV x that has a PDF $p(x)$, we write $x \sim p(x)$ (and $\tilde{x} \sim q(\tilde{x})$ for the postulated case). Statistical expectation is denoted $\mathsf{E}\{\cdot\}$ and, unless stated otherwise, calculated over all randomness in the argument using true or postulated PDFs, depending on which type of RVs are present. Integrals w.r.t. real-valued variables are always over $\mathbb{R}$ (for vectors over the appropriate product space) and we tend to omit the integration limits for notational simplicity. For a complex variable $z = x + \mathrm{j}y$, we denote $\int(\,)\mathrm{d}z = \int(\,)\mathrm{d}x\,\mathrm{d}y$, and similarly for complex vectors. Logarithms are natural logs and denoted ln unless stated otherwise.

II. System Model

Consider the system model depicted in Fig. 1 and the signal model of $\boldsymbol{y} \in \mathbb{C}^N$ written in (1) where $\boldsymbol{H} \in \mathbb{C}^{N\times M}$ is the channel matrix and $\boldsymbol{x} \in \mathbb{C}^M$ the signal of interest. The receive-side distortion plus noise component is divided into two parts, namely $\boldsymbol{w} = \boldsymbol{n} + \boldsymbol{\omega} \in \mathbb{C}^N$ where $\boldsymbol{n}$ is caused by thermal noise and $\boldsymbol{\omega}$ represents hardware impairments arising from the non-ideal behavior of the radio-frequency (RF) transceivers. Similarly, $\boldsymbol{v} = \boldsymbol{m} + \boldsymbol{\nu} \in \mathbb{C}^M$ where $\boldsymbol{m}$ and $\boldsymbol{\nu}$ are related to thermal noise and hardware impairments or distortions, respectively, at the transmit-side. In practice, the effect of $\boldsymbol{m}$ is often negligible compared to $\boldsymbol{\nu}$. In conventional MIMO literature it is common to consider only the thermal noise at the receiver, which translates to assuming $\boldsymbol{\omega} = \boldsymbol{\nu} = \boldsymbol{m} = \boldsymbol{0}$ in our more generic system model.

Let us denote the PDF of the transmit vector $\boldsymbol{x}$ by $p(\boldsymbol{x})$ and assume it factorizes as

$$p(\boldsymbol{x}) = \prod_{m=1}^{M} p(x_m), \tag{2}$$

so that independent streams are transmitted at each transmit antenna. Furthermore, let $p(x_m)$ be a zero-mean distribution with variance $\bar{\gamma}_m$. For later convenience, we let $\boldsymbol{\Gamma}$ be a diagonal matrix whose non-zero elements are given by $\bar{\gamma}_1, \ldots, \bar{\gamma}_M$, that is, $\boldsymbol{\Gamma} = \mathsf{E}\{\boldsymbol{x}\boldsymbol{x}^{\mathsf{H}}\}$. The channel $\boldsymbol{H}$ is assumed to have independent identically distributed (IID) CSCG elements with variance[1] $1/M$. The thermal noise samples at the transceivers are modeled as CSCG random vectors $\boldsymbol{m}$ and $\boldsymbol{n}$ that have independent elements. For simplicity, we assume that any given noise or hardware impairment component is independent of any other RVs in the system. The transmit-and receive-side impairments $\boldsymbol{\nu}$ and $\boldsymbol{\omega}$ are taken to be CSCG random vectors with covariance matrices $\boldsymbol{R}_{\boldsymbol{\nu}}$ and $\boldsymbol{R}_{\boldsymbol{\omega}}$, respectively. The distortion plus noise vectors $\boldsymbol{v}$ and $\boldsymbol{w}$ are thus CSCG random vectors whose covariance matrices we denote $\boldsymbol{R}_{\boldsymbol{v}}$ and $\boldsymbol{R}_{\boldsymbol{w}}$, respectively. Notice that these matrices can be functions of the statistics of some other RVs albeit we suppress the explicit statement of such dependence at this point for notational convenience. The SNR without transmit-side noise is defined as $\mathrm{tr}(\boldsymbol{\Gamma})/\mathrm{tr}(\boldsymbol{R}_{\boldsymbol{w}})$.

The PDF of the received signal, conditioned on $\boldsymbol{x}$, $\boldsymbol{v}$ and $\boldsymbol{H}$, is given by

$$p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{v},\boldsymbol{H}) = g\left(\boldsymbol{y}|\boldsymbol{H}(\boldsymbol{x}+\boldsymbol{v});\boldsymbol{R}_{\boldsymbol{w}}\right), \tag{3}$$

and the receiver is assumed to know $\boldsymbol{H}$ and the true distribution $p(\boldsymbol{x})$ of the channel input. However, the additional transmit-side term $\boldsymbol{v}$ is in general unknown at the receive-side and, thus, the PDF (3) cannot be directly used for detection and decoding. Herein, we consider two different scenarios for the joint decoding operation at the receiver:

1) The receiver knows $\boldsymbol{H}$, the PDFs of the noise plus distortion terms $\boldsymbol{v}$ and $\boldsymbol{w}$ as well as the distribution of the data vector $\boldsymbol{x}$. *Matched joint decoding* is then based on the conditional PDF

$$p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{H}) = \mathsf{E}_{\boldsymbol{v}}\left\{g\left(\boldsymbol{y}|\boldsymbol{H}(\boldsymbol{x}+\boldsymbol{v});\boldsymbol{R}_{\boldsymbol{w}}\right)\right\} \tag{4}$$

$$= g(\boldsymbol{y}|\boldsymbol{H}\boldsymbol{x};\boldsymbol{R}_{\boldsymbol{w}} + \boldsymbol{H}\boldsymbol{R}_{\boldsymbol{v}}\boldsymbol{H}^{\mathsf{H}}), \tag{5}$$

where the second equality follows by first using (62) to calculate the expectation w.r.t. $\boldsymbol{v}$ and simplifying the end result using (63) and (64). Note that the effective noise

[1]Typically the total power emitted from the transmit antennas in MIMO systems is constant; that is, $\mathrm{tr}(\boldsymbol{\Gamma}) = \bar{\gamma}$, where $\bar{\gamma}$ is some fixed power budget that does not depend on M. Hence the elements of $\boldsymbol{\Gamma}$ need to be functions of M to satisfy the transmit power normalization. For the following analysis, however, it is more convenient to treat the elements of $\boldsymbol{\Gamma}$ to be independent of M and let the transmit power normalization be a part of the channel. Clearly, both approaches are mathematically fully equivalent.

covariance matrix in (5) depends now on the instantaneous channel realization $\boldsymbol{H}$.

2) The receiver has perfect knowledge of $\boldsymbol{H}$ and the PDF of the data vector $\boldsymbol{x}$. Instead of (4), however, the device uses a postulated channel law

$$q(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{H}) = g(\boldsymbol{y}|\boldsymbol{H}\boldsymbol{x};\tilde{\boldsymbol{R}}), \tag{6}$$

for *mismatched joint decoding* [15], [16]. In contrast to $\boldsymbol{R}_{\boldsymbol{w}}$ in (5), that is a random matrix, the postulated covariance matrix $\tilde{\boldsymbol{R}}$ in (6) is arbitrary but fixed.

If matched joint decoding is employed, the conventional metric for evaluating the (ergodic) achievable rate of the system for given input distribution $p(\boldsymbol{x})$ is the MI between the channel inputs and outputs, namely,

$$I(\boldsymbol{y};\boldsymbol{x}) = \mathsf{E}\left\{\ln p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{H})\right\} - \mathsf{E}\left\{\ln p(\boldsymbol{y}|\boldsymbol{H})\right\}, \tag{7}$$

where $p(\boldsymbol{y}|\boldsymbol{H}) = \mathsf{E}_{\boldsymbol{x}}\{p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{H})\}$ and the expectation is w.r.t. all RVs in the system model, including the channel matrix $\boldsymbol{H}$. From the system design perspective, however, it might be impractical to use (5) due to complexity of implementation, resulting in mismatched decoding. To lower bound the true maximum rate that can be achieved reliably over channel (1) when decoding rule (6) is used at the receiver, we use GMI that is discussed in the next section.

III. Mismatched Joint Decoding: Generalized Mutual Information

A. Definition and Key Properties

Let us assume that the received signal is given by (1) but the receiver uses (6) for decoding. Given $p(\boldsymbol{x})$, the (ergodic) GMI between the channel inputs and outputs is defined as [15], [16]

$$I_{\mathsf{GMI}}(\boldsymbol{y};\boldsymbol{x}) = \sup_{s>0} I^{(s)}_{\mathsf{GMI}}(\boldsymbol{y};\boldsymbol{x}), \tag{8}$$

where, denoting $q^{(s)}(\boldsymbol{y}|\boldsymbol{H}) = \mathsf{E}_{\boldsymbol{x}}\{q(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{H})^s\}$, the s-dependent part reads

$$I^{(s)}_{\mathsf{GMI}}(\boldsymbol{y};\boldsymbol{x}) = \mathsf{E}\left\{\ln q(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{H})^s\right\} - \mathsf{E}\left\{\ln q^{(s)}(\boldsymbol{y}|\boldsymbol{H})\right\}. \tag{9}$$

Notice that since we consider ergodic rates, the expectations in (9) are w.r.t. all RVs in the system model, including the channel matrix $\boldsymbol{H}$.

The GMI in (8) gives the achievable rate for which reliable transmission over the channel (1) is possible given decoding metric (6) and ensemble of codebooks where the code words are independent with IID elements. An important property of GMI is that it always provides a valid lower bound for the maximum rate of the channel, namely, if I is the maximum ergodic rate that can be reliably transmitted over the channel (1) using input distribution $p(\boldsymbol{x})$ and decoding rule (6), then[2] $I \geq I_{\mathsf{GMI}}$. It is important to notice that the task of computing I for an arbitrary channel with an arbitrary decoding rule is in general an open research problem and the very reason why we have to resort to alternative approaches such as GMI.

[2]For the special case of IID codebooks and discrete memoryless channels with mismatched decoding, the lower bound provided by GMI is indeed tight [16]. However, there are examples (see e.g., [15], [16] and references therein) where rates higher than GMI can be obtained through other choice of code distribution. The downside of the techniques employed in the latter case is that they are often limited to finite alphabet channels and are much more cumbersome to use than the GMI, which can be easily applied to very general channel models. For more discussion on GMI, see for example, [15]–[21] and the references therein.

B. Special Case of Gaussian Signaling

We are first interested in evaluating the s-dependent part of the normalized GMI per transmit stream $M^{-1}I^{(s)}_{\mathsf{GMI}}(\boldsymbol{y};\boldsymbol{x})$ for given $s>0$. The optimization over the free parameter s is carried out after the suitable expressions are found. Note that (9) is a valid lower bound on the achievable rate for any $s>0$ and the optimization is carried out to obtain the tightest bound possible. The first term in (9) can be written as

$$\begin{aligned}
&\frac{1}{M}\mathsf{E}\left\{\ln q(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{H})^s\right\} \\
&\quad= -\overbrace{\frac{s}{M}[N\ln\pi + \ln\det\tilde{\boldsymbol{R}}]}^{=c^{(s)}} \\
&\qquad - \frac{s}{M}\mathsf{E}\left\{(\boldsymbol{H}\boldsymbol{v}+\boldsymbol{w})^{\mathsf{H}}\tilde{\boldsymbol{R}}^{-1}(\boldsymbol{H}\boldsymbol{v}+\boldsymbol{w})\right\} \\
&\quad= -c^{(s)} - \frac{s}{M}\left[\mathrm{tr}(\tilde{\boldsymbol{R}}^{-1}\boldsymbol{R}_{\boldsymbol{w}}) + \frac{1}{M}\mathrm{tr}(\tilde{\boldsymbol{R}}^{-1})\mathrm{tr}(\boldsymbol{R}_{\boldsymbol{v}})\right].
\end{aligned} \tag{10}$$

The first equality follows from (6) by the fact that $\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x} = \boldsymbol{H}\boldsymbol{v}+\boldsymbol{w}$ when $\boldsymbol{x}$ is given. The second equality is a consequence of the assumption that the channels and noise vectors are all mutually independent and $\boldsymbol{H}$ has zero-mean IID entries with variance $1/M$. Notice that (10) is independent of $p(\boldsymbol{x})$ and hence valid for all channel inputs. Evaluating the second term in (9) is more complicated but *for the special case of Gaussian inputs* we have the result shown below.

Example 1: For the special case of Gaussian inputs; that is, $p(\boldsymbol{x}) = g(\boldsymbol{x}|\boldsymbol{0};\boldsymbol{\Gamma})$,

$$\begin{aligned}
\frac{1}{M}I^{(s)}_{\mathsf{GMI}}(\boldsymbol{y};\boldsymbol{x}) = \frac{1}{M}\mathsf{E}_{\boldsymbol{H}}\Big\{&\ln\det(\tilde{\boldsymbol{R}} + s\boldsymbol{H}\boldsymbol{\Gamma}\boldsymbol{H}^{\mathsf{H}}) \\
&+ s\,\mathrm{tr}\left[\left(\boldsymbol{R}_{\boldsymbol{w}} + \boldsymbol{H}(\boldsymbol{R}_v+\boldsymbol{\Gamma})\boldsymbol{H}^{\mathsf{H}}\right)\left(\tilde{\boldsymbol{R}} + s\boldsymbol{H}\boldsymbol{\Gamma}\boldsymbol{H}^{\mathsf{H}}\right)^{-1}\right] \\
&- s\,\mathrm{tr}(\tilde{\boldsymbol{R}}^{-1}\boldsymbol{R}_{\boldsymbol{w}}) - \frac{s}{M}\mathrm{tr}(\tilde{\boldsymbol{R}}^{-1})\mathrm{tr}(\boldsymbol{R}_{\boldsymbol{v}}) - \ln\det\tilde{\boldsymbol{R}}\Big\}.
\end{aligned} \tag{11}$$

The result is obtained by first using (62) and then simplifying with (63) and (64). Inserting the RHS of (1) into the obtained expression and taking the expectations w.r.t. the noise terms $\boldsymbol{v}$ and $\boldsymbol{w}$ completes the derivation.

Example 1 shows that for Gaussian signals we only need to average over the channel matrix $\boldsymbol{H}$ to obtain the s-dependent part of GMI. This is doable with Monte Carlo simulation. However, finding the optimal s is time consuming even in this case and a simple analytical expression that does not explicitly depend on the form of the marginals in (2) would be highly desirable. With this in mind, we adopt the following restriction to our system model from the physical characteristics of typical real transmitters for simplifying the analysis.

Assumption 1: The covariance matrix for the transmit-side distortion plus noise term $\boldsymbol{v}$ is diagonal so that we may write $\boldsymbol{R}_{\boldsymbol{v}} = \boldsymbol{R}_{\boldsymbol{m}} + \boldsymbol{R}_{\boldsymbol{\nu}} = \mathrm{diag}(r^{(1)}_{\boldsymbol{v}},\ldots,r^{(M)}_{\boldsymbol{v}})$. Hence, $\boldsymbol{v}$ has

independent (but not necessarily identically distributed) entries drawn according to $p(v_m) = g(v_m|0; r_{\boldsymbol{v}}^{(m)})$.

