Nearly Doubling the Throughput of Multiuser MIMO Systems Using Codebook Tailored Limited Feedback Protocol

G. Wunder, J. Schreck, P. Jung
Fraunhofer Heinrich Hertz Institute, Berlin

Abstract

We present a new robust feedback and transmit strategy for multiuser MIMO downlink communication systems, termed Rate Approximation (RA), and analyze its performance. The new scheme combines flexibility and robustness needed for reliable communications with the user terminal under a limited feedback constraint. It responds to two important observations: One is that it is not so significant to approximate the channel but rather the (potential) rate itself so as to mimic the optimal scheduling decision at the base station. The second observation is that a fixed transmit codebook at the transmitter is often better when simultaneously the channel state information is more accurate. Both observations are incorporated in the new scheme where the transmit and feedback codebook are strictly separated and user rates are delivered to the base station subject to a controlled uniform error regardless what the scheduling decision is. The scheme is analyzed and proved to have better performance below a certain interference plus noise margin and better scaling properties than the classical Jindal formula when considered in the very same setting. LTE system simulations sustain the analytic results showing performance gains of up to 50% compared to zeroforcing when using multiple antennas at both the base station and the terminal, and up to 70% when using single antennas at the terminal. Finally, a new feedback protocol is developed which inherently considers the transmit codebook and which is able to deal with the complexity issue at the terminal.

I. Introduction

Multiuser multiple input multiple output (MU–MIMO) communication systems have been in the focus of intensive research over many years. The optimal transmission technique for these systems is dirty paper coding (DPC) [1] which, under perfect channel state information at the transmitter (CSIT), achieves superior performance gains over linear schemes. However, in practical systems CSIT is obtained via a rate–constrained feedback channel, which is known to be a sensitive part of the overall system and must be carefully designed. An extensive overview of so–called limited feedback MU–MIMO systems can be found in [2]. References [3], [4] set the standard for performance evaluation in MU–MIMO systems.

In this paper we revisit the limited feedback problem in MU–MIMO systems. We consider linear beamforming techniques and assume that the transmit beamforming vectors are defined by a fixed transmit codebook. In contrast to previous work we use a different codebook for the feedback and apply a new feedback strategy which we call *Rate Approximation* (RA). The main idea is that the terminal selects a channel quantization vector from a feedback codebook considering any possible combination of beamforming vectors from the transmit codebook. As we will show, this enables the base station to approximate the user rates subject to a small uniform a priori error. Given the feedback message, the base station is then permitted to assert any beamforming vector from the transmit codebook for some optimization purpose (not just the beamforming vector dictated by the user if scheduled).

Let us provide a striking example: Consider a setup where the users are served on unitary beamforming vectors. It is well known that such a scheme achieves the optimal throughput for a large number of users [5]. However, it has been mostly overlooked yet that the scheme performs excellent for a finite number of users as well, in particular with limited feedback under the new RA strategy. This is illustrated in Figure 1. In the analytical section (see Corollary 1) we make exact this observations precise in this paper.

Organization and Main Results: In Section II we introduce the system model and in Section III the RA scheme is introduced. In Section IV we analyze the performance of the RA scheme. Our baseline is ZF and the operating point is such that the base station always serves as many (independent) users as transmit antennas. This model is essentially in accordance with [3] and enables us to make stringent comparisons. The analysis is performed in three steps:

1) We consider perfect CSIT and show that unitary beamforming (UB) can achieve a significant performance gain over ZF for a large fraction of a practically relevant signal–to–noise ratio (SNR) range.
2) We analyze the a priori rate error at the base station (i.e. before any scheduling decision) for each individual terminal evoked by our RA feedback strategy and prove that it is not

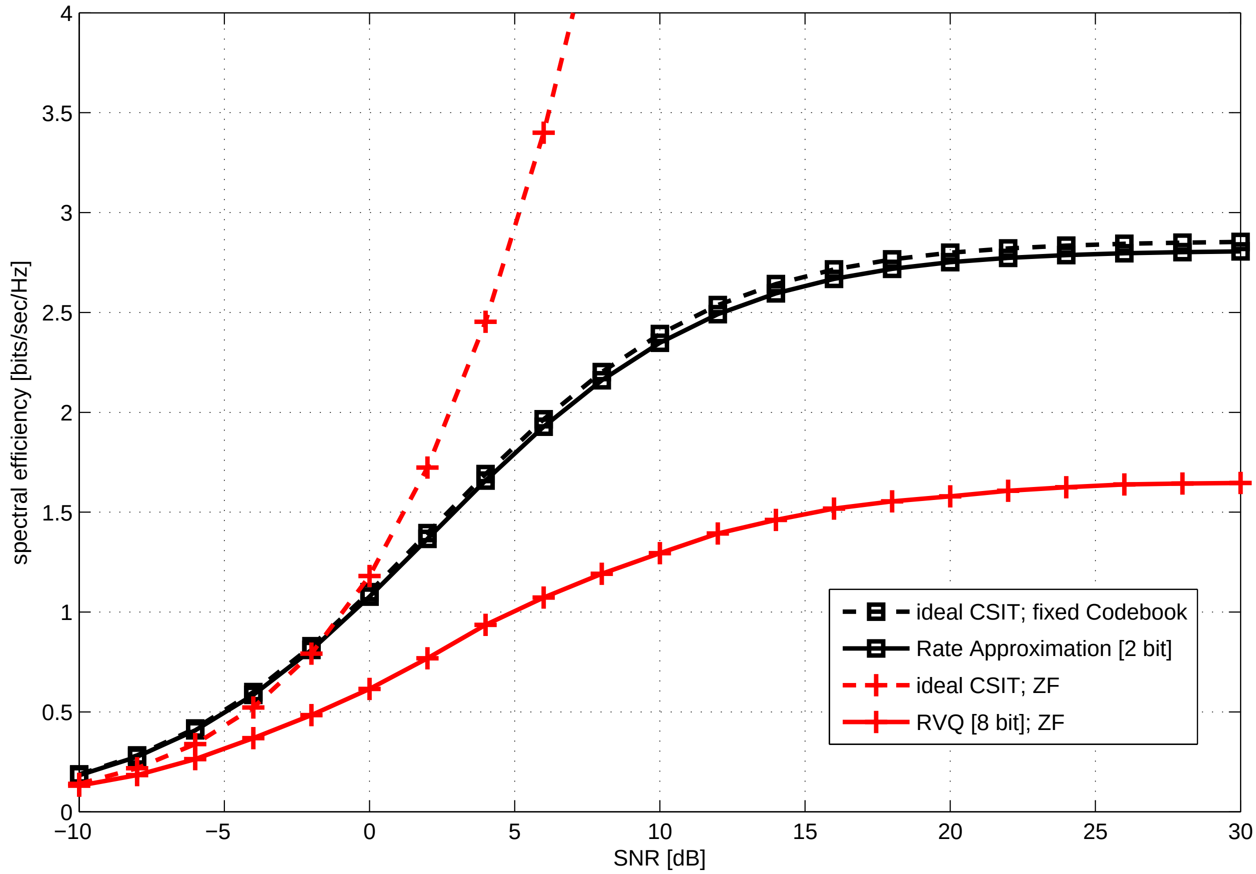

Fig. 1. Spectral efficency vs. average SNR: comparison of ZF with perfect CSIT and beamforming under a fixed codebook with partial CSIT due to Rate Approximation. The bounds for ZF are given by [3]. It is observed that RA not only performs better for realistic SNR but also is immensely robust to channel uncertainty.

only typically much smaller than that of ZF but has also better scaling properties compared to the classical result by [3] which is even improving in the number of transmit antennas. Consequently, together with 1) we characterize the single cell SNR operating points where a fixed codebook has better performance than ZF, clearly setting a new standard.