The physical meaning of this assumption is that hardware impairments at different transmitter branches arise in separate electrical components and there are no mechanisms which generate significant correlation between the elements of the distortion noise vector. Furthermore, it is actually not necessary for the replica analysis but it helps simplify the end result to a form whose numerical evaluation is computationally easy.

C. Analytical Results via the Replica Method

If the goal is to calculate the expectations related to the latter term in (9) analytically and for general input distributions, we need to employ somewhat more advanced analytical tools than the basic probability calculus used in Example 1. As we shall see shortly, employing the replica method provides a formula that is applicable to a variety of input constellations, such as Gaussian or QAM. To begin, let us first denote

$$-\frac{1}{M}\mathsf{E}\ln q^{(s)}(\boldsymbol{y}|\boldsymbol{H}) = c^{(s)} + f(s), \tag{12}$$

where $c^{(s)}$ is defined in (10) and the latter term, equivalent of the so-called *free energy* in statistical mechanics, reads

$$f(s) = -\frac{1}{M}\mathsf{E}\left\{\ln \mathsf{E}_{\tilde{\boldsymbol{x}}}\left\{\mathrm{e}^{-[\boldsymbol{H}(\boldsymbol{x}+\boldsymbol{v}-\tilde{\boldsymbol{x}})+\boldsymbol{w}]^{\mathsf{H}} s\tilde{\boldsymbol{R}}^{-1}[\boldsymbol{H}(\boldsymbol{x}+\boldsymbol{v}-\tilde{\boldsymbol{x}})+\boldsymbol{w}]}\right\}\right\}. \tag{13}$$

Now the inner expectation over the postulated channel input $\tilde{\boldsymbol{x}}$ is w.r.t. a generic PDF (2) and cannot be solved using (62) as before. The outer expectation is w.r.t. the rest of the RVs in the system, namely $\{\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{w}, \boldsymbol{H}\}$. Due to (9) and (10) the expression to be optimized in the GMI formula thus becomes

$$\frac{1}{M} I_{\mathsf{GMI}}^{(s)}(\boldsymbol{y}; \boldsymbol{x}) = f(s) - \frac{s}{M}\left[\mathrm{tr}(\tilde{\boldsymbol{R}}^{-1}\boldsymbol{R}_{\boldsymbol{w}}) + \frac{1}{M}\mathrm{tr}(\tilde{\boldsymbol{R}}^{-1})\mathrm{tr}(\boldsymbol{R}_{\boldsymbol{v}})\right]. \tag{14}$$

Remark 1: By (13) and (14), it is clear that if the receiver assumes that the additive noise in the system is spatially white $\tilde{\boldsymbol{R}} = \tilde{r}\boldsymbol{I}_N$ with some finite sample variance $\tilde{r}$, the GMI remains the same for all $\tilde{r} > 0$ since the optimization over $s > 0$ in (8) can be replaced by an optimization over a new variable $\tilde{s} = s/\tilde{r} > 0$. Thus, if the receiver uses $\tilde{\boldsymbol{R}} = \tilde{r}\boldsymbol{I}_N$ for decoding, the GMI is the same for all $\tilde{r} > 0$ when the transmit and receive covariance matrices $\boldsymbol{R}_{\boldsymbol{v}}$ and $\boldsymbol{R}_{\boldsymbol{w}}$ are fixed.

The main obstacle in evaluating (14) is clearly $f(s)$. This term happens to be, however, of a form that can be tackled by the replica method, as outlined in Appendix B. The following result is derived in Appendix C under the assumption of the so-called *replica symmetric* (RS) ansatz (see Step 3 in Appendix B) when the system approaches the *large system limit* (LSL), that is, $M, N \to \infty$ with finite and fixed ratio $\alpha = M/N > 0$. The limit notation is omitted below and the results should therefore be interpreted as approximations for systems that have finite dimensions.

Proposition 1: Let $m = 1, \ldots, M$ and denote

$$\chi_m = x_m + v_m, \tag{15}$$
$$\tilde{\chi}_m = \tilde{x}_m, \tag{16}$$

where $x_m, \tilde{x}_m \sim p(x_m)$ and $v_m \sim g(v_m|0; r_{\boldsymbol{v}}^{(m)})$ are independent for all m by assumption. Let

$$p(z_m|\chi_m) = g(z_m|\chi_m; \eta^{-1}), \tag{17}$$
$$q(z_m|\tilde{\chi}_m) = g(z_m|\tilde{\chi}_m; \xi^{-1}), \tag{18}$$

be the PDF of an output z_m of an additive white Gaussian noise (AWGN) channel whose input is either (15) or (16), respectively, and corrupted by additive noise with variance η^{-1} or ξ^{-1}, respectively. The parameters η, ξ satisfy

$$\eta = \frac{1}{\alpha}\frac{\left[\frac{1}{N}\mathrm{tr}(\tilde{\boldsymbol{\Omega}}^{-1})\right]^2}{\frac{1}{N}\mathrm{tr}(\tilde{\boldsymbol{\Omega}}^{-1}\boldsymbol{\Omega}\tilde{\boldsymbol{\Omega}}^{-1})}, \tag{19}$$

$$\xi = \frac{1}{\alpha N}\mathrm{tr}(\tilde{\boldsymbol{\Omega}}^{-1}), \tag{20}$$

for the given matrices

$$\boldsymbol{\Omega} = \boldsymbol{R}_{\boldsymbol{w}} + \varepsilon\boldsymbol{I}_N, \tag{21}$$
$$\tilde{\boldsymbol{\Omega}} = s^{-1}\tilde{\boldsymbol{R}} + \tilde{\varepsilon}\boldsymbol{I}_N, \tag{22}$$

and variables

$$\varepsilon = \frac{1}{M}\sum_{m=1}^{M}\mathsf{E}\left\{|v_m + x_m - \langle\tilde{x}_m\rangle_q|^2\right\}, \tag{23}$$

$$\tilde{\varepsilon} = \frac{1}{M}\sum_{m=1}^{M}\mathsf{E}\left\{|\tilde{x}_m - \langle\tilde{x}_m\rangle_q|^2\right\}. \tag{24}$$

The notation $\langle\tilde{x}_m\rangle_q$ above refers to a decoupled posterior mean estimator

$$\langle\tilde{x}_m\rangle_q = \frac{\mathsf{E}_{\tilde{x}_m}\{\tilde{x}_m q(z_m|\tilde{x}_m)\}}{q(z_m)}, \tag{25}$$

where $q(z_m) = \mathsf{E}_{\tilde{\chi}_m}\{q(z_m|\tilde{\chi}_m)\}$. If we also write $p(z_m) = \mathsf{E}_{\chi_m}\{p(z_m|\chi_m)\}$, the free energy $f(s)$ defined in (14) is given under the assumption of the RS ansatz by

$$\begin{aligned} f_{\mathsf{RS}}(s) = {} & \frac{1}{\alpha N}\left[\ln\det\tilde{\boldsymbol{\Omega}} + \mathrm{tr}(\tilde{\boldsymbol{\Omega}}^{-1}\boldsymbol{\Omega}) - \ln\det(s^{-1}\tilde{\boldsymbol{R}})\right] \\ & - \left(\ln\frac{\pi}{\xi} + \frac{\xi}{\eta} + \frac{1}{M}\sum_{m=1}^{M}\int p(z_m)\ln q(z_m)\mathrm{d}z_m\right) \\ & - \xi\varepsilon + \frac{\xi(\xi-\eta)}{\eta}\tilde{\varepsilon}. \end{aligned} \tag{26}$$

If multiple solutions to the coupled fixed point (19)–(24) are found, the one minimizing (26) should be chosen.

Proof: An outline of the derivation is given in Appendix C.

The above result extends some previous works such as [26], [27] in the direction of correlated noise at the receiver and additive transmit-side impairments. It is thus clear that the original GMI term (9) of the MIMO system that suffers from transceiver hardware impairments has an interpretation in terms of an

equivalent *decoupled*[3] scalar system. This decoupled channel has only additive distortions but unlike in the conventional case of replica analysis [26], [27], the transmit-side has its own noise term. It should be remarked, however, that the implicit assumption here is that $f_{\mathsf{RS}}(s) = f(s)$; that is, the system is not replica symmetry breaking (RSB). We leave the RSB case as a possible future work and check the validity of the solution with selected numerical simulations.

For simplicity of presentation, we consider next a few practical special cases of Proposition 1 where the transmit power is the same for all antennas and the noise and distortions at the transmit-side are spatially uncorrelated, namely, $\boldsymbol{\Gamma} = \bar{\gamma}\boldsymbol{I}_M$ and $\boldsymbol{R}_{\boldsymbol{v}} = r_{\boldsymbol{v}}\boldsymbol{I}_M$. The receiver postulates spatially white noise $\tilde{\boldsymbol{R}} = \tilde{r}\boldsymbol{I}_N$ with some variance $\tilde{r} > 0$. This allows us to write

$$\frac{1}{M} I_{\mathsf{GMI}}(\boldsymbol{y};\boldsymbol{x}) = \sup_{\tilde{s}>0} \left\{ f(\tilde{s}) - \alpha^{-1}\tilde{s}\left[N^{-1}\mathrm{tr}(\boldsymbol{R}_{\boldsymbol{w}}) + r_{\boldsymbol{v}}\right] \right\}, \tag{27}$$

where $f(\tilde{s})$ is given by (13) with $s\tilde{\boldsymbol{R}}^{-1}$ replaced by $\tilde{s}\boldsymbol{I}_N$. Furthermore, in this case all variables are identically distributed for all $m = 1, 2, \ldots, M$ so we may omit the subscripts related to m in the following. We still need to fix the input distribution (2) to obtain the parameters (23) and (24). For this, we give two concrete examples: 1) Gaussian signaling; and 2) discrete channel inputs, such as, QAM.

Example 2: Let the channel inputs (2) be IID Gaussian, namely, $p(\boldsymbol{x}) = g(\boldsymbol{x}|\boldsymbol{0}; \bar{\gamma}\boldsymbol{I}_M)$ so that $p(\tilde{\chi}_m) = p(x_m) = g(x|0;\bar{\gamma})$ and $p(\chi_m) = g(\chi_m|0;\bar{\gamma} + r_{\boldsymbol{v}})$ in Proposition 1. The parameter ξ can then be obtained explicitly as

$$\xi = \frac{\bar{\gamma}\tilde{s}(1-\alpha) - \alpha + \sqrt{4\alpha\bar{\gamma}\tilde{s} + \left[\bar{\gamma}\tilde{s}(1-\alpha) - \alpha\right]^2}}{2\alpha\bar{\gamma}}, \tag{28}$$

while η and ε are obtained by solving the coupled fixed point equations

$$\eta = \frac{1}{\alpha\left[N^{-1}\mathrm{tr}(\boldsymbol{R}_{\boldsymbol{w}}) + \varepsilon\right]}, \tag{29}$$

$$\varepsilon = \frac{\eta r_{\boldsymbol{v}} + \bar{\gamma}(\eta + \xi^2\bar{\gamma})}{\eta(1+\xi\bar{\gamma})^2} = \frac{\bar{\gamma} + r_{\boldsymbol{v}}}{(1+\xi\bar{\gamma})^2} + \frac{1}{\eta(1+1/\xi\bar{\gamma})^2}. \tag{30}$$

Additional algebra shows that for IID Gaussian inputs, the free energy (26) reduces to

$$f_{\mathsf{RS}}(\tilde{s}) = \frac{1}{\alpha}\left(\frac{\xi}{\eta} + \ln\tilde{s} + \ln\frac{1}{\alpha\xi}\right) - \xi\varepsilon + \ln(1+\xi\bar{\gamma}) + \frac{\xi r_{\boldsymbol{v}}}{1+\xi\bar{\gamma}}. \tag{31}$$

Note that the expression for parameter $\tilde{\varepsilon}$ in (24) is not explicitly given here but it is implicitly a part of (28) due to relations (20) and (22).

The computational formula for obtaining the GMI with the above example is detailed in Table I. Notice that there are two non-trivial steps in the algorithm: 1) the optimization over $\tilde{s} > 0$; and 2) the problem of solving a system of two nonlinear equations with two unknowns. The first difficulty is not specific to the current study and is present in any work that considers GMI as means to analyze mismatched decoding. The computational complexity of the second problem is negligible compared to the original task of taking an expectation over the channel matrices in (11). Indeed, a typical solution for η and ε is obtained after some tens of iterations of an iterative substitution method.

TABLE I
HOW TO OBTAIN GMI FOR GAUSSIAN SIGNALING FROM EXAMPLE

1) Choose the parameters that define the MIMO system of interest, namely, antenna ratio $\alpha = M/N$, transmit- and receive-side distortion plus noise covariance matrices $\boldsymbol{R}_{\boldsymbol{v}} = r_{\boldsymbol{v}}\boldsymbol{I}_M$ and $\boldsymbol{R}_{\boldsymbol{w}}$, respectively, and the average transmit power per antenna $\bar{\gamma}$. Let also the optimization parameter $\tilde{s} > 0$ be given.
2) Plug the values of $\{\alpha, \bar{\gamma}, \tilde{s}\}$ to (28) and obtain ξ.
3) Insert ξ along with the rest of the necessary parameters in (29) and (30), and solve η numerically, e.g., using an iterative substitution method.
4) Use the solutions of ξ and η in (31) to obtain the free energy.
5) Optimize (27) over $\tilde{s} > 0$.

For the high-SNR case where $\bar{\gamma} \to \infty$ for a fixed covariance matrix $\boldsymbol{R}_{\boldsymbol{w}}$, the result in Example 2 can be further simplified as shown in Example 3 below.

Example 3: Let us consider the case of Gaussian signaling as given in Example 2 in the limit $\bar{\gamma} \to \infty$. We assume for simplicity (see, e.g., [10]) that $\boldsymbol{R}_{\boldsymbol{w}} = r_{\boldsymbol{w}}\boldsymbol{I}_N$ and $r_{\boldsymbol{v}} = \bar{\gamma}\kappa^2$ where $\kappa > 0$ and $r_{\boldsymbol{w}} > 0$ are fixed and finite parameters. At high-SNR, there are two possibilities for the parameter $\tilde{s} = s/\tilde{r}$ in the GMI: 1) the optimal value of $\tilde{s}$ is a strictly positive constant; and 2) the value of $\tilde{s}$ goes to zero when $\bar{\gamma} \to \infty$. For the first case, $M^{-1}I^{(s)}_{\mathsf{GMI}}(\boldsymbol{y};\boldsymbol{x}) \to -\infty$ so to obtain a consistent solution for the fixed point equations, the parameter $\tilde{s}$ has to be inversely proportional to $\bar{\gamma}$, i.e., $\tilde{s} = s_{\bar{\gamma}}/\bar{\gamma}$ where $s_{\bar{\gamma}}$ is a strictly positive finite constant. Then $\xi \to 0$ as $\bar{\gamma} \to \infty$, and the normalized GMI reduces to

$$\frac{1}{M} I^{\infty}_{\mathsf{GMI}}(\boldsymbol{y};\boldsymbol{x}) = \sup_{s_{\bar{\gamma}}>0} \left\{ \frac{1}{\alpha}\ln\left(\frac{s_{\bar{\gamma}}}{\alpha\xi_{\bar{\gamma}}}\right) + \ln(1+\xi_{\bar{\gamma}}) + \frac{\kappa^2\xi_{\bar{\gamma}}}{1+\xi_{\bar{\gamma}}} - \frac{s_{\bar{\gamma}}\kappa^2}{\alpha} \right\}, \tag{32}$$

in the limit $\bar{\gamma} \to \infty$. The auxiliary parameter $\xi_{\bar{\gamma}} \triangleq \xi\bar{\gamma} > 0$ is given by

$$\xi_{\bar{\gamma}} = \frac{s_{\bar{\gamma}}(1-\alpha) - \alpha + \sqrt{4\alpha s_{\bar{\gamma}} + \left[s_{\bar{\gamma}}(1-\alpha) - \alpha\right]^2}}{2\alpha}. \tag{33}$$

Compared to the finite-SNR case in Example 2, the GMI is now directly given by (32) and there are no fixed-point equations that need to be solved.

The next example provides explicit formulas for the computation of GMI given finite discrete constellations, such as, PSK or QAM.

Example 4: Let $\mathcal{A}$ be a discrete modulation alphabet with fixed and finite cardinality $|\mathcal{A}|$ and consider the GMI (27). Let the channel inputs x_m be drawn independently and uniformly

[3]This decoupling property is ubiquitous in replica analysis (see for example [26], [27]) as well as in random matrix theory (see [32], [33] and references therein), and is one of the key reasons why the asymptotic methods provide computationally feasible solutions for complex problems.

from $\mathcal{A}$. The parameters of the decoupled channel model in Proposition 1 can be obtained by first solving ξ and $\tilde{\varepsilon}$ from

$$\xi = \frac{\tilde{s}}{\alpha(1+\tilde{s}\tilde{\varepsilon})}, \tag{34}$$

$$\tilde{\varepsilon} = \bar{\gamma} - \int q(z)\,|\langle \tilde{x}\rangle_q|^2\,\mathrm{d}z, \tag{35}$$

using the following definitions for the decoupled estimator and the postulated channel probability

$$\langle \tilde{x}\rangle_q = \frac{1}{q(z)|\mathcal{A}|}\sum_{\tilde{x}\in\mathcal{A}} \tilde{x}\, g(z|\tilde{x};\xi^{-1}), \tag{36}$$

$$q(z) = \frac{1}{|\mathcal{A}|}\sum_{x\in\mathcal{A}} g(z|x;\xi^{-1}), \tag{37}$$

respectively. Note that this implies solving two parameters from two nonlinear equations and can be done, for example, by using an iterative substitution method. After obtaining the solutions for ξ (and $\tilde{\varepsilon}$), the rest of the parameters can be obtained by solving the two coupled equations

$$\eta = \frac{1}{\alpha\left[N^{-1}\mathrm{tr}(\boldsymbol{R}_{\boldsymbol{w}})+\varepsilon\right]}, \tag{38}$$

$$\varepsilon = \mathsf{E}\left\{\left|v + x - \langle \tilde{x}\rangle_q\right|^2\right\}, \tag{39}$$

for η and ε, where the expectation is w.r.t. the true joint probability of $\{x, v, z\}$. Finally, the free energy reads

$$f_{\mathsf{RS}}(\tilde{s}) = \frac{1}{\alpha}\left(\frac{\xi}{\eta} + \ln\tilde{s} + \ln\frac{1}{\alpha\xi}\right) - \xi\varepsilon + \frac{\xi(\xi-\eta)}{\eta}\tilde{\varepsilon} - \left(\frac{\xi}{\eta} + \ln\frac{\pi}{\xi} + \int p(z)\ln q(z)\mathrm{d}z\right), \tag{40}$$

where we denoted

$$p(z) = \frac{1}{|\mathcal{A}|}\sum_{x\in\mathcal{A}} g(z|x;\eta^{-1}+r_{\boldsymbol{v}}), \tag{41}$$

for the decoupled PDF of the received signal.

Notice that the form of η in Example 4 is the same as in Example 2, but the parameter ε has now a different structure. Compared to the Gaussian case, the equivalent result for IID discrete channel inputs looks in general more cumbersome. First of all, we need to solve now two sets of equations instead of just one. They both contain terms that involve $|\mathcal{A}|$ summations and there are also two expectations left to evaluate, one in (35) and another in (39). However, both expectations involve only scalar variables. This is in stark contrast to the original problem that involved computing $|\mathcal{A}|^M$ summations for every channel and noise/distortion realization and taking expectation over the channel and noise that are multidimensional integrals. This makes direct Monte Carlo computation of the GMI for discrete signaling in practice infeasible for large constellations and numbers of antennas.

IV. Matched Joint Decoding

A. Definition and the Special Case of Gaussian Signaling

Let us now consider the case of matched decoding where the correct channel transition probability (5) is utilized at the receiver. The first entropy term in (7) reads

$$\mathsf{E}\left\{\ln p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{H})\right\} = -\mathsf{E}_{\boldsymbol{H}}\left\{\ln\det(\boldsymbol{R}_{\boldsymbol{w}} + \boldsymbol{H}\boldsymbol{R}_{\boldsymbol{v}}\boldsymbol{H}^{\mathsf{H}})\right\} - c, \tag{42}$$

where $c = N\ln(\mathrm{e}\pi)$. It should be remarked that there is still an expectation left w.r.t. the channel realizations $\boldsymbol{H}$ in (42). This could be evaluated, for example, using Monte Carlo methods or random matrix theory [32], [33]. For the special case of Gaussian inputs, the identities in Appendix A allow us to partially calculate also the latter entropy term in (7), providing the following result that is useful for Monte Carlo simulations.

Example 5: Let $p(\boldsymbol{x}) = g(\boldsymbol{x}|\boldsymbol{0};\boldsymbol{\Gamma})$.Then,

$$\frac{1}{M}I(\boldsymbol{y};\boldsymbol{x}) = \frac{1}{M}\mathsf{E}_{\boldsymbol{H}}\left\{\ln\det\left(\boldsymbol{R}_{\boldsymbol{w}} + \boldsymbol{H}(\boldsymbol{\Gamma}+\boldsymbol{R}_{\boldsymbol{v}})\boldsymbol{H}^{\mathsf{H}}\right)\right\} - \frac{1}{M}\mathsf{E}_{\boldsymbol{H}}\left\{\ln\det(\boldsymbol{R}_{\boldsymbol{w}} + \boldsymbol{H}\boldsymbol{R}_{\boldsymbol{v}}\boldsymbol{H}^{\mathsf{H}})\right\}, \tag{43}$$

is the normalized ergodic MI for matched decoding.

The above expression is relatively easy to compute also by brute-force Monte Carlo methods since there is only an expectation over the fading. Unfortunately, to the best of our knowledge, the latter entropy term in (7) is mathematically intractable for rigorous methods like random matrix theory when $p(\boldsymbol{x})$ is an arbitrary distribution that satisfies (2). For example, given discrete inputs as in Example 4, calculating $\mathsf{E}\{\ln p(\boldsymbol{y}|\boldsymbol{H})\}$ and combining it with (42) reduces the MI to (44), shown at the bottom of the page. This form is computationally very complex and can be evaluated using Monte Carlo methods only for small number of antennas and simple constellations. To obtain a result for general input distribution $p(\boldsymbol{x})$ that has lower computational complexity, we resort to the replica method (see Appendix B). As before, the results that follow have been written in a simplified form where the assumption of LSL is suppressed for notational simplicity.

B. Analytical Results via the Replica Method

Proposition 2: Let us write for notational convenience

$$\chi_m = x_m + v_m, \qquad m = 1,\ldots,M, \tag{45}$$

$$I(\boldsymbol{y};\boldsymbol{x}) = M\ln|\mathcal{A}| - N - \frac{1}{|\mathcal{A}|}\sum_{\boldsymbol{x}\in\mathcal{A}^M}\mathsf{E}_{\boldsymbol{v},\boldsymbol{w},\boldsymbol{H}}\left\{\ln\left(\sum_{\tilde{\boldsymbol{x}}\in\mathcal{A}^M}\mathrm{e}^{-[\boldsymbol{H}(\boldsymbol{x}-\tilde{\boldsymbol{x}}+\boldsymbol{v})+\boldsymbol{w}]^{\mathsf{H}}(\boldsymbol{R}_{\boldsymbol{w}}+\boldsymbol{H}\boldsymbol{R}_{\boldsymbol{v}}\boldsymbol{H})^{-1}[\boldsymbol{H}(\boldsymbol{x}-\tilde{\boldsymbol{x}}+\boldsymbol{v})+\boldsymbol{w}]}\right)\right\} \tag{44}$$

where $x_m \sim p(x_m)$ and $v_m \sim g(v_m|0; r_{\boldsymbol{v}}^{(m)})$ are independent for all m. Let

$$p(z_m|\chi_m) = g(z_m|\chi_m; \eta^{-1}), \tag{46}$$

be a conditional PDF of an AWGN channel whose input is (45) and noise variance is η^{-1}. The conditional mean estimator of χ_m received over this channel reads

$$\langle \chi_m \rangle = \frac{\mathsf{E}_{\chi_m}\left\{\chi_m p(z_m|\chi_m)\right\}}{\mathsf{E}_{\chi_m}\left\{p(z_m|\chi_m)\right\}}, \tag{47}$$

where the parameter η is given, along with another parameter ε, as the solution to the coupled fixed point equations

$$\eta = \frac{1}{\alpha N}\mathrm{tr}\left[(\boldsymbol{R}_{\boldsymbol{w}} + \varepsilon \boldsymbol{I}_N)^{-1}\right], \tag{48}$$

$$\varepsilon = \frac{1}{M}\sum_{m=1}^{M}\left[\bar{\gamma}_m + r_{\boldsymbol{v}}^{(m)} - \mathsf{E}\left|\langle \chi_m \rangle\right|^2\right]. \tag{49}$$

If we also define a second set of parameters η' and ε' that are solutions to the coupled fixed point equations

$$\eta' = \frac{1}{\alpha N}\mathrm{tr}\left[(\boldsymbol{R}_{\boldsymbol{w}} + \varepsilon' \boldsymbol{I}_N)^{-1}\right], \tag{50}$$

$$\varepsilon' = \frac{1}{M}\sum_{m=1}^{M}\frac{r_{\boldsymbol{v}}^{(m)}}{1+\eta' r_{\boldsymbol{v}}^{(m)}}, \tag{51}$$

the per-stream MI is finally given by

$$\frac{1}{M}I(\boldsymbol{y};\boldsymbol{x}) = \frac{\ln\det(\boldsymbol{R}_{\boldsymbol{w}} + \varepsilon \boldsymbol{I}_N) - \ln\det(\boldsymbol{R}_{\boldsymbol{w}} + \varepsilon' \boldsymbol{I}_N)}{\alpha N}$$
$$-(\eta\varepsilon - \eta'\varepsilon') + \frac{1}{M}\sum_{m=1}^{M}\left[I(z_m;\chi_m) - \ln\left(1+\eta' r_{\boldsymbol{v}}^{(m)}\right)\right], \tag{52}$$

where

$$I(z_m;\chi_m) = -1 - \ln\frac{\pi}{\eta} - \int p(z_m)\ln p(z_m)\mathrm{d}z_m, \tag{53}$$

is the MI of the Gaussian channel defined by (45) and (46).

Proof: The result can be obtained using Appendix B for two separate MIMO channels. For the first one, we replace everywhere $\boldsymbol{x}_a \to \boldsymbol{x}_a + \boldsymbol{v}_a,\ a = 0, 1, \ldots, u$ and an application of the RM provides the equations (45)–(49). The formulas (50)–(53), on the other hand, are obtained by substituting $\boldsymbol{x}_a \to \boldsymbol{v}_a,\ a = 0, 1, \ldots, u$ in Appendix B.

Just like Proposition 1 in Section III, Proposition 2 is valid for any input distribution that satisfies (2). The solutions to the coupled (48) and (49) as well as (50) and (51) can be obtained numerically, e.g., using an iterative substitution method.

For concreteness, we again give examples for Gaussian and discrete signaling when the noise plus distortion is spatially white $\boldsymbol{R}_{\boldsymbol{v}} = r_{\boldsymbol{v}}\boldsymbol{I}_M$ and transmit power is uniformly allocated $\boldsymbol{\Gamma} = \bar{\gamma}\boldsymbol{I}_M$. This makes the channels $m = 1, 2, \ldots, M$ identically distributed so we omit the subscript m in the following.

Example 6: Let $\boldsymbol{R}_{\boldsymbol{v}} = r_{\boldsymbol{v}}\boldsymbol{I}_M$ and consider the special case of Gaussian inputs $p(\boldsymbol{x}) = g(\boldsymbol{x}|\boldsymbol{0}; \bar{\gamma}\boldsymbol{I}_M)$. Then

$$I(z;\chi) = \ln\left[1 + \eta(\bar{\gamma} + r_{\boldsymbol{v}})\right], \tag{54}$$

$$\varepsilon = \frac{\bar{\gamma} + r_{\boldsymbol{v}}}{1 + \eta(\bar{\gamma} + r_{\boldsymbol{v}})}, \tag{55}$$

and the rest of the parameters are given in Proposition 2.

We next consider the high-SNR case $\bar{\gamma} \to \infty$ as in Example 3 and compare it to the result obtained in [10] using completely different mathematical methods.

Example 7: For the case $\boldsymbol{R}_{\boldsymbol{w}} = r_{\boldsymbol{w}}\boldsymbol{I}$, $\boldsymbol{R}_{\boldsymbol{v}} = \kappa^2\bar{\gamma}\boldsymbol{I}$ (see, e.g., [10]) we find that if $\alpha \leq 1$ then $\bar{\gamma} \to \infty$ yields $\eta = \eta'$ and $\varepsilon = \varepsilon'$. The high SNR limit is therefore

$$\frac{1}{M}I^{\infty}(\boldsymbol{y};\boldsymbol{x}) = \log\left(\frac{1+\kappa^2}{\kappa^2}\right), \qquad \alpha \leq 1. \tag{56}$$

For the case $\alpha > 1$, both η and η' tend to zero at high SNR while ε and ε' grow without bound. This is not yet sufficient to solve (52). However, combining this with the relations $\eta'\varepsilon' = \eta\varepsilon$ and $\varepsilon' = \varepsilon\frac{\kappa^2}{1+\kappa^2}$, that hold in the limit $\bar{\gamma} \to \infty$ for $\alpha > 1$, provides the second part of the high SNR result

$$\frac{1}{M}I^{\infty}(\boldsymbol{y};\boldsymbol{x}) = \frac{1}{\alpha}\log\left(\frac{1+\kappa^2}{\kappa^2}\right), \qquad \alpha > 1. \tag{57}$$

The asymptotic mutual information expressions in (56) and (57) coincide exactly with the results obtained previously in [10], as expected.