3) We outline an advanced vector quantization problem for the RA scheme replacing the common chordal distance with a new distance function which inherently uses the structure of the transmit codebook.

In Section V we underline our results with LTE system simulations showing the benefit obtained by the proposed RA scheme and develop a suboptimal feedback protocol dealing with the complexity issue. This feedback protocol is proposed to replace the common approach for

LTE. Finally, in Sec. VI the conclusion is drawn with emphasis on the impact on future standards.

Notation: Bold letters denote column vectors and bold capital letters matrices. The inner product between vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ is defined as $\langle \boldsymbol{a}, \boldsymbol{b} \rangle = \boldsymbol{a}^H \boldsymbol{b}$, where $\boldsymbol{a}^H$ is the conjugate transpose of the vector $\boldsymbol{a}$. $\mathbb{S}^{n-1}$ is the unit sphere in $\mathbb{C}^n$ for $0 < n \in \mathbb{N}$; $\boldsymbol{a}_{i,j}$ denotes the j–th component of the vector $\boldsymbol{a}_i$ and $[\boldsymbol{a}_i]_{i=1,\dots,n} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \dots, \boldsymbol{a}_n]$ is the matrix with column vectors $\boldsymbol{a}_i$. The non–negative integers are denoted as $\mathbb{N}_+$.

II. System Setup

A. System Model

We consider the MU–MIMO downlink channel of a cellular system where a base station, equipped with n_t transmit antennas, serves multiple users, equipped with n_r receive antennas, on the same time and frequency resource with a single data stream. The users are collected in the set $\mathcal{U}$. Let $\boldsymbol{x} \in \mathbb{C}^{n_t \times 1}$ be the transmitted signal. User m receives the transmitted signal through the channel $\boldsymbol{H}_m \in \mathbb{C}^{n_r \times n_t}$ and applies a receive filter $\boldsymbol{u}_m \in \mathbb{C}^{n_r \times 1}$ to recover its intended signal,

$$y_m = \langle \boldsymbol{u}_m, \boldsymbol{H}_m \boldsymbol{x} \rangle + n_m = \langle \hat{\boldsymbol{h}}_m, \boldsymbol{x} \rangle + n_m$$

where $n_m = \langle \boldsymbol{u}_m, \boldsymbol{n}_m \rangle \sim \mathcal{CN}(0, \sigma^2)$ is additive white Gaussian noise. The effective channel from the base station to user m is given by $\hat{\boldsymbol{h}}_m^H = \langle \boldsymbol{u}_m, \boldsymbol{H}_m \rangle$; the normalized effective channel will be denoted by $\boldsymbol{h}_m = \hat{\boldsymbol{h}}_m / \|\hat{\boldsymbol{h}}_m\|_2$ and the effective channel norm is $\mu_m = \|\hat{\boldsymbol{h}}_m\|_2$. Further, we assume each user m has perfect knowledge of its own channel $\boldsymbol{H}_m$ and that the channels are constant over one transmission interval; no fading model is imposed.

In each transmission interval the base station selects a subset $\mathcal{S} \subseteq \mathcal{U}$ of users for transmission and assigns each user $m \in \mathcal{S}$ a beamforming vector out of a transmit codebook $\mathcal{C} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \dots\}$, known to the base station and all users a priori. Sometimes, we will also assume that the codebook elements constitute an orthonormal basis (ONB) of $\mathbb{C}^{n_t}$, such that for every $\boldsymbol{f} \in \mathbb{C}^{n_t}$

$$\sum_{i=1}^{n_t} |\langle \boldsymbol{w}_i, \boldsymbol{f} \rangle|^2 = \|\boldsymbol{f}\|_2^2. \tag{1}$$

In that case we speak of UB. The assignment of users to beamforming vectors is defined by a mapping

$$\pi : \mathcal{S} \to \mathcal{C},$$

that maps each element in $\mathcal{S}$ to a codebook element from $\mathcal{C}$. We assume that $|\mathcal{S}| \leq n_s \leq n_t$, where n_s is the maximum number of scheduled users per resource. In the sequel we do not state the domain of π explicitly, if it is clear from the context. Define the complex information symbols intended for user m as $d_m \in \mathbb{C}$, then the transmitted signal can be given by the superposition

$$\boldsymbol{x} = \sum_{m\in\mathcal{S}} \sqrt{\frac{P}{|\mathcal{S}|}} \boldsymbol{w}_{\pi(m)} d_m.$$

In the following the power allocation is uniform, that is, the base station distributes its available power equally among all users. The transmitted signal must satisfy an average sum power constraint, i.e. $\mathbb{E}\left[\|\boldsymbol{x}\|_2^2\right] \leq P$. Define

$$\lambda_m^2 = \frac{P\mu_m^2}{n_t\sigma^2}$$

as the receive SNR of user $m \in \mathcal{U}$, which is a *measurable quantity at the receiver*. The achieved sum rate under any user selection $\mathcal{S}$ and any mapping $\pi : \mathcal{S} \to \mathcal{C}$ is

$$R\left(\pi, \boldsymbol{H}\right) = \sum_{m\in\mathcal{S}} r_m(\pi, \lambda_m \boldsymbol{h}_m) \tag{2}$$

where the effective channels are defined by the composite matrix $\boldsymbol{H} = [\lambda_m \boldsymbol{h}_m]_{m\in\mathcal{S}}$ for some set of users $\mathcal{S} \subseteq \mathcal{U}$. The per user contributions to the sum rate are given by the Shannon rates

$$r_m(\pi, \lambda_m \boldsymbol{h}_m) = \log\left(1 + \frac{\lambda_m^2 \phi^{\boldsymbol{h}}_{m,\pi(m)}}{\frac{|\mathcal{S}|}{n_t} + \lambda_m^2 \sum_{l\in\mathcal{S}\setminus\{m\}} \phi^{\boldsymbol{h}}_{m,\pi(l)}}\right) \tag{3}$$

where the effective channel gains are defined as

$$\phi^{\boldsymbol{h}}_{m,\pi(m)} = |\langle \boldsymbol{h}_m, \boldsymbol{w}_{\pi(m)}\rangle|^2.$$

Throughout the paper we assume maximum sum rate scheduling, for instance, with perfect CSIT the optimal scheduling decision $\pi_H : \mathcal{S} \to \mathcal{C}$ is

$$(\mathcal{S}, \pi_H) = \underset{\substack{\mathcal{S}\subseteq\mathcal{U} \\ \pi:\mathcal{S}\to\mathcal{C}}}{\arg\max}\, R\left(\pi, \boldsymbol{H}\right). \tag{4}$$

However, due to the rate–constrained feedback channel, the base station takes its decisions based solely on partial CSIT from each user. Clearly, these decisions should match the optimal decision as good as possible. This challenge can be most efficiently approached by the following RA scheme.