Example 8: If the channel inputs are from a discrete alphabet $\mathcal{A}$ as in Example 4, the parameter ε in (49) is obtained using

$$\langle \chi \rangle = \frac{1}{p(z)}\sum_{x\in\mathcal{A}}\left[\frac{1}{|\mathcal{A}|}g(z|x;\eta^{-1}+r_{\boldsymbol{v}})\left(\frac{x+\eta r_{\boldsymbol{v}} z}{1+\eta r_{\boldsymbol{v}}}\right)\right], \tag{58}$$

$$\mathsf{E}\left|\langle\chi\rangle\right|^2 = \int p(z)\mathsf{E}\left\{\left|\langle\chi\rangle\right|^2\right\}\mathrm{d}z, \tag{59}$$

in Proposition 2. Here $p(z)$ is given by (41) and $\langle\chi\rangle$ denotes the conditional mean estimator of (45) from the observations (46).The related MI term reads by definition

$$I(z;\chi) = \ln\left(\frac{\eta}{\mathrm{e}\pi}\right) - \int p(z)\ln p(z)\,\mathrm{d}z. \tag{60}$$

Both (49) and (60) need, in general, to be solved numerically.

V. Numerical Examples

In the following, assume for simplicity that $\boldsymbol{\Gamma} = \bar{\gamma}\boldsymbol{I}$, $\boldsymbol{R}_{\boldsymbol{w}} = \boldsymbol{I}$ and $\boldsymbol{R}_{\boldsymbol{v}} = \kappa^2\bar{\gamma}\boldsymbol{I}$, where $\kappa = 10^{\mathsf{EVM}/20}$ and EVM denotes the EVM of the transmitter in decibels. The SNR without transmit-side noise is therefore simply $\bar{\gamma}$, or in decibels, $\bar{\gamma}_{\mathsf{dB}} = 10\log_{10}(\bar{\gamma})$. Furthermore, all cases assume a symmetric antenna setup $\alpha = M/N = 1$ for simplicity.

The first numerical experiment plotted in Fig. 2 examines the accuracy of the asymptotic analytical results when applied to finite-sized systems. The EVM is fixed to a rather pessimistic value $\mathsf{EVM} = -10$ dB to highlight the differences between the ideal and imperfect hardware configurations. The normalized

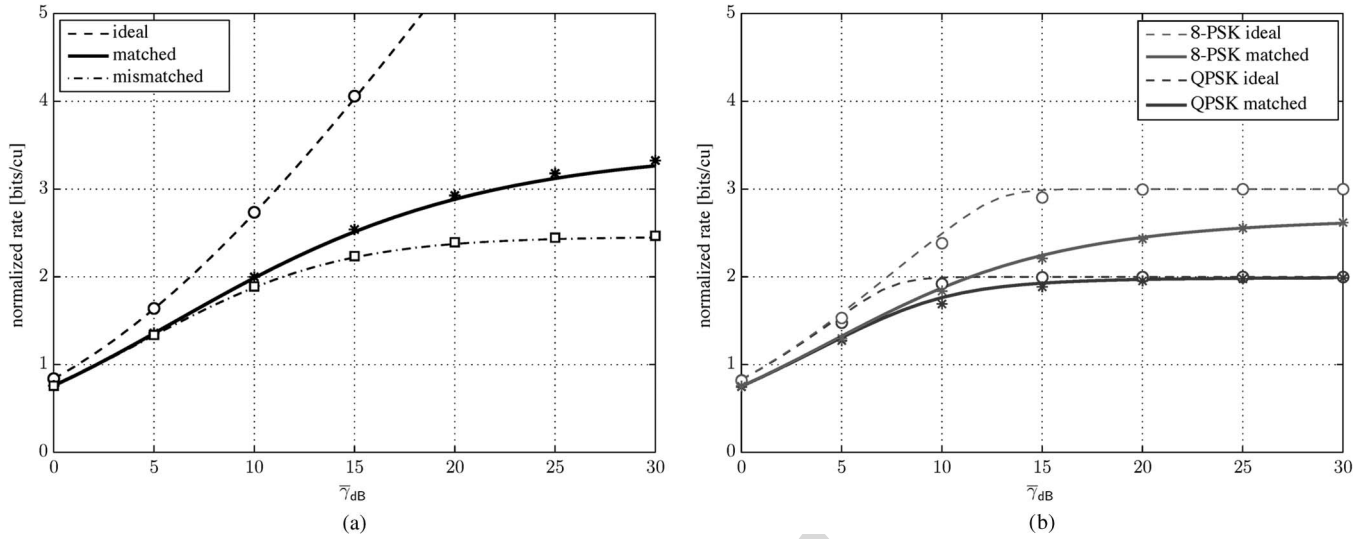

Fig. 2. Normalized rate $M^{-1}I(\boldsymbol{y};\boldsymbol{x})$ in bits per channel use (cu) vs. SNR for MIMO transmission. Lines for replica results and markers for Monte Carlo simulations for $M = N = 4$ antenna configuration. Selected cases of ideal hardware $\mathsf{EVM} = -\infty$ dB and hardware impairments ($\mathsf{EVM} = -10$ dB) with matched and mismatched decoding are plotted. (a) Gaussian signaling. (b) Discrete signaling.

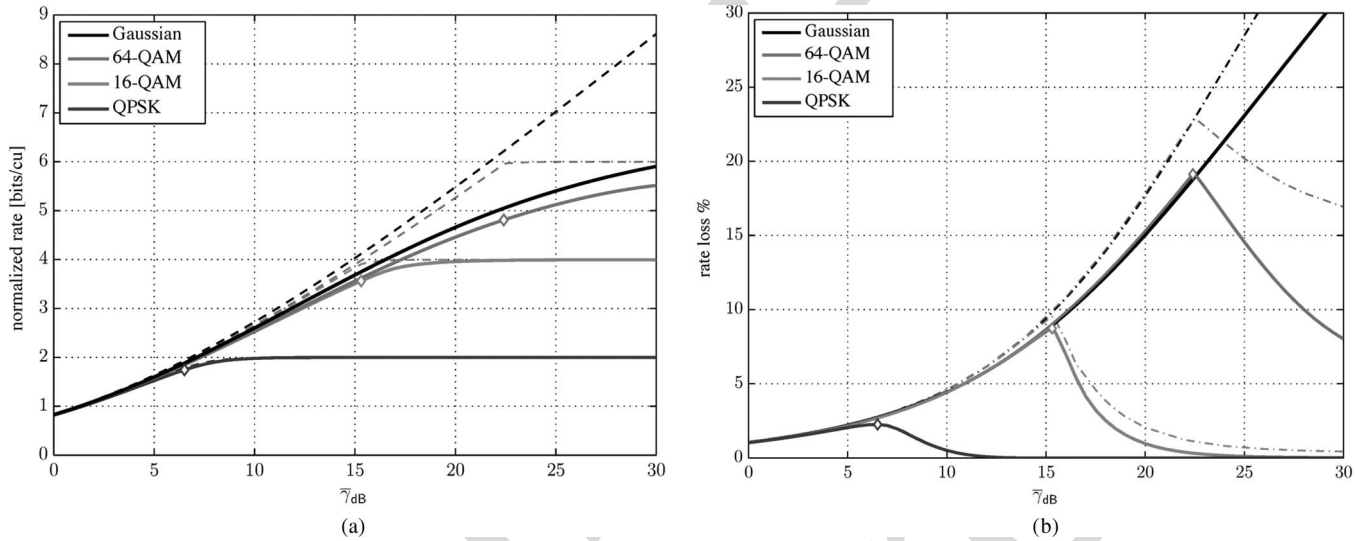

Fig. 3. Performance of a MIMO system with $M = N$ antennas and given ideal ($\mathsf{EVM} = -\infty$ dB) or non-ideal hardware ($\mathsf{EVM} = -20$ dB) for different signaling methods. Markers depict the points where discrete constellations and matched decoding with hardware impairments experience the maximum rate losses compared to the ideal cases. (a) Normalized rate $M^{-1}I(\boldsymbol{y};\boldsymbol{x})$ given ideal hardware (dashed lines) or non-ideal hardware and matched decoding (solid lines). (b) Rate loss percentage compared to ideal hardware for matched (solid lines) and mismatched (dash-dotted lines) decoding.

rate is shown using the asymptotic replica analysis (lines) and Monte Carlo simulations (markers) for a finite-size symmetric antenna setup with $M = N = 4$. In the case of Gaussian signaling, plotted in Fig. 2(a), the analytical approximations for the normalized rate $M^{-1}I(\boldsymbol{y};\boldsymbol{x})$ given by Examples 2 and 5 are quite good when compared to the finite size simulations based on Examples 1 and 5. For discrete signaling depicted in Fig. 2(b) we have plotted only the case of matched decoding due to the computational complexity of Monte Carlo simulations in the mismatched case. The gap between asymptotic result presented in Example 8 and Monte Carlo averaging of (44) is similar to the Gaussian case for both constellations. Fig. 2 shows that the analytical approximation given by the replica method is reasonably good already at $M = N = 4$, even though formally the limit $M, N \to \infty$ is required by the analysis. Note that Monte Carlo simulation of (44) has exponential computational complexity and the system size cannot be increased much higher than $M = 4$. Therefore, the rest of the examples are generated using only the analytical results given in the previous sections.

Fig. 3 illustrates the performance of an $M = N$ MIMO system for a more realistic EVM value $\mathsf{EVM} = -20$ dB. For the case of matched decoding we used Examples 6 and 8, while Examples 2 and 4 were used to obtain the curves representing mismatched decoding. In Fig. 3(a), the normalized rate $M^{-1}I(\boldsymbol{y};\boldsymbol{x})$ is depicted as a function of SNR $\bar{\gamma}$ in decibels. For clarity of presentation, we have plotted only the ideal case and the case of non-ideal hardware with matched decoding. The Gaussian curves (black lines) here are the same as the simulation curves in [10, Fig. 2] given the parameter value

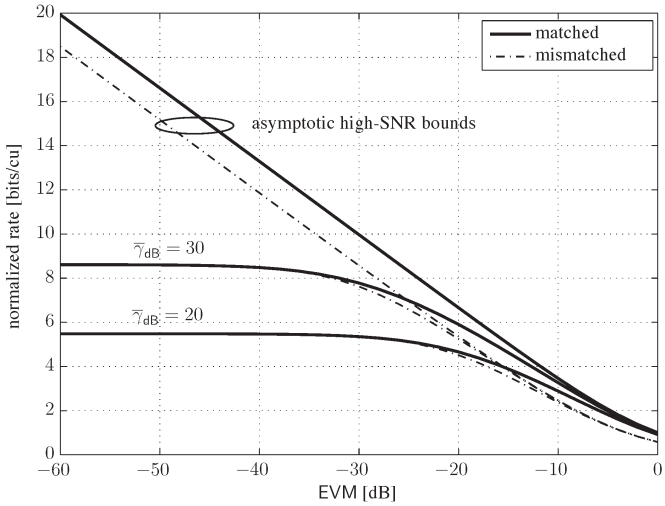

Fig. 4. Normalized rate $M^{-1}I(\boldsymbol{y};\boldsymbol{x})$ in bits per channel use vs. EVM in decibels for MIMO transmission with Gaussian signaling. Solid lines for matched decoding and dash-dotted lines for mismatched decoding.

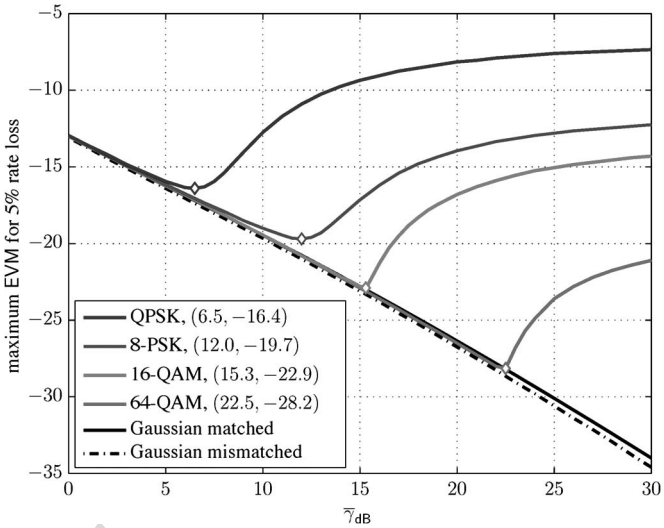

Fig. 5. Maximum allowed EVM in decibels for matched decoding so that the system experiences at most 5% loss in rate compared to the case with ideal hardware ($\mathsf{EVM} = -\infty$ dB). Markers depict the worst case EVM requirement for the discrete constellations and parenthesis in the legend provide the respective values as $(\bar{\gamma}_{\mathrm{dB}}, \mathsf{EVM})$. All discrete cases correspond to matched joint decoding at the receiver.

$\kappa = 0.1$. Apart from 64-QAM and Gaussian signaling, the figure seems to imply that lower order constellations exhaust the source entropy before the transmit-side noise has any significant effect for this choice of EVM. To see more clearly the effect of transmit noise, Fig. 3(b) shows the rate loss (in percentage) for the case with transmit noise $\mathsf{EVM} = -20$ dB when compared to the ideal case $\mathsf{EVM} = -\infty$ dB. The solid lines represent again matched decoding while dash-dotted lines are for mismatched decoding. As expected, mismatched decoding reduces the achievable rate when compared to matched decoding, but the effect is relatively minor when compared to the total rate loss caused by the presence of transmit noise itself. The markers depict the points where maximum relative rate loss is experienced for matched decoding. The same markers are also plotted in Fig. 3(a) for comparison.

In Fig. 4 we have plotted the asymptotic high-SNR results given in Examples 3 and 7. Note that given a finite value of EVM, the normalized rates for matched and mismatched decoding have a gap in this case. For more realistic, but still quite high SNR values of 20 dB and 30 dB, the two decoding strategies converge to the same value roughly when $\bar{\gamma}_{\mathrm{dB}} < -\mathsf{EVM}$. The apparent discrepancy is explained by recalling that the asymptotic cases assume $\bar{\gamma} \to \infty$ for a fixed and nonzero EVM and, thus, as a finite SNR approximation implies $\bar{\gamma} \gg 1/\kappa^2$. As may be observed from the lower right corner of the figure, the SNR values 20 dB and 30 dB have also a similar behavior near $\bar{\gamma} \gg 1/\kappa^2$. Thus, the high-SNR result is consistent with the finite-SNR cases.

It is important to guarantee certain performance when designing a system. The maximum EVM that leads to at most 5% rate loss (as compared to having ideal hardware) for a fixed input distribution and different given SNRs is plotted in Fig. 5. For Gaussian signaling we have plotted both the matched and mismatched cases while discrete cases assume matched joint decoding for simplicity. As expected, the EVM requirement for Gaussian signaling is a monotonically decreasing, but not linear, function of SNR. A simple linear approximation that provides a lower bound for the case of Gaussian signaling with matched decoding is given by

$$\mathsf{EVM} = -0.7 \cdot \bar{\gamma}_{\mathrm{dB}} - 13, \tag{61}$$

in decibels for the depicted region. This can be used as a simple rule-of-thumb for worst-case maximum allowed EVM in the system, although we recommend that EVM target values obtained in this way are always rounded down to 1–5 dB precision to include extra safety margin. For discrete constellations, the EVM requirement first follows the Gaussian case but then starts to get looser at higher SNRs. This is expected, as can be observed from Fig. 3(a), since the maximum achievable rate for a discrete constellation saturates at a certain SNR when the input distribution runs out of entropy. After this point, the rate loss can be held fixed for increasing SNR by increasing the transmit-side noise variance, or EVM, accordingly.