III. Rate Approximation

Before introducing and analyzing the RA scheme we review some basic results for ZF, which is our baseline performance.

A. Baseline Performance: ZF

For the ZF scheme with perfect CSIT, each beamforming vector $\boldsymbol{w}_{ZF,m} \in \mathbb{C}^{n_t}$ is chosen to be in the null space of the matrix $[\boldsymbol{h}_l]_{l \in \mathcal{S} \setminus \{m\}}$ (see e.g. [3]). The sum rate is then given by

$$R_{ZF}(\boldsymbol{H}) = \sum_{m \in \mathcal{S}} \log\left(1 + \lambda_m^2 \phi_{m,ZF}^{\boldsymbol{h}}\right), \tag{5}$$

which is almost surely non–zero for any receive SNR $\lambda_m > 0$. Here, we have defined $\phi_{m,ZF}^{\boldsymbol{h}} = |\langle \boldsymbol{h}_m, \boldsymbol{w}_{ZF,m} \rangle|^2$. Furthermore, define $R(\boldsymbol{H}, \mathcal{V}_{\text{RVQ}})$ as the sum rate of ZF under limited feedback based on random vector quantization (RVQ) [6], where $\mathcal{V}_{\text{RVQ}} = \{\boldsymbol{\nu}^{[1]}, \ldots, \boldsymbol{\nu}^{[2^B]}\}$ is a codebook of random vectors isotropically distributed on the unit sphere. In contrast to the RA scheme (introduced later on) user m chooses his feedback message to minimize the chordal distance

$$d_C(\boldsymbol{h}_m, \boldsymbol{\nu}) = \sqrt{1 - |\langle \boldsymbol{\nu}, \boldsymbol{h}_m \rangle|^2}. \tag{6}$$

From [3, Theorem 2] we have the following performance bound for ZF with RVQ. Suppose that $\hat{\boldsymbol{h}}_{m,i} \sim \mathcal{CN}(0,1)$ independent across users and antennas. Then, limited feedback with B feedback bits per user incurs a throughput loss relative to ZF with perfect CSIT according to

$$\begin{aligned} \Delta R_{\text{RVQ}} &= \mathbb{E}_{\boldsymbol{H}}\left[R_{ZF}(\boldsymbol{H})\right] - \mathbb{E}_{\boldsymbol{H},\mathcal{V}}\left[R(\boldsymbol{H}, \mathcal{V}_{\text{RVQ}})\right] \\ &< n_t \log(1 + \frac{P}{\sigma^2} 2^{-\frac{B}{n_t - 1}}). \end{aligned} \tag{7}$$

B. Rate Approximation Scheme

Assume the base station has partial CSIT $(\vartheta_m, \boldsymbol{\nu}_m)$ for each user $m \in \mathcal{U}$. The channel direction information (CDI) $\boldsymbol{\nu}_m \in \mathcal{V}$ is given by an element of a feedback codebook $\mathcal{V} = \{\boldsymbol{\nu}^{[1]}, \ldots, \boldsymbol{\nu}^{[2^B]}\}$ that consists of a collection of normalized vectors $\boldsymbol{\nu}^{[i]} \in \mathbb{S}^{n_t - 1}$ and is a priori known to all users and the base station. The scalar $\vartheta_m \in \mathbb{R}$ can be interpreted as the channel quality information (CQI). In the sequel we assume that the CQI is perfectly transferred to the base station, which is a typical assumption. A natural choice for the CQI is $\vartheta_m = \lambda_m$ if CDI can be perfectly tracked, but under partial CSIT the choice is in general crucial.

If the beamforming vectors are restricted to a fixed codebook $\mathcal{C}$ the scheduling decision based on partial CSIT $\boldsymbol{V} = [\vartheta_m \boldsymbol{\nu}_m]_{m\in\mathcal{S}}$ can be found by solving

$$(\mathcal{S}, \pi_V) = \underset{\substack{\mathcal{S}\subseteq\mathcal{U} \\ \pi:\mathcal{S}\to\mathcal{C}}}{\arg\max}\, R(\pi, \boldsymbol{V}), \tag{8}$$

which is a combinatorial problem that can be solved either by a brute force search over the user sets $\mathcal{S} \subseteq \mathcal{U}$, with $|\mathcal{S}| \leq n_s$ and the mappings $\pi : \mathcal{S} \to \mathcal{C}$ or, alternatively, more efficiently in a greedy fashion [7], [8].

Let us now derive the RA scheme. The aim of the RA scheme is that the scheduling decision under partial CSIT matches with the optimal scheduling decision (4) as good as possible. In the following we demonstrate how the error can be calculated thereby circumventing the optimal combinatorial solution. Define the average rate gap between ZF with perfect CSIT and beamforming based on a fixed codebook with perfect CSIT as

$$\Delta R_{\text{CSIT}} := \mathbb{E}_{\boldsymbol{H}}\left[R_{ZF}(\boldsymbol{H})\right] - \mathbb{E}_{\boldsymbol{H}}\left[R(\pi_H, \boldsymbol{H})\right]$$

and the average rate gap between the real sum rates $R(\pi, \boldsymbol{H})$ and the approximated sum rates $R(\pi, \boldsymbol{V})$ as

$$\Delta R(\pi) := \mathbb{E}_{\boldsymbol{H}}\left[R(\pi, \boldsymbol{H})\right] - \mathbb{E}_{\boldsymbol{H}}\left[R(\pi, \boldsymbol{V})\right]. \tag{9}$$

Now, the rate gap between ZF with perfect CSIT and beamforming based on a fixed codebook with partial CSIT can be bounded from above by

$$\begin{aligned}
\Delta R_{\text{ZF}} &= \mathbb{E}_{\boldsymbol{H}}\left[R_{\text{ZF}}(\boldsymbol{H})\right] - \mathbb{E}_{\boldsymbol{H}}\left[R(\pi_V, \boldsymbol{H})\right] \\
&= \Delta R_{\text{CSIT}} + \mathbb{E}_{\boldsymbol{H}}\left[R(\pi_H, \boldsymbol{H})\right] - \mathbb{E}_{\boldsymbol{H}}\left[R(\pi_V, \boldsymbol{H})\right] \\
&= \Delta R_{\text{CSIT}} + \Delta R(\pi_H) + \mathbb{E}_{\boldsymbol{H}}\left[R(\pi_H, \boldsymbol{V})\right] - \mathbb{E}_{\boldsymbol{H}}\left[R(\pi_V, \boldsymbol{H})\right] \\
&\leq \Delta R_{\text{CSIT}} + \Delta R(\pi_H) - \Delta R(\pi_V) && (10) \\
&\leq \Delta R_{\text{CSIT}} + 2\mathbb{E}_{\boldsymbol{H}}\left[\sum_{m\in\mathcal{S}} \max_{\substack{\mathcal{S}_m\subset\mathcal{S} \\ \pi:\mathcal{S}_m\to\mathcal{C}}} |r_m(\pi, \lambda_m \boldsymbol{h}_m) - r_m(\pi, \vartheta_m \boldsymbol{\nu}_m)|\right], && (11)
\end{aligned}$$

where (10) must hold since π_V is the optimal mapping of users to beamforming vectors under channels $\boldsymbol{V}$. In (11) we defined the set

$$\mathcal{S}_m := \{\{m\}, \{m, m+1\}, \ldots, \{m, m+1, \ldots, m+n_s-1\}\},$$

that is, the set of possible user selections which include user m. Moreover, we exploited that the rate gap $\Delta R\left(\pi_H\right) - \Delta R(\pi_V)$ is bounded from above by the worst case rate gap

$$\Delta R_{\text{RA}} := 2 \cdot \mathbb{E}_{\boldsymbol{H}} \left[\sum_{m \in \mathcal{S}} \max_{\substack{\mathcal{S}_m \subset \mathcal{S} \\ \pi : \mathcal{S}_m \to \mathcal{C}}} \left| r_m(\pi, \lambda_m \boldsymbol{h}_m) - r_m(\pi, \vartheta_m \boldsymbol{\nu}_m) \right| \right]. \tag{12}$$

From (11) we observe that to bound ΔR_{ZF} it is sufficient if each user individually minimizes its individual rate gaps $|r_m(\pi, \lambda_m \boldsymbol{h}_m) - r_m(\pi, \vartheta_m \boldsymbol{\nu}_m)|$ for any $\mathcal{S}_m$. This observation is the fundament for the following RA scheme.

To determine its feedback message each user $m \in \mathcal{U}$ must find a tuple $(\vartheta_m, \boldsymbol{\nu}_m) \in (\mathbb{R}, \mathcal{V})$ that minimizes the RA distance[1]

$$d\left(\lambda_m \boldsymbol{h}_m, \vartheta_m \boldsymbol{\nu}_m\right) = \max_{\substack{\mathcal{S}_m \\ \pi : \mathcal{S}_m \to \mathcal{C}}} \left| r_m\left(\pi, \lambda_m \boldsymbol{h}_m\right) - r_m\left(\pi, \vartheta_m \boldsymbol{\nu}_m\right) \right|. \tag{13}$$

That is, each user $m \in \mathcal{U}$ finds its feedback message by solving

$$(\vartheta_m, \boldsymbol{\nu}_m) = \arg\min_{\substack{\vartheta \in \mathbb{R} \\ \boldsymbol{\nu} \in \mathcal{V}}} d\left(\lambda_m \boldsymbol{h}_m, \vartheta \boldsymbol{\nu}\right). \tag{14}$$

The RA scheme can be easily extended to users with multiple receive antennas $n_r > 1$. In this case for each scheduling decision $\pi : \mathcal{S}_m \to \mathcal{C}$ the optimal receive filter can be considered in the RA distance, i.e. the real rates are given by

$$r_m(\pi, \boldsymbol{h}_m) = \max_{\boldsymbol{u} \in \mathbb{C}^{n_r}} r_m\left(\pi, \langle \boldsymbol{u}, \boldsymbol{H}_m \rangle^H\right).$$

Although not apparent at this point let us formulate some decisive advantages of the RA scheme: first, in the RA distance $d(\cdot, \cdot)$ the transmit codebook matters which seems good engineering practice as we use all the available information. Second, the terminals provide an uniform error which indicates how well the rates are approximated and leads to inherent robustness. This becomes particularly beneficial in the LTE multi antenna case where CSI information is averaged over the subcarriers (see Simulations in Section V). Third, the RA scheme is amendable to codebook optimization [9] due to our new distance function. Finally, let us establish that the RA distance is indeed different compared to the chordal distance by the following example.

Toy Example. *To illustrate how the RA scheme operates we consider a toy example in* $\mathbb{R}^2$. *We compare the feedback decisions taken under the RA distance* (13) *and minimum chordal*

[1] A closer look reveals that it is neither in all cases a distance on $\mathbb{C}^{n_t}$ nor on the Grassmann manifold.

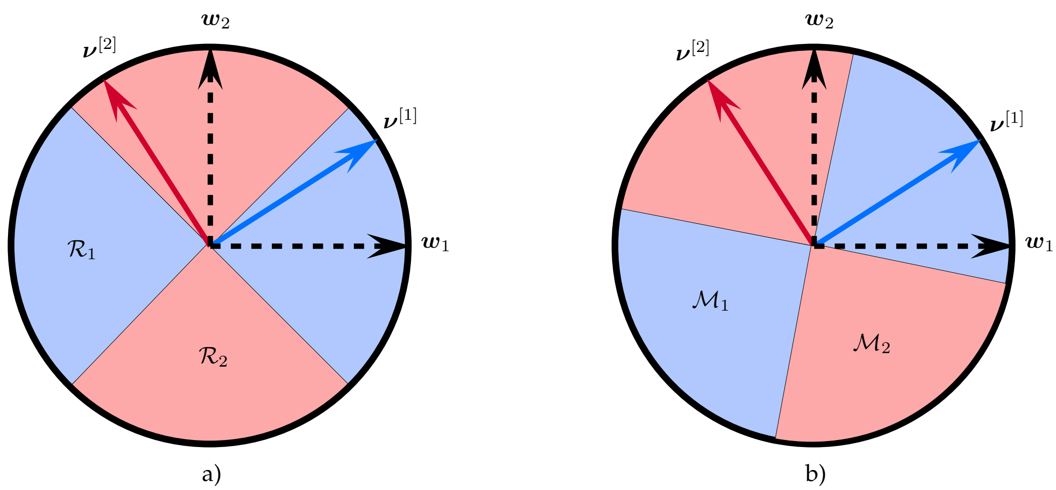

Fig. 2. Sets $\mathcal{R}_i$ and $\mathcal{M}_i$ in the toy example. a) RA distance: If a user experiences a channel $\boldsymbol{h} \in \mathcal{R}_i$ the quantized channel $\boldsymbol{v}_i$ is fed back. b) Minimum chordal distance: A user experiencing a channel $\boldsymbol{h} \in \mathcal{M}_i$ feeds back channel quantization $\boldsymbol{v}_i$.