VI. Conclusion and Future Work

Considering a 'binoisy' channel model, we have derived asymptotic expressions for the achievable rate of SU-MIMO systems suffering from transceiver hardware impairments. For matched decoding, where the receiver is designed and implemented explicitly based on the generalized system model, expressions for the ergodic mutual information between the channel inputs and outputs have been given. In addition, a simplified receiver that neglected the hardware imperfections and performed mismatched detection and decoding has been studied via generalized mutual information. The mathematical expressions provided in the paper cover practical discrete modulation schemes, such as, quadrature amplitude modulation, as well as Gaussian signaling. The numerical results showed that for realistic system parameters, the effects of transmit-side noise and mismatched decoding become significant only at high modulation orders. Furthermore, the effect of mismatched

decoding was found to be relatively minor compared to the total rate loss caused by the presence of transmit noise itself. The results were also used to identify the maximum EVM values that allows for certain system operation.

A. Future Work

For the ease of exposition, the present paper considered the analysis of a relatively simple SU-MIMO system where the channel had IID Gaussian elements. An extension of the present replica analysis to correlated or Ricean fading channels can be done by following, for example, the analysis in [29] and [30], respectively. Establishing the effects of transmit-side noise for the cases of correlated and non-Rayleigh fading channels is an important avenue for future work.

As a further extension, it is important to investigate whether similar phenomena as observed in the present paper are present also for more complicated signal models with discrete channel inputs. Such systems already analyzed in the ideal setting with the replica method include, for example, multiuser MIMO and base station collaboration [34], channels with interference and precoding [35] and K-hop relay channels [36]. Combining the ideas from the present paper and [34]–[36] would provide a possible approach to solving such cases.

Appendix A
Useful Results

Here we collect useful results that are used often in the paper. All matrix operations below are implicitly assumed to be well-defined. The *Gaussian integration* formula for vector $\boldsymbol{x} \in \mathbb{C}^N$ is given by (see, e.g., [37, Appendix I])

$$\frac{1}{\pi^N} \int \mathrm{e}^{-\boldsymbol{x}^{\mathsf{H}} \boldsymbol{M} \boldsymbol{x} + 2\Re\{\boldsymbol{b}^{\mathsf{H}} \boldsymbol{x}\}} \mathrm{d}\boldsymbol{x} = \frac{1}{\det(\boldsymbol{M})} \mathrm{e}^{\boldsymbol{b}^{\mathsf{H}} \boldsymbol{M}^{-1} \boldsymbol{b}}, \tag{62}$$

and used in Sections II–IV and Appendix C. Similarly, the *matrix inversion lemma* [38]

$$(\boldsymbol{W}^{-1} + \boldsymbol{U}\boldsymbol{T}^{-1}\boldsymbol{V}^{\mathsf{H}})^{-1} = \boldsymbol{W} - \boldsymbol{W}\boldsymbol{U}(\boldsymbol{T} + \boldsymbol{V}^{\mathsf{H}}\boldsymbol{W}\boldsymbol{U})^{-1}\boldsymbol{V}^{\mathsf{H}}\boldsymbol{W}, \tag{63}$$

and the related *determinant identity*

$$\det(\boldsymbol{W}^{-1} + \boldsymbol{U}\boldsymbol{T}^{-1}\boldsymbol{V}^{\mathsf{H}}) = \det(\boldsymbol{T} + \boldsymbol{V}^{\mathsf{H}}\boldsymbol{W}\boldsymbol{U}) \det(\boldsymbol{W}^{-1}) \det(\boldsymbol{T}^{-1}), \tag{64}$$

are employed several times in the paper.

Appendix B
Replica Method

Consider a function Z that maps RVs to positive real numbers[4] and define two sets of RVs, $V \in \mathcal{V}$ and $X \in \mathcal{X}$, with joint probability $P_{V,X}$. Assume for convenience that $P_{V,X}$ can be described in terms of a joint PDF $p(V, X)$ and denote the marginal PDFs of X and V $p_X(X)$ and $p_V(V)$, respectively. Then, both in statistical mechanics and communication theory, we often encounter a formula

$$\begin{aligned} f &= -\frac{1}{M} \mathsf{E}_V \left\{ \ln \mathsf{E}_X \left\{ Z(V, X) \right\} \right\} \\ &= -\frac{1}{M} \int_{\mathcal{V}} p_V(V) \ln Z(V) \mathrm{d}V, \end{aligned} \tag{65}$$

where $Z(V) = \int_{\mathcal{X}} p_X(X) Z(V, X) \mathrm{d}X$. In physics jargon, the variables V are said to be *quenched* and the quantity (65) is the average *free energy density* of a system whose *partition function* is $Z(V)$. Two concrete examples of (65) are:

1) Let $Z(V, X) = g(\boldsymbol{y} | \boldsymbol{H}\boldsymbol{x}; \boldsymbol{R}_{\boldsymbol{w}})$ be the conditional PDF of the observation in an ideal MIMO channel with $V = \{\boldsymbol{y}, \boldsymbol{H}\}$ and $X = \{\boldsymbol{x}\}$, where $\boldsymbol{x}$ has IID elements from a discrete modulation set $\mathcal{A}$, such as PSK or QAM. Then (65) represents a normalized version of the second term in (7), namely, the (normalized) total *entropy* of the received signal $\boldsymbol{y}$ given a realization of $\boldsymbol{H}$ and averaged over all possible realizations of $\boldsymbol{H}$.
2) Let $Z(V, X) = \mathrm{e}^{\beta \boldsymbol{\sigma}^{\mathsf{H}} \boldsymbol{J} \boldsymbol{\sigma}}$, where $\beta > 0$ denotes the inverse temperature, $V = \boldsymbol{J} \in \mathbb{R}^{M \times M}$ a coupling matrix and $X = \boldsymbol{\sigma} \in \{\pm 1\}^M$ a spin configuration. If $p_V(V)$ is a uniform probability over $\boldsymbol{\sigma}$ and $\boldsymbol{J}$ has, e.g., IID Gaussian elements, then (65) is the average *free energy density* of a mean-field Ising spin glass in the absence of external field (up to trivial constants).

In both cases, f captures important properties of the system at hand and obtaining a computable formula for (65) would be of great interest. This seems infeasible though since the number of terms in the expectation is exponential in M.

A. Outline of the Replica Method

One method for solving (65) is the *replica method* (RM) from equilibrium statistical mechanics. While the RM is extremely versatile, it unfortunately lacks mathematical rigor in some parts (see, e.g., [24]–[26]). However, due to its success both in physics and engineering, it is generally agreed to be at least a valuable starting point for analysis of problems that seem otherwise too difficult to handle. A cursory overview of literature about the RM inside a specific field or topic may paint the picture that the RM is a fixed set of mathematical methods which can be applied to any suitable problem at hand. This is not entirely accurate and conceptually the RM can be seen more like a systematic way of turning a very difficult problem into a more manageable one than a set of specific tools that actually solve the problem. Indeed, the mathematical methods that are used at different stages of the RM can often be chosen from a variety of choices, although it is very common to have some form of *large deviations theory* as part of the analysis (see Step 2 below). Thus, instead of trying to be entirely general, we describe next (one form of) the steps taken in the RM in the context of the first example above.

[4] In the following we refrain differentiating random variables and their realizations for notational convenience. Also, Z and, as a result, f can depend on some parameters (non-random variables) that are not explicitly stated.

Step 1 (Replica Trick): Consider (65) and write it as

$$\begin{aligned} f &= -\frac{1}{M}\lim_{u\to 0^+}\frac{\partial}{\partial u}\ln \mathsf{E}_V\left\{[Z(V)]^u\right\} \\ &= -\frac{1}{M}\lim_{u\to 0^+}\frac{\partial}{\partial u}\ln \mathsf{E}_V\left\{\left(\sum_{\boldsymbol{x}\in\mathcal{A}^M} p_X(\boldsymbol{x})Z(V,\boldsymbol{x})\right)^u\right\} \\ &= -\frac{1}{M}\lim_{u\to 0^+}\frac{\partial}{\partial u}\ln \Xi(u), \end{aligned} \tag{66}$$

where $u\in\mathbb{R}$ and we denoted $\Xi(u)=\mathsf{E}_V\{[Z(V)]^u\}$. Then, *assume* that we can treat u as an integer when we take the expectation, namely,

$$\begin{aligned} \Xi(u) &= \mathsf{E}_V\left\{\prod_{a=1}^{u}\sum_{\boldsymbol{x}_a\in\mathcal{A}^N} p_X(\boldsymbol{x}_a)Z(V,\boldsymbol{x}_a)\right\} \\ &= \frac{1}{\pi^{uN}(\det \boldsymbol{R}_{\boldsymbol{w}})^u} \\ &\quad\times \mathsf{E}_V\left\{\sum_{\{\boldsymbol{x}_a\}}\prod_{a=1}^{u}\left[\mathrm{e}^{-(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x}_a)^{\mathsf{H}}\boldsymbol{R}_{\boldsymbol{w}}^{-1}(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x}_a)}p_X(\boldsymbol{x}_a)\right]\right\}, \end{aligned} \tag{67}$$

where the summation in the last expression is over the set $\{\boldsymbol{x}_a\}_{a=1}^u$. After taking the expectations, if we manage to write (67) in a form that does not explicitly force u to be an integer, *invoke analytical continuity* to extend u to real numbers.

The step above is at the very heart of the RM. It is important to realize that the equalities in (66) are provably true if differentiation under the integral sign is permitted and $u\in\mathbb{R}$. The part lacking rigorous mathematical justification is (67), especially when combined with the next two steps. Somewhat surprisingly, however, the end results of RM can sometimes be proved to be exact. Examples of such cases are: MIMO channel with Gaussian inputs, random energy model (REM) and Sherrington-Kirkpatrick model of spin glasses (see, e.g., [24]–[27], and references therein).

Step 2 (Large System Limit): Let the system approach the LSL, that is, the dimensions of the channel matrix $\boldsymbol{H}$ grow without bound at a finite and fixed ratio $\alpha=M/N>0$. Furthermore, assume that the limits w.r.t. u and M commute, so that we can first calculate the expectations in (67) in the LSL and then let $u\to 0$, as in (66).

The LSL assumption is natural in equilibrium statistical mechanics (e.g., the second example above), where the systems contain usually very large numbers of interacting particles M. In communication theory, the equivalent would be, e.g., a MIMO systems with large antenna arrays or a CDMA with large number of simultaneous users. It is in fact quite common to write the LSL assumption directly as a part of the replica trick in (66). The steps are separated here since the replica trick could also be used for finite sized systems. Due to mathematical difficulty of such cases, however, both steps are usually found together. The assumption of commuting limits is typically postulated *a priori* and rigorous justification of this step is beyond the scope of the present paper.

Let us denote the true transmitted vector $\boldsymbol{x}_0$, so that $\boldsymbol{y}=\boldsymbol{H}\boldsymbol{x}_0+\boldsymbol{w}$ is the generating model for the observation $\boldsymbol{y}$ and we can equivalently write $V=\{\boldsymbol{w},\boldsymbol{x}_0,\boldsymbol{H}\}$. Returning then to Step 1, we note that although the *replicated* vectors $\{\boldsymbol{x}_a\}_{a=1}^u$ act as IID RVs drawn according to p_X in (67) when conditioned on V, they can be correlated if not conditioned on V. We examine this through the empirical correlations between the vectors in the set $X_{u+1}=\{\boldsymbol{x}_a\}_{a=0}^u$ using *overlap matrix* $\boldsymbol{Q}\in\mathbb{C}^{(u+1)\times(u+1)}$, whose (a,b)th element[5] is given by $Q_{a,b}=M^{-1}\boldsymbol{x}_b^{\mathsf{H}}\boldsymbol{x}_a$. Then, the structure that is imposed on $\boldsymbol{Q}$ divides the replica analysis into two rough categories as described below.

Step 3 (Replica Symmetry): The *RS ansatz* or *RS assumption* means that the indexes $a=1,\ldots,u$ are permutation symmetric and $\boldsymbol{Q}$ can be written in terms of four parameters, for example, $Q_{0,0}=p$, $Q_{0,a}=m, a\geq 1$, $Q_{a,a}=q, a\geq 1$, and $Q_{a,b}=q, a\neq b\geq 1$. Note that $\boldsymbol{Q}=\boldsymbol{Q}^{\mathsf{H}}$ by construction. If $\boldsymbol{Q}$ is not of the RS form, it is said to have replica symmetry breaking (RSB) structure whose analysis is much more involved [24], [25].

The importance of the RS assumption will become clear when we present a rough sketch of the analysis of an ideal MIMO channel. We also note that the overlap matrix given in Step 3 allows the "zeroth" index to be treated separately to take into account the possibility that either $\boldsymbol{x}_0$ has different distribution than $\boldsymbol{x}_a$ when $a\geq 1$, or the decoder uses mismatched statistics, i.e., $Z(V,\boldsymbol{x}_a)$ does not match the probability law of the observation $\boldsymbol{y}=\boldsymbol{H}\boldsymbol{x}_0+\boldsymbol{w}$ as in Appendix C. For the simplified case considered below, however, we have $p=Q$ and $m=q$ since the indexes $a=0,1,\ldots,u$ can be treated on equal footing and two parameters are sufficient to define the RS form of $\boldsymbol{Q}$.

Next we give a brief and informal example of replica analysis for an ideal MIMO channel. The reader may be surprised to find out that most of the discussion below deals with details about how to obtain the necessary formulas when we follow the three stages above and not about those stages per se.

B. Average Over the Channel and Noise

The starting point of our replica calculation is (67), where we use the generating model of $\boldsymbol{y}$ to write in the exponential $\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x}_a=\boldsymbol{w}-\boldsymbol{H}(\boldsymbol{x}_0-\boldsymbol{x}_a)$. The first task is then to compute the expectation w.r.t. $\boldsymbol{w}$ and $\boldsymbol{H}$ for a fixed set $X_{u+1}=\{\boldsymbol{x}_a\}_{a=0}^u$ that satisfies the correlations of the RS overlap matrix $\boldsymbol{Q}$. Note that we cannot assume anymore that the vectors in X_{u+1} are independent since we changed the order of expectations in (67) and the average over X_{u+1} is carried out (later) without conditioning on $\boldsymbol{w}$ and $\boldsymbol{H}$. With this in mind, it follows that given X_{u+1}, the set $\{\boldsymbol{H}\boldsymbol{x}_a\}$ consists of CSCG RVs with correlations $\mathsf{E}_{\boldsymbol{H}}\{(\boldsymbol{H}\boldsymbol{x}_a)(\boldsymbol{H}\boldsymbol{x}_b)^{\mathsf{H}}\}=M^{-1}\boldsymbol{x}_b^{\mathsf{H}}\boldsymbol{x}_a\boldsymbol{I}_N=Q_{a,b}\boldsymbol{I}_N$ that are deterministic in the LSL. Thus, we can replace $\{\boldsymbol{H}(\boldsymbol{x}_0-\boldsymbol{x}_a)\}_{a=1}^u$ by a set of CSCG RVs $\{\boldsymbol{\Delta}_a\}_{a=1}^u$ and use Gaussian integration (62) to average over both $\boldsymbol{w}$ and $\{\boldsymbol{\Delta}_a\}_{a=1}^u$ to obtain (for more details, see Appendix C-B).

$$\Xi(u) = \int \mathrm{e}^{NG^{(u)}(\boldsymbol{Q})}\mu(\boldsymbol{Q})\mathrm{d}\boldsymbol{Q}, \tag{68}$$

$$\begin{aligned} G^{(u)}(\boldsymbol{Q}) = &-u\ln\det\left[\boldsymbol{R}_{\boldsymbol{w}}+(Q-q)\boldsymbol{I}_N\right] \\ &-u\ln\pi-\ln(u+1), \end{aligned} \tag{69}$$

[5]The row/column indexes of $\boldsymbol{Q}$ are $0,1,\ldots,u$ so that the correlations are measured also w.r.t. the true transmitted vector $\boldsymbol{x}_0$. Furthermore, due to (2), the empirical correlations can be expected to converge to the true ones in the LSL postulated in Step 2.

where $\boldsymbol{Q}$ should be understood to be in its RS parametrized form and $\mu(\boldsymbol{Q})$ is the PDF of the overlap matrix $\boldsymbol{Q}$.