distance (6) *by user* $m \in \mathcal{U}$*. Assume* $n_s = 2$*,* $n_t = 2$ *and* $n_r = 1$ *such that the channel vector is given by* $\boldsymbol{h}_m \in \mathbb{R}^2$*. The transmit codebook is given by the columns of the identity matrix* $\mathcal{C} = \{(1 \quad 0)^T, (0 \quad 1)^T\}$ *and the feedback codebook is given by a rotated version of the transmit codebook. The CQI is equal to the receive SNR* $\vartheta_m^2 = \lambda_m^2$*. For the RA distance we define the sets*

$$\mathcal{R}_i = \left\{ \boldsymbol{h}_m \in \mathbb{R}^2 : \boldsymbol{\nu}^{[i]} = \arg\min_{\boldsymbol{\nu} \in \mathcal{V}} d(\lambda_m \boldsymbol{h}_m, \lambda_m \boldsymbol{\nu}) \right\}.$$

That is, all channels that result in the codebook element $\boldsymbol{v}_i \in \mathcal{V}$*. In a similar manner for the minimum chordal distance we define the sets*

$$\mathcal{M}_i = \left\{ \boldsymbol{h}_m \in \mathbb{R}^2 : \boldsymbol{\nu}^{[i]} = \arg\min_{\boldsymbol{\nu} \in \mathcal{V}} d_C(\boldsymbol{h}_m, \boldsymbol{\nu}) \right\}.$$

Figure 2 a) shows the two sets $\mathcal{R}_1$ *and* $\mathcal{R}_2$ *and Figure 2 b) shows the sets* $\mathcal{M}_1$ *and* $\mathcal{M}_2$*. We observe that the RA distance results in a different CDI feedback decision compared to the minimum chordal distance. The RA feedback decision is obvious more oriented on the transmit codebook.*

In the next section we show that ΔR_{RA} can be made small and that there is an SNR range for which ΔR_{CSIT} is negative.

IV. Performance Analysis

A. Problem Formulation

Many papers prove that a particular transmission scheme achieves the optimal multiuser multiplexing gain. That is, for sufficiently large $|\mathcal{U}|$ the sum rate scales like $n_t \log\log|\mathcal{U}|$. For instance this was shown for random beamforming [5], ZF [10] and UB [11]. However, since rates and the number of users are finite in a practical system, the significance of these asymptotic results can at least be questioned. Putting it the other way around: two methods achieving the optimal gain might behave completely different in a practical system. This implies that it does not say much about individual rates of users.

Our analysis is different and more inspired by the finite user results in [3], [12]. For the sake of analytical tractability, we first constrain the number of users and selected users by $|\mathcal{U}| = |\mathcal{S}| = n_t$, thereby circumventing the user selection process and enabling stringent comparison to Jindal's results in [3], and also resort to UB. Later on, we will abandon this assumption and incorporate user selection and general codebooks as well.

Our goal hereafter is to upperbound the rate loss with respect to ZF with perfect CSI; this is done in two steps: 1.) first the base station is assumed to have ideal CSIT, but operates solely on the set of beamforming vectors of the transmit codebook (this corresponds to evaluating ΔR_{CSIT}) and 2.) the individual rate error due to the rate–constrained feedback channel is bounded from above by involving the RA scheme (this corresponds to evaluating ΔR_{RA}).

B. ZF vs. UB: Perfect CSIT

So far we have not used any fading model. In the following theorem, the channel is modeled as isotropic fading [13]. In particular, we fix the channel gains μ_m for $m \in \mathcal{U}$ and let the only randomness be an independent phase ambiguity in the channel coefficients, i.e. $\hat{\boldsymbol{h}}_m = \mu_m \boldsymbol{h}_m$ where $\boldsymbol{h}_m \in \mathbb{S}^{n_t-1}$ is an isotropically distributed random complex unit vector. Note that this assumption has no impact on the generality of the result; in fact it only allows us to streamline the presentation throughout the proof.

The following theorem is qualitatively known but we make it mathematically more precise.

Theorem 1. *If $\mathcal{U} = \mathcal{S} = \{1, 2, \dots, n_t\}$, then for isotropic fading with any non–random $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_{n_t}$ (respectively non–random $\lambda_1^2 \geq \lambda_2^2 \geq \ldots \geq \lambda_{n_t}^2$) the rate gap between ZF and UB* *with perfect CSI is bounded from above by*

$$\Delta R_{CSIT} \leq \sum_{m\in\mathcal{S}} \log\left(1+\min_{\epsilon>0}\lambda_m^2 \frac{(1+\lambda_m^2)\,\frac{1}{n_t-1} - \left(1+\frac{\lambda_m^2}{n_t-1}\right)\frac{c(\epsilon)(1-\epsilon)\log(n_t-m+1)}{(1+\epsilon)^2 n_t}}{1+\lambda_m^2}\right)$$

where $c(\epsilon) := (1-e^{-(n_t-m+1)^{\epsilon}} - \frac{1}{\epsilon n_t})$.

The proof can be found in Appendix A. Theorem 1 states that for some $\lambda_m \leq \lambda^*$ the rate gap ΔR_{CSIT} is indeed negative; this can be seen by observing that at low SNR (i.e. $\lambda_m \to 0$) for all $m < n_t$ the difference is roughly

$$\frac{1}{n_t-1} - \frac{\log(n_t-m+1)}{n_t} < 0$$

by choosing e.g. $c(\epsilon) = \frac{\log(\log(n_t))}{\log(n_t)}$ and n_t large enough. Therefore, for low SNR unitary beamforming can indeed outperform ZF.

C. Uniform RA Error with UB

Let us first provide a general expression for the maximum in (12) which gives us a hint how the RA scheme operates. Note that from now on when the RA scheme operates on a unitary transmit codebook we will denote this scheme by RA–UB.

Lemma 1. *If* $\mathcal{U} = \mathcal{S} = \{1,2,\ldots,n_t\}$ *and* $\mathcal{C} \subseteq \mathcal{V}$ *then for some pair* $\boldsymbol{h}_m, \boldsymbol{\nu}_m$ *under the RA–UB scheme*

$$\max_{\substack{\mathcal{S}_m \\ \pi:\mathcal{S}_m\to\mathcal{C}}} |r_m(\pi,\lambda_m\boldsymbol{h}_m) - r_m(\pi,\vartheta_m\boldsymbol{\nu}_m)| \tag{15}$$

$$\leq \max_{\boldsymbol{w}\neq\boldsymbol{w}^*} \log\left(1+\frac{\lambda_m^2\left(\left||\langle\boldsymbol{h}_m,\boldsymbol{w}\rangle|^2 - |\langle\boldsymbol{\nu}_m,\boldsymbol{w}\rangle|^2\right| + \frac{\langle\boldsymbol{\nu},\boldsymbol{w}\rangle}{\langle\boldsymbol{\nu},\boldsymbol{w}^*\rangle}\left||\langle\boldsymbol{\nu}_m,\boldsymbol{w}^*\rangle|^2 - |\langle\boldsymbol{h}_m,\boldsymbol{w}^*\rangle|^2\right|\right)}{1+\lambda_m^2\left(1-\max\{|\langle\boldsymbol{h}_m,\boldsymbol{w}\rangle|^2, |\langle\boldsymbol{\nu}_m,\boldsymbol{w}\rangle|^2\}\right)}\right)$$

where we defined $\boldsymbol{w}^*$ *by* $|\langle\boldsymbol{h}_m,\boldsymbol{w}\rangle|^2 \leq |\langle\boldsymbol{h}_m,\boldsymbol{w}^*\rangle|^2$ *for all* $\boldsymbol{w}\in\mathcal{C}$. *The strategy is to pick* $\boldsymbol{\nu}_m$ *close to* $\boldsymbol{h}_m$ *as long as* $|\langle\boldsymbol{\nu}_m,\boldsymbol{w}^*\rangle|^2 \geq |\langle\boldsymbol{h}_m,\boldsymbol{w}^*\rangle|^2$.

The proof can be found in Appendix B. While the error term is still not easily accessible the former lemma allows us to devise the following corollaries.

Corollary 1. *If* $\mathcal{C} \equiv \mathcal{V}$, *then under the RA–UB scheme*

$$\Delta R_{RA} \leq \sum_{m=1}^{n_t} \mathbb{E}_{\boldsymbol{H}}\left[\log\left(1+\max_{\boldsymbol{w}\neq\boldsymbol{w}^*}\frac{\lambda_m^2|\langle\boldsymbol{h}_m,\boldsymbol{w}\rangle|^2}{1+\lambda_m^2(1-|\langle\boldsymbol{h}_m,\boldsymbol{w}\rangle|^2)}\right)\right]$$

with CDI $\boldsymbol{\nu}_m = \boldsymbol{w}^*$ *and CQI defined by*

$$\vartheta_m^2 := \frac{\lambda_m^2 |\langle \boldsymbol{h}_m, \boldsymbol{\nu}_m \rangle|^2}{1 + \lambda_m^2 (1 - |\langle \boldsymbol{h}_m, \boldsymbol{\nu}_m \rangle|^2)}. \tag{16}$$

Proof: Since, $\mathcal{C} \equiv \mathcal{V}$ we have that $|\langle \boldsymbol{\nu}_m, \boldsymbol{w} \rangle|^2 = 0$ for all $\boldsymbol{w} \neq \boldsymbol{w}^*$. Plugging (15) in (12) proves the claim. □

Corollary 1 gives a universal lower bound on the performance of any RA–UB scheme. Observe that a CQI similar to (16) was proposed in [8] in the context of ZF and also in [5] using random beamforming. Equation (16) is also in accordance with [10], where it was shown that the CQI should reflect the channel magnitude and the quality of the channel quantization. For the RA scheme, in the simulations, we will also consider the CQI

$$\vartheta_m^2 = \lambda_m^2 |\langle \boldsymbol{h}_m, \boldsymbol{\nu}_m \rangle|^2, \tag{17}$$

which can be interpreted as the effective channel of user m over the quantized channel $\boldsymbol{\nu}_m$. Equation (17) captures two important aspects. On the one hand, if the CDI is equal to the channel direction, the user gets no penalty (i.e. $|\langle \boldsymbol{h}_m, \boldsymbol{\nu}_m \rangle|^2 = 1$) on the other hand if the CDI is orthogonal to the channel direction, the effective channel is zero (i.e. $|\langle \boldsymbol{h}_m, \boldsymbol{\nu}_m \rangle|^2 = 0$). Hence, the CQI (17) reflects the receive SNR and the quantization error.

Another consequence of Lemma 1 is that the sum rate $R(\pi_V, \boldsymbol{H})$ achieves the optimal throughout scaling for asymptotically many users. We use the (standard) Rayleigh fading channel model, i.e. $\hat{\boldsymbol{h}}_{m,i} \sim \mathcal{CN}(0,1)$ for $i = 1, \ldots, n_t$ and $m \in \mathcal{U}$.

Corollary 2. *Let the unitary transmit codebook be a subset of the feedback codebook* $\mathcal{C} \subseteq \mathcal{V}$ *and each channel is Rayleigh, then under the RA–UB scheme we have*

$$\frac{\mathbb{E}_{\boldsymbol{H}} \left[R(\pi_V, \boldsymbol{H}) \right]}{n_t \log \log |\mathcal{U}|} \to 1, \ |\mathcal{U}| \to \infty.$$

Proof: From [5, Lemma 1] we know that under perfect CSIT the sum rate $R(\pi_H, \boldsymbol{H})$ defined in (2) scales as $n_t \log \log |\mathcal{U}|$. From Lemma 1 we have that the rate gap $\Delta R_{\text{RA}} = R(\pi_H, \boldsymbol{H}) - R(\pi_V, \boldsymbol{H})$ remains bounded when $|\mathcal{U}|$ increases. This concludes the proof. □

So far we are still not able to effectively upperbound ΔR_{RA} which is now settled. The following lemma shows that ΔR_{RA} remains bounded when the SNR increases and that the rate error depends solely on the function

$$D_m(B) := \mathbb{E}_{\boldsymbol{H}}\left[\min_{\substack{1>\tilde{\vartheta}_m>0\\ \boldsymbol{\nu}\in\mathcal{V}}}\max_{\boldsymbol{w}\in\mathcal{C}}\left|\tilde{\lambda}_m|\langle\boldsymbol{h},\boldsymbol{w}\rangle|^2-\tilde{\vartheta}_m|\langle\boldsymbol{\nu},\boldsymbol{w}\rangle|^2\right|\frac{1}{1-\tilde{\lambda}_m}\right], \tag{18}$$

where we defined $\tilde{\lambda}_m = \frac{\lambda_m}{1+\lambda_m}$, $\tilde{\vartheta}_m = \frac{\vartheta_m}{1+\vartheta_m}$ and B is the number of feedback bits, i.e. 2^B is the number of elements in the feedback codebook $\mathcal{V}$.

Lemma 2. *If $\mathcal{U} = \mathcal{S} = \{1, 2, \ldots, n_t\}$ and $\mathcal{C} \subseteq \mathcal{V}$ then under the RA–UB scheme,*

$$\Delta R_{RA} \leq 2\sum_{m=1}^{n_t}\log\left(1+\min_{\epsilon>0}\frac{(1+\epsilon)\,D_m(B)}{1+\frac{\epsilon}{n_t-1}D_m(B)}\right)$$

The proof can be found in Appendix C.

The following lemma gives a fundamental bound on $D_m(B)$.