Remark 2: Firstly, note that due to the RS assumption (Step 3), the function (69) is of a form that does not restrict u to be an integer, as desired. This is one of the reasons why we need to express matrix $\boldsymbol{Q}$ in a parametrized way instead of using it "as-is." Secondly, there is some universality in this derivation and the form (68) is a typical result of replica analysis. In some cases, however, different techniques are needed. One example is non-IID "mixing matrix" that requires direct matrix integration [39], [40].

C. Distribution of the Overlap Matrix and Large Deviations

The second major step in the analysis is to find an explicit formula for $\mu(\boldsymbol{Q})$, i.e., for the probability weight of the set $\{\boldsymbol{x}_a\}_{a=0}^{u}$ that satisfies $Q_{a,b} = M^{-1}\boldsymbol{x}_b^{\mathsf{H}}\boldsymbol{x}_a$. The form of (68) suggest that we should try to represent $\mu(\boldsymbol{Q})$ as an exponential whose argument is linear in N (or M) so that we can employ Laplace's method or the method of steepest descent to evaluate the integral w.r.t. $\boldsymbol{Q}$. If $\boldsymbol{x}_a \in \mathbb{R}^M$, due to (2), the elements of $\boldsymbol{x}_a$ are IID for all $a = 0, 1, \ldots, u$ and μ follows the large deviation principle [25], [41]. Informally this implies[6] $\mu(\boldsymbol{Q}) \asymp \mathrm{e}^{-Mc^{(u)}(\boldsymbol{Q})}$, where the *rate function*

$$c^{(u)}(\boldsymbol{Q}) = \sup_{\tilde{\boldsymbol{Q}}} \left\{ \mathrm{tr}(\boldsymbol{Q}\tilde{\boldsymbol{Q}}) - \lim_{M\to\infty} \frac{1}{M} \ln \phi^{(u)}(\tilde{\boldsymbol{Q}}) \right\}, \quad (70)$$

describes the exponential behavior of the probability,

$$\phi^{(u)}(\tilde{\boldsymbol{Q}}) = \mathsf{E}_{X_{u+1}} \left\{ \exp\left(\sum_{a,b=0}^{u} \tilde{Q}_{a,b} \boldsymbol{x}_b^{\mathsf{H}} \boldsymbol{x}_a \right) \right\}, \quad (71)$$

is the moment generating function (MGF) associated with $\mu(Q)$ and the supremum is over all $(u+1)\times(u+1)$ matrices $\tilde{\boldsymbol{Q}}$ that have the same RS form as Q, that is, $\tilde{Q}_{0,0} = \tilde{p}$, $\tilde{Q}_{0,a} = \tilde{m}, a \geq 1$, $\tilde{Q}_{a,a} = \tilde{Q}, a \geq 1$, and $\tilde{Q}_{a,b} = \tilde{q}, a \neq b \geq 1$. Thus, we can assess (68) in the LSL up to the leading order by using the exponential form of μ and Laplace's method, namely,

$$\begin{aligned} \Xi(u) &\asymp \int \mathrm{e}^{M\alpha^{-1}G^{(u)}(\boldsymbol{Q})} \mathrm{e}^{-Mc^{(u)}(\boldsymbol{Q})} \mathrm{d}\boldsymbol{Q} \\ &= \int \exp\left(N \left[\alpha^{-1} G^{(u)}(\boldsymbol{Q}) - c^{(u)}(\boldsymbol{Q}) \right] \right) \mathrm{d}\boldsymbol{Q} \quad (72) \\ &\asymp \exp\left(M \sup_{\boldsymbol{Q},\tilde{\boldsymbol{Q}}} \left\{ T^{(u)}(\boldsymbol{Q}, \tilde{\boldsymbol{Q}}) \right\} \right), \quad (73) \end{aligned}$$

where we denoted for notational convenience

$$T^{(u)}(\boldsymbol{Q}, \tilde{\boldsymbol{Q}}) = \frac{1}{\alpha} G^{(u)}(\boldsymbol{Q}) - \mathrm{tr}(\boldsymbol{Q}\tilde{\boldsymbol{Q}}) + \lim_{M\to\infty} \frac{1}{M} \ln \phi^{(u)}(\tilde{\boldsymbol{Q}}). \quad (74)$$

For complex vectors $\{\boldsymbol{x}_a\}$, the end result is essentially the same and the solution to the supremum is found among the critical points of the argument (see e.g., [29], [31], [39]). The large deviations analysis also guarantees that $\tilde{\boldsymbol{Q}}$ is in general a real symmetric matrix and if $(\boldsymbol{Q}^*, \tilde{\boldsymbol{Q}}^*)$ is the solution of the optimization problem in (73) then $T^{(u)}(\boldsymbol{Q}^*, \tilde{\boldsymbol{Q}}^*) \in \mathbb{R}$, as expected since f is in our case real.

However, in RM there is some ambiguity as to whether the correct point in the saddle-point approximation (73) minimizes or maximizes the exponential when we let $u \to 0$ [24], [25]. Thus, in RM, we seek in practice the critical points and (66) is thus of the form

$$f = -\lim_{u\to 0^+} \frac{\partial}{\partial u} \operatorname*{extr}_{\boldsymbol{Q},\tilde{\boldsymbol{Q}}} \left\{ T^{(u)}(\boldsymbol{Q}, \tilde{\boldsymbol{Q}}) \right\}, \quad (75)$$

where $\mathrm{extr}_X\{h(X)\}$ denotes finding the critical points of a function $h(X)$.

D. Decoupled MGF and Critical Points

The second part of replica analysis where the RS assumption plays an important role (for the first one, see Remark 2) is when we try to solve (71) and find the critical points of $T^{(u)}(\boldsymbol{Q}, \tilde{\boldsymbol{Q}})$. For the simplified setup in this section where $\boldsymbol{Q}$ and $\tilde{\boldsymbol{Q}}$ are represented with parameter $\{Q, q\}$ and $\{\tilde{Q}, \tilde{q}\}$, respectively, the MGF can be expressed as (see, e.g., [27] for details)

$$\phi^{(u)}(\tilde{\boldsymbol{Q}}) = \prod_{m=1}^{M} \left[\left(\frac{\tilde{q}}{\pi} \right)^{-u} \int \left[\mathsf{E}_{x_m} g(z_m | x_m; \tilde{q}^{-1}) \right]^{u+1} \mathrm{d}z_m \right], \quad (76)$$

where z_m are just dummy variables. On the other hand, finding the critical points involves taking eight partial derivatives for the RS case in *Step 3* (for the simplified case here, four is enough). Then, one should pick the solution that satisfies the conditions at the critical point while providing the global extremum of (66).In the case considered here, we can actually get rid of two parameters since $p = M^{-1}\mathsf{E}\|\boldsymbol{x}\|^2$ and $\tilde{p} = 0$ always at the critical point. Note that if we did not parametrize $\boldsymbol{Q}$, the critical points would be described by $u(u+1)$ equations and $\Xi(u)$ would depend explicitly on the fact that u is an integer. This is one of the reasons why even the full-RSB solution (see [24], [25]) uses a round-about way of presenting $\boldsymbol{Q}$ instead of using it "as-is."

Finally, we remark that it is quite common (see, e.g., [27]) to represent the end result in terms of new variables. For example, if we have equal transmit powers for each antennas $\bar{\gamma} = \bar{\gamma}_m$ in the simplified case considered here, then the parameters $\eta = \tilde{q}$ and $\varepsilon = Q - q = \bar{\gamma} - q$ fully describe the RS matrices $\boldsymbol{Q}$ and $\tilde{\boldsymbol{Q}}$ at the critical point. The former variable is inverse noise variance of a decoupled Gaussian channel

$$z = x + w, \qquad p(w) = g(w|0; \eta^{-1}), \quad (77)$$

and the latter variable ε is the MMSE of this channel when the inputs are drawn according to $p_X(x)$. The rest of RM is straightforward, albeit tedious algebra to arrive at (66).

[6] We use notation $a_M \asymp b_M$ to denote "equality up to the leading exponential order," that is $\lim_{M\to\infty} M^{-1} \ln(a_M / b_M) = 0$.

APPENDIX C
REPLICA ANALYSIS FOR MISMATCHED CASE

The analysis herein follows the main steps of RM as listed in Appendix B. Reader who is not familiar with the RM is encouraged to use discussion there as a guide to the derivations below.

A. Replica Trick

Let us consider the function $f(s)$ (free-energy) defined in (13). We then postulate that it can be expressed in the LSL using the standard replica trick (cf. Appendix B)

$$f(s) = -\lim_{M\to\infty}\frac{1}{M}\lim_{u\to 0}\frac{\partial}{\partial u}\ln \Xi^{(u,M)}(s), \tag{78}$$

where we defined for later convenience

$$\Xi^{(u,M)}(s) = \mathsf{E}\left\{\prod_{a=1}^{u} \mathrm{e}^{-[\boldsymbol{w}+\boldsymbol{H}(\boldsymbol{\chi}_0-\boldsymbol{\chi}_a)]^{\mathsf{H}}\boldsymbol{\Sigma}^{-1}[\boldsymbol{w}+\boldsymbol{H}(\boldsymbol{\chi}_0-\boldsymbol{\chi}_a)]}\right\}, \tag{79}$$

and denoted[7] $\boldsymbol{\Sigma} = s^{-1}\tilde{\boldsymbol{R}}$ along with $\boldsymbol{\chi}_0 = \boldsymbol{x}_0 + \boldsymbol{v}_0$ and $\boldsymbol{\chi}_a = \boldsymbol{x}_a, a = 1,\ldots,u$. Here $\boldsymbol{x}_0$ is the original transmit vector in (1) and $\{\boldsymbol{x}_a\}_{a=1}^{u}$ are replicated data vectors, which are IID drawn according to $p(\boldsymbol{x})$ when conditioned on $\{\boldsymbol{x}_0, \boldsymbol{v}_0, \boldsymbol{w}, \boldsymbol{H}\}$. On the other hand, $\boldsymbol{v}_0$ represents the noise plus distortion component at the transmit-sidethat is CSCG with covariance matrix $\boldsymbol{R}_{\boldsymbol{v}}$. Starting with (79), the goal is then to obtain a functional expression for $\Xi^{(u,M)}(s)$ in the LSL that does not enforce u to be an integer and then use (78) to obtain the desired quantity. In the following, explicit limit notations are often omitted for notational convenience.

B. Average Over the Channel and Noise

To proceed with the evaluation of (79), we first make the RS assumption

$$p = M^{-1}\|\boldsymbol{\chi}_0\|^2, \tag{80}$$
$$m = M^{-1}\boldsymbol{\chi}_0^{\mathsf{H}}\boldsymbol{\chi}_a, \qquad a = 1,\ldots,u, \tag{81}$$
$$Q = M^{-1}\|\boldsymbol{\chi}_a\|^2, \qquad a = 1,\ldots,u, \tag{82}$$
$$q = M^{-1}\boldsymbol{\chi}_a^{\mathsf{H}}\boldsymbol{\chi}_b, \qquad a \neq b \in \{1,\ldots,u\}. \tag{83}$$

and remind the reader that if we average first over $\boldsymbol{H}$, the empirical correlations between $\{\boldsymbol{x}_a\}_{a=0}^{u}$ are not zero in general as discussed in Appendix B. Thus, noticing that

$$\mathsf{E}_{\boldsymbol{H}}\left\{[\boldsymbol{H}(\boldsymbol{\chi}_0-\boldsymbol{\chi}_a)]\,[\boldsymbol{H}(\boldsymbol{\chi}_0-\boldsymbol{\chi}_b)]^{\mathsf{H}}\right\} = \begin{cases}[p-(m+m^*)+Q]\,\boldsymbol{I}_N, & a=b,\\ [p-(m+m^*)+q]\,\boldsymbol{I}_N, & a\neq b,\end{cases} \tag{84}$$

we may replace $\{\boldsymbol{H}(\boldsymbol{\chi}_0-\boldsymbol{\chi}_a)\}_{a=1}^{u}$ in (79) in the LSL by CSCG vectors $\{\boldsymbol{\Delta}_a\}_{a=1}^{u}$ that are constructed as

$$\boldsymbol{\Delta}_a = \boldsymbol{d}_a\sqrt{Q-q} + \boldsymbol{t}\sqrt{p-(m+m^*)+q} \tag{85}$$
$$= \boldsymbol{d}_a\sqrt{A} + \boldsymbol{t}\sqrt{B}, \tag{86}$$

where $\{\boldsymbol{t}, \{\boldsymbol{d}_a\}_{a=1}^{u}\}$ are IID standard complex Gaussian RVs independent of $\boldsymbol{w}$. Plugging (86) into $\Xi^{(u,M)}(s)$ and recalling that $\boldsymbol{\Sigma}$ is a fixed predefined matrix gives

$$\Xi^{(u,M)}(s) = \frac{1}{\det(\boldsymbol{R}_{\boldsymbol{w}})}\mathsf{E}\int\frac{\mathrm{d}\boldsymbol{w}}{\pi^N}\mathrm{e}^{-\boldsymbol{w}^{\mathsf{H}}\left(\boldsymbol{R}_{\boldsymbol{w}}^{-1}+u\boldsymbol{\Sigma}^{-1}\right)\boldsymbol{w}} \times\int\frac{\mathrm{d}\boldsymbol{t}}{\pi^N}\mathrm{e}^{-\boldsymbol{t}^{\mathsf{H}}(\boldsymbol{I}+uB\boldsymbol{\Sigma}^{-1})\boldsymbol{t}-2\Re\left\{\boldsymbol{w}^{\mathsf{H}}(u\sqrt{B}\boldsymbol{\Sigma}^{-1})\boldsymbol{t}\right\}} \times\left[\int\mathrm{e}^{-\boldsymbol{d}^{\mathsf{H}}(\boldsymbol{I}+A\boldsymbol{\Sigma}^{-1})\boldsymbol{d}+2\Re\left\{\left[-\sqrt{A}\boldsymbol{\Sigma}^{-1}(\boldsymbol{w}+\sqrt{B}\boldsymbol{t})\right]^{\mathsf{H}}\boldsymbol{d}\right\}}\frac{\mathrm{d}\boldsymbol{d}}{\pi^N}\right]^{u}. \tag{87}$$