Lemma 3. *If $|\mathcal{V}| = 2^B$ and $\tilde{\vartheta}_m = \tilde{\lambda}_m$ for all $m \in \mathcal{U}$ and the transmit codebook $\mathcal{C}$ is unitary, then*

$$D_m(B) \leq c(n_t)\mathbb{E}_{\boldsymbol{H}}\left[\lambda_m\right]2^{-\frac{B}{n_t-1}},$$

with

$$c(n_t) = \left(\Theta(\mathcal{B}_2^{n_t-1})\binom{2n_t-2}{n_t-1}\frac{\Gamma(1+\frac{n_t-1}{2})\sqrt{n_t}}{(n_t-1)!\pi^{\frac{n_t-1}{2}}}\right)^{\frac{1}{n_t-1}}.$$

and $B \geq \frac{(n_t-1)}{2}\log[(n_t-1)\sqrt{n_t-1}]$. For $n_t - 1$ small tight bounds are known for the covering density $\Theta(\mathcal{B}_2^{n_t-1})$, e.g. $\Theta(\mathcal{B}_2^2) \leq 1.2091$ (Kershner, 1939), $\Theta(\mathcal{B}_2^3) \leq 1.4635$ (Bambah, 1954), $\Theta(\mathcal{B}_2^4) \leq 1.7655$ (Delone & Ryshkov, 1963). For $n_t - 1 \geq 3$ the Rogers bound [14] $\Theta(\mathcal{B}_2^d) < 4(n_t-1)\log(n_t-1)$ can be used.

The complete proof can be found in the Appendix D. Note that $c(n_t)$ is close to unity but falls below unity not before $n_t \geq 8$ as required for improved scaling compared to Jindal's result. This is an artefact of the proof technique as the following illustration for the case $n_t = 3$ shows.

Without loss of generality, we assume the unitary transmit codebook is given by the standard ONB and define the real positive vectors

$$\boldsymbol{\phi}_m^{\boldsymbol{h}} = \left(\phi_{m,\pi(1)}^{\boldsymbol{h}}, \phi_{m,\pi(2)}^{\boldsymbol{h}}, \phi_{m,\pi(3)}^{\boldsymbol{h}}\right)^T \text{ and } \boldsymbol{\phi}_m^{\boldsymbol{\nu}} = \left(\phi_{m,\pi(1)}^{\boldsymbol{h}}, \phi_{m,\pi(2)}^{\boldsymbol{\nu}}, \phi_{m,\pi(3)}^{\boldsymbol{\nu}}\right)^T,$$

per definition, this vectors have unit ℓ_1–norm, i.e., $\|\boldsymbol{\phi}_m^{\boldsymbol{h}}\|_1 = \|\boldsymbol{\phi}_m^{\boldsymbol{\nu}}\|_1 = 1$ and define points on the standard 2–simplex. Further, $\max_\pi\left|\phi_{m,\pi(m)}^{\boldsymbol{h}} - \phi_{m,\pi(m)}^{\boldsymbol{\nu}}\right| = \|\boldsymbol{\phi}_m^{\boldsymbol{h}} - \boldsymbol{\phi}_m^{\boldsymbol{\nu}}\|_\infty$ defines a distance

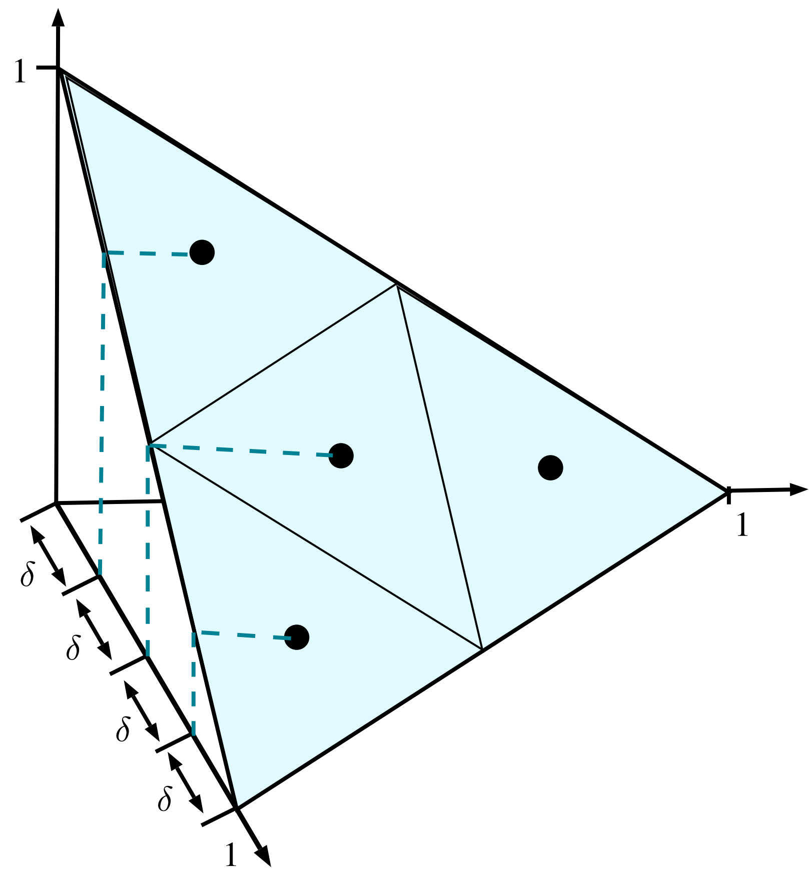

Fig. 3. The standard 2–simplex in 3 dimensions. The projection of the quantization points $\mathcal{Q}$ on the coordinate axes implies a worst case quantization error δ.

between two points on the standard 2–simplex. Hence, for a given feedback codebook $\mathcal{V}$ we can define the Voronoi region around the point $\phi_m^{\boldsymbol{\nu}}$ with $\boldsymbol{\nu} \in \mathcal{V}$ as

$$V(\boldsymbol{\phi}_m^{\boldsymbol{\nu}}) = \{\boldsymbol{x} \in \mathbb{R}_+^3 : \|\boldsymbol{x} - \boldsymbol{\phi}_m^{\boldsymbol{\nu}}\|_\infty < \|\boldsymbol{x} - \boldsymbol{\phi}_m^{\boldsymbol{\xi}}\|_\infty, \forall \boldsymbol{\xi} \in \mathcal{V}, \boldsymbol{\xi} \neq \boldsymbol{\nu}\}.$$

If $B \in \{1, 2, 4, \ldots\}$ and $n_t = 3$, the feedback codebook can be chosen such that the Voronoi regions are 2–simplices with edge length $\tilde{\delta} \leq \sqrt{2}$. Now, using the symmetry of the covering and projecting the quantization points back on the coordinate axes (see Figure 3) we get $\max_{\boldsymbol{x} \in V(\boldsymbol{\phi}_m^{\boldsymbol{\nu}})} \|\boldsymbol{x} - \boldsymbol{\phi}_m^{\boldsymbol{\nu}}\|_\infty = \tilde{\delta}/\sqrt{8} = \delta$.