Next, Gaussian integration (62) is applied on the integral w.r.t. $\boldsymbol{d}$. Using also (63) we arrive at

$$\Xi^{(u,M)}(s) = \mathsf{E}\int\frac{\mathrm{d}\boldsymbol{w}}{\pi^N}\frac{\mathrm{e}^{-\boldsymbol{w}^{\mathsf{H}}\left(\boldsymbol{R}_{\boldsymbol{w}}^{-1}+u(A\boldsymbol{I}_N+\boldsymbol{\Sigma})^{-1}\right)\boldsymbol{w}}}{\left[\det(\boldsymbol{I}+A\boldsymbol{\Sigma}^{-1})\right]^{u}\det(\boldsymbol{R}_{\boldsymbol{w}})} \times\int\mathrm{e}^{-\boldsymbol{t}^{\mathsf{H}}\left[\boldsymbol{I}_N+uB(A\boldsymbol{I}_N+\boldsymbol{\Sigma})^{-1}\right]\boldsymbol{t}+2\Re\left\{\left[-u\sqrt{B}(A\boldsymbol{I}_N+\boldsymbol{\Sigma})^{-1}\boldsymbol{w}\right]^{\mathsf{H}}\boldsymbol{t}\right\}}\frac{\mathrm{d}\boldsymbol{t}}{\pi^N}. \tag{88}$$

Application of (62) and (63) again for the integral w.r.t. $\boldsymbol{t}$ provides

$$\begin{aligned}\Xi^{(u,M)}(s) &= \mathsf{E}\left\{\frac{\left[\det(\boldsymbol{I}+A\boldsymbol{\Sigma}^{-1})\right]^{-u}}{\det\left[\boldsymbol{I}_N+uB(A\boldsymbol{I}_N+\boldsymbol{\Sigma})^{-1}\right]\det(\boldsymbol{R}_{\boldsymbol{w}})} \times\int\mathrm{e}^{-\boldsymbol{w}^{\mathsf{H}}\left(\boldsymbol{R}_{\boldsymbol{w}}^{-1}+u[(A+uB)\boldsymbol{I}_N+\boldsymbol{\Sigma}]^{-1}\right)\boldsymbol{w}}\frac{\mathrm{d}\boldsymbol{w}}{\pi^N}\right\}\\ &= \mathsf{E}\left\{\frac{\left(\det\left[\boldsymbol{I}_N+u\boldsymbol{R}_{\boldsymbol{w}}\left((A+uB)\boldsymbol{I}_N+\boldsymbol{\Sigma}\right)^{-1}\right]\right)^{-1}}{\det\left[\boldsymbol{I}_N+uB(A\boldsymbol{I}_N+\boldsymbol{\Sigma})^{-1}\right]\left[\det(\boldsymbol{I}+A\boldsymbol{\Sigma}^{-1})\right]^{u}}\right\},\end{aligned} \tag{89}$$

where the second line is also obtained through Gaussian integration. The above holds for any $\boldsymbol{R}_{\boldsymbol{w}}$ and $\boldsymbol{\Sigma}$ that are Hermitian and invertible. The determinants in (89) can be further simplified using (64), so that recalling $\boldsymbol{\Sigma} = s^{-1}\tilde{\boldsymbol{R}}$ and defining two auxiliary matrices

$$\boldsymbol{\Omega}(p,m,q) = \boldsymbol{R}_{\boldsymbol{w}} + (p-(m+m^*)+q)\,\boldsymbol{I}_N, \tag{90}$$
$$\tilde{\boldsymbol{\Omega}}(Q,q) = s^{-1}\tilde{\boldsymbol{R}} + (Q-q)\boldsymbol{I}_N, \tag{91}$$

that are both Hermitian, we finally have

$$\Xi^{(u,M)}(s) = \det(s^{-1}\tilde{\boldsymbol{R}})^{u}\mathsf{E}\left\{\mathrm{e}^{G^{(u)}(p,m,q,Q)}\right\}, \tag{92}$$

$$G^{(u)}(p,m,q,Q) = (1-u)\ln\det\tilde{\boldsymbol{\Omega}}(Q,q) - \ln\det\left[\tilde{\boldsymbol{\Omega}}(Q,q)+u\boldsymbol{\Omega}(p,m,q)\right], \tag{93}$$

[7] We remind the reader that for the case of mismatched decoding, the postulated covariance matrix $\tilde{\boldsymbol{R}}$ is fixed by definition so that $\boldsymbol{\Sigma} = s^{-1}\tilde{\boldsymbol{R}}$ is also a fixed predefined matrix. This is in contrast to the case of matched decoding (5), where the effective covariance matrix $\boldsymbol{R}_{\boldsymbol{w}} + \boldsymbol{H}\boldsymbol{R}_{\boldsymbol{v}}\boldsymbol{H}$ is random and depends directly on the channel matrix $\boldsymbol{H}$.

Using the differentiation rule $\frac{\partial}{\partial x}\ln\det\boldsymbol{A}=\mathrm{tr}(\boldsymbol{A}^{-1}\frac{\partial \boldsymbol{A}}{\partial x})$, where the partial derivative should be understood as an elementwise operation on $\boldsymbol{A}$, we also obtain for later use the equalities

$$\frac{\partial}{\partial p}G^{(u)}(\boldsymbol{Q}) = -u\mathrm{tr}\left((\tilde{\boldsymbol{\Omega}}+u\boldsymbol{\Omega})^{-1}\right), \tag{94}$$

$$\frac{\partial}{\partial m}G^{(u)}(\boldsymbol{Q}) = \frac{\partial}{\partial m^*}G^{(u)}(\boldsymbol{Q}) = u\mathrm{tr}\left((\tilde{\boldsymbol{\Omega}}+u\boldsymbol{\Omega})^{-1}\right), \tag{95}$$

$$\frac{\partial}{\partial q}G^{(u)}(\boldsymbol{Q}) = u(u-1)\mathrm{tr}\left(\tilde{\boldsymbol{\Omega}}^{-1}\boldsymbol{\Omega}(\tilde{\boldsymbol{\Omega}}+u\boldsymbol{\Omega})^{-1}\right), \tag{96}$$

$$\frac{\partial}{\partial Q}G^{(u)}(\boldsymbol{Q}) = u\mathrm{tr}\left(\tilde{\boldsymbol{\Omega}}^{-1}\boldsymbol{\Omega}(\tilde{\boldsymbol{\Omega}}+u\boldsymbol{\Omega})^{-1}\right) - u\mathrm{tr}(\tilde{\boldsymbol{\Omega}}^{-1}), \tag{97}$$

where the dependencies to $\{p,m,q,Q\}$ were omitted on the RHSs of the equations for notational implicity.

C. Distribution of the Overlap Matrix and Large Deviations

Let us now write the general form of empirical correlations between $\{\boldsymbol{\Delta}_a\}$ as

$$\begin{aligned}\frac{1}{M}\mathsf{E}_{\boldsymbol{H}}\left\{\boldsymbol{\Delta}_b^{\mathsf{H}}\boldsymbol{\Delta}_a\right\} &= \left(\frac{\|\boldsymbol{\chi}_0\|^2}{M}-\frac{\boldsymbol{\chi}_b^{\mathsf{H}}\boldsymbol{\chi}_0}{M}-\frac{\boldsymbol{\chi}_0^{\mathsf{H}}\boldsymbol{\chi}_a}{M}+\frac{\boldsymbol{\chi}_b^{\mathsf{H}}\boldsymbol{\chi}_a}{M}\right)\\ &= (Q_{0,0}-Q_{0,b}-Q_{a,0}+Q_{a,b}),\end{aligned} \tag{98}$$

where $Q_{a,b}$ are the elements of the overlap matrix $\boldsymbol{Q}\in\mathbb{C}^{(u+1)\times(u+1)}$ and have the obvious definitions. We then need to find a suitable formula for the rate function (70). By the RS assumption,

$$\mathrm{tr}(\boldsymbol{Q}\tilde{\boldsymbol{Q}}) = p\tilde{p}+u\tilde{m}(m+m^*)+uQ\tilde{Q}+u(u-1)q\tilde{q}, \tag{99}$$

since $\tilde{\boldsymbol{Q}}$ is real symmetric and we may write (78) as in (100), shown at the bottom of the page, where the per-antenna rate function reads

$$\phi_m^{(u)}(\tilde{\boldsymbol{Q}}) = \mathsf{E}_{\{\chi_{a,m}\}}\left\{\exp\left[\sum_{a=0}^{u}\sum_{b=0}^{u}\tilde{Q}_{a,b}\chi_{b,m}^*\chi_{a,m}\right]\right\}, \tag{101}$$

and $\chi_a=[\chi_{a,1}\cdots\chi_{a,M}]^\top$.

D. Decoupled MGF and Critical Points

The first set of equations for the critical point arises from the equality

$$\frac{\partial}{\partial x}\mathrm{tr}(\boldsymbol{Q}\tilde{\boldsymbol{Q}}) = \frac{1}{M}\frac{\partial}{\partial x}G^{(u)}(\boldsymbol{Q}), \tag{102}$$

for $x\in\{p,m,q,Q\}$. The partial derivatives on the LHS are trivial due to (99) and the RHSs we already obtained in (94)–(97). If we drop the explicit dependence of $\boldsymbol{\Omega}$ and $\tilde{\boldsymbol{\Omega}}$ on $\{p,m,q,Q\}$ for notational simplicity, the RS conjugate parameters satisfy

$$\tilde{p} = -u\frac{1}{M}\mathrm{tr}\left[(\tilde{\boldsymbol{\Omega}}+u\boldsymbol{\Omega})^{-1}\right] = -u\tilde{m}, \tag{103}$$

$$\tilde{m} = \frac{1}{M}\mathrm{tr}\left[(\tilde{\boldsymbol{\Omega}}+u\boldsymbol{\Omega})^{-1}\right], \tag{104}$$

$$\tilde{q} = \frac{1}{M}\mathrm{tr}\left[\tilde{\boldsymbol{\Omega}}^{-1}\boldsymbol{\Omega}(\tilde{\boldsymbol{\Omega}}+u\boldsymbol{\Omega})^{-1}\right], \tag{105}$$

$$\tilde{Q} = \frac{1}{M}\mathrm{tr}\left[\tilde{\boldsymbol{\Omega}}^{-1}\boldsymbol{\Omega}(\tilde{\boldsymbol{\Omega}}+u\boldsymbol{\Omega})^{-1}\right] - \frac{1}{M}\mathrm{tr}\left(\tilde{\boldsymbol{\Omega}}^{-1}\right). \tag{106}$$

Note that the above implies that in the limit $u\to 0$, we have $\tilde{p}\to 0$, and $\tilde{m}\to -(\tilde{Q}-\tilde{q})$, so that the relevant critical point can be written by using two instead of four "tilde-parameters."

The next task is to obtain an explicit expression for the per-component moment generating function (MGF) in (101) that does not require u to be an integer. Since this part is closely similar to the analysis carried out, e.g., in [27] we omit the details of the derivations. Following the notation of [27], we let $\xi=\tilde{m}$ and $\eta=\tilde{m}^2/\tilde{q}$ which is sufficient to describe $\tilde{\boldsymbol{Q}}$ here. Then, if we denote $\chi_m=x_m+v_m$ and $\tilde{\chi}_m=\tilde{x}_m$, the scalar MGF (101) can be written as

$$\begin{aligned}\phi_m^{(u)}(\tilde{\boldsymbol{Q}}) = \left(\frac{\pi}{\xi}\right)^u \mathsf{E}\Bigg\{\int \mathrm{d}z_m \mathrm{e}^{u\xi\left(|z_m|^2-|\chi_m|^2\right)}\\ \times p(z_m|\chi_m)\left[\mathsf{E}_{\tilde{\chi}_m}q(z_m|\tilde{\chi}_m)\right]^u\Bigg\},\end{aligned} \tag{107}$$

where $p(z_m|\chi_m)=g(z_m|\chi_m;,\eta^{-1})$ and $q(z_m|\tilde{\chi}_m)=g(z_m|\tilde{\chi}_m;\xi^{-1})$. As a consequence of the above, u does not need to be an integer anymore and the limit $u\to 0$ is well defined. From the partial derivatives of $\{\tilde{p},\tilde{m},\tilde{q},\tilde{Q}\}$ we obtain the second set of conditions at the critical point

$$p = \lim_{M\to\infty}\frac{1}{M}\sum_{m=1}^{M}\mathsf{E}|x_m+v_m|^2, \tag{108}$$

$$Q = \lim_{M\to\infty}\frac{1}{M}\sum_{m=1}^{M}\mathsf{E}\left\langle|\tilde{x}_m|^2\right\rangle_q, \tag{109}$$

$$m = \lim_{M\to\infty}\frac{1}{M}\sum_{m=1}^{M}\mathsf{E}(x_m+v_m)\left\langle\tilde{x}_m^*\right\rangle_q, \tag{110}$$

$$q = \lim_{M\to\infty}\frac{1}{M}\sum_{m=1}^{M}\mathsf{E}\left\langle\tilde{x}_m^*\right\rangle_q\left\langle\tilde{x}_m^*\right\rangle_q, \tag{111}$$

where $x_m,\tilde{x}_m\sim p(x_m)$, $v_m\sim g(v_m|0;r_{\boldsymbol{v}}^m)$,

$$\left\langle f(\tilde{x}_m)\right\rangle_q = \mathsf{E}_{\tilde{x}_m}f(\tilde{x}_m)\frac{q(z_m|\tilde{x}_m)}{q(z_m)}, \tag{112}$$

$$\begin{aligned}f_{\mathsf{RS}} = -\lim_{M\to\infty}\frac{1}{M}\ln\det(\boldsymbol{\Sigma}) - \underset{\boldsymbol{Q},\tilde{\boldsymbol{Q}}}{\mathrm{extr}}\lim_{M\to\infty}\Bigg\{\frac{1}{M}\lim_{u\to 0}\frac{\partial}{\partial u}G^{(u)}(\boldsymbol{Q})\\ -\lim_{u\to 0}\frac{\partial}{\partial u}\left[p\tilde{p}+u(m\tilde{m}^*+\tilde{m}m^*)+uQ\tilde{Q}+u(u-1)q\tilde{q}\right]+\frac{1}{M}\sum_{m=1}^{M}\lim_{u\to 0}\frac{\partial}{\partial u}\ln\phi_m^{(u)}(\tilde{\boldsymbol{Q}})\Bigg\}\end{aligned} \tag{100}$$

and $q(z_m) = \mathsf{E}_{\tilde{x}_m} q(z_m|\tilde{x}_m)$. The interpretation is that (112) represents the conditional mean estimator for postulated channel $q(z_m|\tilde{\chi}_m)$ when the true channel is given by $p(z_m|\chi_m)$. Then the true $\varepsilon = p - (m + m^*) + q$, and postulated $\tilde{\varepsilon} = Q - q$ MMSE reduce to (23) and (24), respectively. Finally, computing the partial derivatives w.r.t. u in (100) and taking the limit $u \to 0$ provides after some algebra the free energy (26).

REFERENCES

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

[2] T. C. W. Schenk, P. F. M. Smulders, and E. R. Fledderus, "Performance of MIMO OFDM systems in fading channels with additive TX and RX impairments," in *Proc. 1st Annu. IEEE BENELUX/DSP Valley Signal Process. Symp.*, Apr. 2005, pp. 41–44.

[3] B. Göransson, S. Grant, E. Larsson, and Z. Feng, "Effect of transmitter and receiver impairments on the performance of MIMO in HSDPA," in *Proc. 9th IEEE Workshop Signal Process. Adv. Wireless Commun.*, Jul. 2008, pp. 496–500.

[4] H. Suzuki, T. V. Anh Tran, I. B. Collings, G. Daniels, and M. Hedley, "Transmitter noise effect on the performance of a MIMO-OFDM hardware implementation achieving improved coverage," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 6, pp. 867–876, Aug. 2008.

[5] H. Suzuki, I. B. Collings, M. Hedley, and G. Daniels, "Practical performance of MIMO-OFDM-LDPC with low complexity double iterative receiver," in *Proc. 20th IEEE Int. Symp. Pers., Indoor, Mobile Radio Commun.*, Sep. 2009, pp. 2469–2473.

[6] C. Studer, M. Wenk, and A. Burg, "MIMO transmission with residual transmit-RF impairments," in *Proc. Int. ITG Workshop Smart Antennas*, Feb. 2010, pp. 189–196.

[7] J. González-Coma, P. M. Castro, and L. Castedo, "Impact of transmit impairments on multiuser MIMO non-linear transceivers," in *Proc. Int. ITG Workshop Smart Antennas*, Feb. 2011, pp. 1–8.

[8] C. Studer, M. Wenk, and A. Burg, "System-level implications of residual transmit-RF impairments in MIMO systems," in *Proc. 5th Eur. Conf. Antennas Propag.*, Apr. 2011, pp. 2686–2689.

[9] J. González-Coma, P. M. Castro, and L. Castedo, "Transmit impairments influence on the performance of MIMO receivers and precoders," in *Proc. 11th Eur. Wireless Conf.*, Apr. 2011, pp. 1–8.

[10] E. Björnson, P. Zetterberg, M. Bengtsson, and B. Ottersten, "Capacity limits and multiplexing gains of MIMO channels with transceiver impairments," *IEEE Commun. Lett.*, vol. 17, no. 1, pp. 91–94, Jan. 2013.

[11] X. Zhang, M. Matthaiou, E. Björnson, M. Coldrey, and M. Debbah, "On the MIMO capacity with residual transceiver hardware impairments," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2014, pp. 5299–5305.

[12] G. Fettweis *et al.*, "Dirty RF: A new paradigm," *Int. J. Wireless Inf. Netw.*, vol. 14, no. 2, pp. 133–148, Jun. 2007.

[13] T. Schenk, *RF Imperfections in High-Rate Wireless Systems: Impact and Digital Compensation*. Berlin, Germany: Springer-Verlag, 2008.

[14] F. Gregorio, J. Cousseau, S. Werner, T. Riihonen, and R. Wichman, "EVM analysis for broadband OFDM direct-conversion transmitters," *IEEE Trans. Veh. Technol.*, vol. 62, no. 7, pp. 3443–3451, Sep. 2013.

[15] N. Merhav, G. Kaplan, A. Lapidoth, and S. Shamai, "On information rates for mismatched decoders," *IEEE Trans. Inf. Theory*, vol. 40, no. 6, pp. 1953–1967, Nov. 1994.

[16] A. Ganti, A. Lapidoth, and I. Telatar, "Mismatched decoding revisited: General alphabets, channels with memory, the wide-band limit," *IEEE Trans. Inf. Theory*, vol. 40, no. 6, pp. 1953–1967, Nov. 2000.

[17] A. Guillén i Fàbregas, A. Martinez, and G. Caire, "Bit-interleaved coded modulation," *Found. Trends Commun. Inf. Theory*, vol. 5, no. 1/2, pp. 1–153, Jan. 2008.

[18] W. Zhang, "A general framework for transmission with transceiver distortion and some applications," *IEEE Trans. Commun.*, vol. 60, no. 2, pp. 384–399, Feb. 2012.

[19] 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.

[20] H. Weingarten, Y. Steinberg, and S. Shamai, "Gaussian codes and weighted nearest neighbor decoding in fading multiple-antenna channels," *IEEE Trans. Inf. Theory*, vol. 50, no. 8, pp. 1665–1686, Aug. 2004.

[21] A. T. Asyhari and A. Guillén i Fàbregas, "Nearest neighbor decoding in MIMO block-fading channels with imperfect CSIR," *IEEE Trans. Inf. Theory*, vol. 58, no. 3, pp. 1483–1517, Mar. 2012.

[22] M. Médard, "The effect upon channel capacity in wireless communications of perfect and imperfect knowledge of the channel," *IEEE Trans. Inf. Theory*, vol. 46, no. 3, pp. 933–946, May 2000.

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

[24] H. Nishimori, *Statistical Physics of Spin Glasses and Information Processing*. New York, NY, USA: Oxford Univ. Press, 2001.

[25] M. Mézard and A. Montanari, *Information, Physics, Computation*. New York, NY, USA: Oxford Univ. Press, 2009.

[26] T. Tanaka, "A statistical-mechanics approach to large-system analysis of CDMA multiuser detectors," *IEEE Trans. Inf. Theory*, vol. 48, no. 11, pp. 2888–2910, Nov. 2002.

[27] D. Guo and S. Verdú, "Randomly spread CDMA: Asymptotics via statistical physics," *IEEE Trans. Inf. Theory*, vol. 51, no. 6, pp. 1983–2010, Jun. 2005.

[28] R. R. Müller, "Channel capacity and minimum probability of error in large dual antenna array systems with binary modulation," *IEEE Trans. Signal Process.*, vol. 51, no. 11, pp. 2821–2828, Nov. 2003.

[29] C.-K. Wen, K.-K. Wong, and J.-C. Chen, "Spatially correlated MIMO multiple-access systems with macrodiversity: Asymptotic analysis via statistical physics," *IEEE Trans. Commun.*, vol. 55, no. 3, pp. 477–488, Mar. 2007.

[30] C.-K. Wen, K.-K. Wong, and J.-C. Chen, "Asymptotic mutual information for Rician MIMO-MA channels with arbitrary inputs: A replica analysis," *IEEE Trans. Commun.*, vol. 58, no. 10, pp. 2782–2788, Oct. 2010.

[31] K. Takeuchi, R. R. Müller, M. Vehkaperä, and T. Tanaka, "On an achievable rate of large Rayleigh block-fading MIMO channels with no CSI," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6517–6541, Oct. 2013.

[32] A. M. Tulino and S. Verdú, "Random matrix theory and wireless communications," *Found. Trends Commun. Inf. Theory*, vol. 1, no. 1, pp. 1–182, Jan. 2004.

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

[34] C.-K. Wen and K.-K. Wong, "On the sum-rate of uplink MIMO cellular systems with amplify-and-forward relaying and collaborative base stations," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 9, pp. 1409–1424, Dec. 2010.

[35] M. A. Girnyk, M. Vehkaperä, and L. K. Rasmussen, "Large-system analysis of correlated MIMO multiple access channels with arbitrary signaling in the presence of interference," *IEEE Trans. Wireless Commun.*, vol. 13, no. 4, pp. 2060–2073, Apr. 2014.

[36] M. A. Girnyk, M. Vehkaperä, and L. K. Rasmussen, "Asymptotic performance analysis of a K-hop amplify-and-forward relay MIMO channel," arXiv:1410.5716.

[37] A. L. Moustakas and S. H. Simon, "On the outage capacity of correlated multiple-path MIMO channels," *IEEE Trans. Inf. Theory*, vol. 53, no. 11, pp. 3887–3903, Nov. 2007.

[38] D. S. Bernstein, *Matrix Mathematics: Theory, Facts, and Formulas*. Princeton, NJ, USA: Princeton Univ. Press, 2009.

[39] R. R. Müller, D. Guo, and A. Moustakas, "Vector precoding for wireless MIMO systems and its replica analysis," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 3, pp. 530–540, Apr. 2008.

[40] M. Vehkaperä, Y. Kabashima, and S. Chatterjee, "Analysis of regularized LS reconstruction and random matrix ensembles in compressed sensing," arXiv:1312.0256.

[41] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*. New York, NY, USA: Springer-Verlag, 1998.

Mikko Vehkaperä (S'03–M'10) received the M.Sc. and Lic.Sc. degrees from University of Oulu, Oulu, Finland, in 2004 and 2006, respectively, and the Ph.D. degree from Norwegian University of Science and Technology (NTNU), Trondheim, Norway, in 2010.

Between 2010 and 2013, he was a Postdoctoral Researcher at School of Electrical Engineering, and the ACCESS Linnaeus Center, KTH Royal Institute of Technology, Stockholm, Sweden. Since 2013, he has been an Academy of Finland Postdoctoral Researcher at Aalto University School of Electrical Engineering. He held visiting appointments at Massachusetts Institute of Technology (MIT), USA, Kyoto University and Tokyo Institute of Technology, Japan, and University of Erlangen-Nuremberg, Germany. His research interests are in the field of wireless communications and signal processing.

Dr. Vehkaperä was a co-recipient for the Best Student Paper award at IEEE International Conference on Networks (ICON2011) for the paper "Delay Constrained Throughput Analysis of CDMA Using Stochastic Network Calculus."

Taneli Riihonen (S'06–M'14) received the M.Sc. and D.Sc. degrees in communications and electrical engineering (both with distinction) from Helsinki University of Technology, Espoo, Finland, in February 2006 and from Aalto University, Helsinki, Finland in August 2014, respectively. During summer 2005, he was a Seasonal Trainee at Nokia Research Center, Helsinki, Finland and, since Fall 2005, he has held various research positions at the Department of Signal Processing and Acoustics, Aalto University School of Electrical Engineering, Helsinki, Finland. He is currently a Visiting Associate Research Scientist at Columbia University in the City of New York, USA. He is serving as an Editor of IEEE COMMUNICATIONS LETTERS since October 2014. His research activity is focused on physical-layer OFDM(A), multiantenna, relaying and full-duplex wireless techniques.

Maksym A. Girnyk (S'09) received the B.Sc. and M.Sc. degrees from the KPI National Technical University of Ukraine, Kyiv, Ukraine in 2005 and 2007, respectively, the M.Sc. degree from Supélec, Paris, France in 2008, as well as the Lic.Eng. and Ph.D. degrees from the KTH Royal Institute of Technology, Stockholm, Sweden in 2012 and 2014, respectively. He was a Visiting Researcher at the Alcatel-Lucent Chair on Flexible Radio at Supélec, Paris, France in 2013 and Aalto University, Aalto, Finland in 2014. His research interests include wireless communications, information theory, random matrix theory, and statistical physics.

Emil Björnson (S'07–M'12) received the M.S. degree in engineering mathematics from Lund University, Lund, Sweden, in 2007. He received the Ph.D. degree in telecommunications from the Department of Signal Processing at KTH Royal Institute of Technology, Stockholm, Sweden, in 2011. From 2012 to 2014, he was a joint postdoc at the Alcatel-Lucent Chair on Flexible Radio, Supélec, Paris, France, and at KTH Royal Institute of Technology. He is currently an Assistant Professor at the Division of Communication Systems at Linköping University, Sweden.

His research interests include multi-antenna cellular communications, massive MIMO techniques, radio resource allocation, green energy efficient systems, and network topology design. He is the first author of the book "Optimal Resource Allocation in Coordinated Multi-Cell Systems" from 2013. He is also dedicated to reproducible research and has made a large amount of simulation code publically available.

Dr. Björnson has received 4 best paper awards (as first author or coauthor) for novel research on optmization and design of multi-cell multi-antenna communications: WCNC 2014, SAM 2014, CAMSAP 2011, and WCSP 2009.

Mérouane Debbah (S'01–AM'03–M'04–SM'08–F'15) received the M.Sc. and Ph.D. degrees from the Ecole Normale Supérieure de Cachan, France. He worked for Motorola Labs (Saclay, France) from 1999 to 2002, and the Vienna Research Center for Telecommunications (Vienna, Austria) until 2003. From 2003 to 2007, he joined the Mobile Communications Department of the Institut Eurecom (Sophia Antipolis, France) as an Assistant Professor. Since 2007, he is a Full Professor at Supelec (Gif-sur-Yvette, France). From 2007 to 2014, he was Director of the Alcatel-Lucent Chair on Flexible Radio. Since 2014, he is Vice-President of the Huawei France R&D Center and Director of the Mathematical and Algorithmic Sciences Lab. His research interests are in information theory, signal processing, and wireless communications. He is an Associate Editor in Chief of the journal, *Random Matrix: Theory and Applications* and was an associate and senior area editor for IEEE TRANSACTIONS ON SIGNAL PROCESSING, respectively in 2011–2013 and 2013–2014. Dr. Debbah is a recipient of the ERC grant MORE (Advanced Mathematical Tools for Complex Network Engineering). He is a WWRF Fellow, and a member of the academic senate of Paris-Saclay. He is the recipient of the Mario Boella Award in 2005, the 2007 IEEE GLOBECOM Best Paper Award, the Wi-Opt 2009 Best Paper Award, the 2010 Newcom++ Best Paper, the WUN CogCom Best Paper 2012 and 2013 Award, the 2014 WCNC Best Paper Award as well as the Valuetools 2007, Valuetools 2008, CrownCom2009, Valuetools 2012 and SAM 2014 best student paper awards. In 2011, he received the IEEE Glavieux Prize Award and in 2012, the Qualcomm Innovation Prize Award.

Lars Kildehøj Rasmussen (S'92–M'93–SM'01) received the M.Eng. degree in 1989 from the Technical University of Denmark, Lyngby, Denmark, and the Ph.D. degree from Georgia Institute of Technology, Atlanta, Georgia, USA, in 1993. He is now a Professor in the Department of Communication Theory at the School of Electrical Engineering, and the ACCESS Linnaeus Center, at the KTH Royal Institute of Technology, Stockholm, Sweden.

He has prior experience from the Institute for Telecommunications Research at the University of South Australia (Adelaide, Australia), the Center for Wireless Communications at National University of Singapore (Singapore), Chalmers University of Technology (Gothenburg, Sweden), and University of Pretoria (Pretoria, South Africa). He is a Co-Founder of Cohda Wireless Pty., Ltd. (http://www.cohdawireless.com/); a Leading Developer of Safe Vehicle and Connected Vehicle design solutions. His research interests include transmission strategies and coding schemes for wireless communications, communications and control, and vehicular communication systems.

He is a member of the IEEE Information Theory Society, Communications Society, and Vehicular Technology Society. He served as Chairman for the Australian Chapter of the IEEE Information Theory Society 2004–2005, and has been a board member of the IEEE Sweden Section Vehicular Technology, Communications, and Information Theory Joint Societies Chapter since 2010. He is an Associate Editor for IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, and *Physical Communications*, as well as a Former Associate Editor of IEEE TRANSACTIONS ON COMMUNICATIONS, 2002–2013.

Risto Wichman received the M.Sc. and D.Sc. (Tech.) degrees in digital signal processing from Tampere University of Technology, Tampere, Finland, in 1990 and 1995, respectively. From 1995 to 2001, he worked at Nokia Research Center as a Senior Research Engineer. In 2002, he joined Department of Signal Processing and Acoustics, Aalto University School of Electrical Engineering, Aalto, Finland, where he is a Full Professor since 2008. His research interests include signal processing techniques for wireless communication systems.