Now we can compute the volumes of the 2–simplices (the standard simplex and the scaled simplex) and proceed as in the proof of Lemma 3 to obtain the result

$$\delta = \max_{\boldsymbol{x} \in V(\boldsymbol{\phi}_m^{\boldsymbol{\nu}})} \|\boldsymbol{x} - \boldsymbol{\phi}_m^{\boldsymbol{\nu}}\|_\infty = 2^{-\frac{B}{n_t - 1} - 1}.$$

Hence, if $n_t = 3$ and each channel is Rayleigh, using the RA–UB scheme the rate loss due to the rate–constrained feedback channel scales like

$$\Delta R_{\mathrm{RA}} \leq 2\mathbb{E}_{\boldsymbol{H}}\left[\sum_{m=1}^{n_t}\log\left(1+\min_{\epsilon>0}\frac{(1+\epsilon)\,\frac{\tilde{\lambda}_m}{1-\tilde{\lambda}_m}2^{-\frac{B}{n_t-1}-1}}{1+\frac{\tilde{\lambda}_m}{1-\tilde{\lambda}_m}\frac{\epsilon}{n_t-1}2^{-\frac{B}{n_t-1}-1}}\right)\right]$$
$$\leq 2\sum_{m=1}^{n_t}\log\left(1+\frac{P}{\sigma^2}2^{-\frac{B}{n_t-1}-1}\right).$$

Therefore, we have an improvement of $n_t - 1$ bits in the exponential term compared to Jindal's result for ZF with feedback based on minimizing the chordal distance (see Section III-A), under the very same assumptions.

D. User Selection and General Codebooks

In this subsection we no longer assume unitary transmit codebooks and allow user selection at the base station.

Theorem 2. *If the base station selects a subset $\mathcal{S} \subseteq \mathcal{U} = \{1, 2, \ldots, n_t\}$ of users per transmission and each channel is Rayleigh, then the RA scheme with B feedback bits per user incurs a throughput loss relative to perfect CSIT bounded from above by*

$$\Delta R_{RA} \leq 4n_s \log\left(1+\frac{Pn_t}{\sigma^2}\mathbb{E}_{\boldsymbol{H}}\left[\min_{\boldsymbol{\nu}\in\mathcal{V}}\max_{\boldsymbol{w}\in\mathcal{C}}\left||\langle\boldsymbol{h},\boldsymbol{w}\rangle|^2-|\langle\boldsymbol{\nu},\boldsymbol{w}\rangle|^2\right|\right]\right).$$

The proof can be found in Appendix E. The expected value

$$\mathbb{E}_{\boldsymbol{H}}\left[\min_{\boldsymbol{\nu}\in\mathcal{V}}\max_{\boldsymbol{w}\in\mathcal{C}}\left||\langle\boldsymbol{h},\boldsymbol{w}\rangle|^2-|\langle\boldsymbol{\nu},\boldsymbol{w}\rangle|^2\right|\right]$$

has been shown to be analytically traceable, in the previous section, for unitary transmit codebooks. However, its examination for arbitrary transmit codebooks $\mathcal{C}$ goes beyond the scope of this paper. It is expected that the scaling advantages of the RA scheme diminish with increasing transmit codebook size. But, clearly for unitary codebooks the previous result still holds which is remarkable, as all the 2^{n_t} possible user rates are uniformly recovered at the base station with better scaling properties than the classical result.

The RA scheme is now applied in a practical scenario.

V. Practical Considerations and Simulations

A. Efficient and Robust Feedback Protocol

Mobile user equipments usually have limited computing capabilities, therefore, most systems require that the complexity at the user side is as low as possible. Hence, solving the full rate approximation problem (i.e. the minimax problem (13)) may not be feasible. Fortunately, our analysis in Section IV yielded another distance function which can be used at the user side to uniformly bound the rate approximation error ΔR_{RA} (12) and is given by

$$d_S(\boldsymbol{h},\boldsymbol{\nu}) = \max_{\boldsymbol{w}\in\mathcal{C}} \left| |\langle \boldsymbol{h},\boldsymbol{w}\rangle|^2 - |\langle \boldsymbol{\nu},\boldsymbol{w}\rangle|^2 \right|. \tag{19}$$

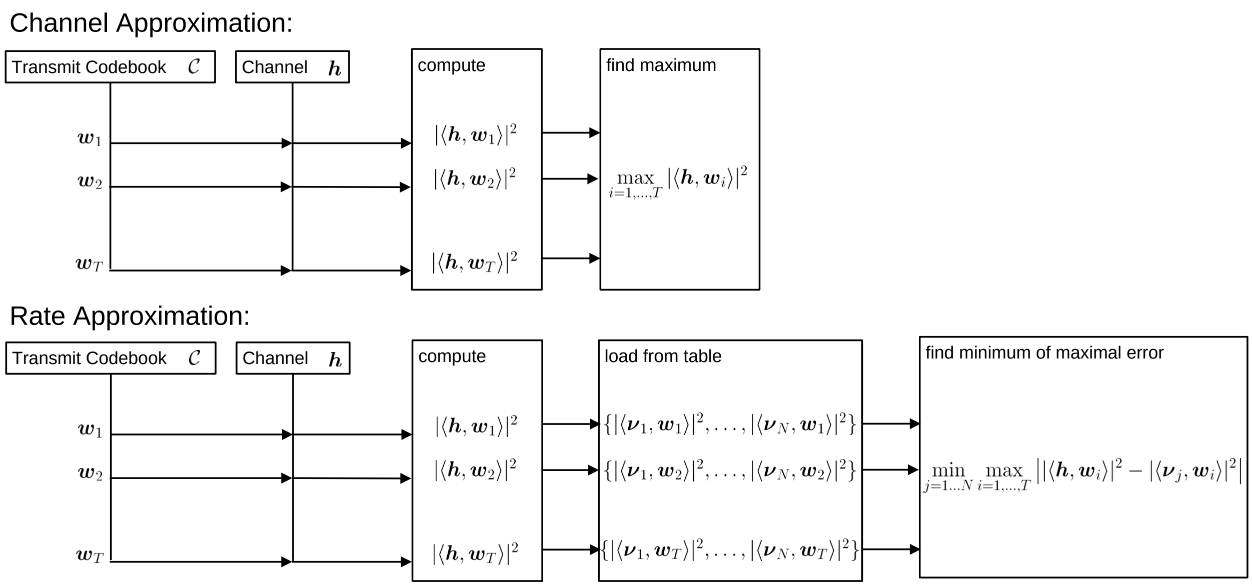

Fig. 4. Schematic comparison of the CDI computation at the user side; for the efficient RA distance (bottom) and chordal distance (top).

Figure 4 (bottom) shows a flow chart for the CDI computation under the proposed feedback protocol using the new distance function. The terms $|\langle \boldsymbol{\nu}_j,\boldsymbol{w}_i\rangle|$, for $i = 1,\ldots,|\mathcal{C}|$ and $j = 1,\ldots,|\mathcal{V}|$, only need to be computed ones and can be stored in the memory of each user. Hence, during the feedback phase user m must only compute $|\langle \boldsymbol{h}_m,\boldsymbol{w}_i\rangle|$ for all $i = 1,\ldots,|\mathcal{C}|$ and the difference $\left| |\langle \boldsymbol{h}_m,\boldsymbol{w}_i\rangle| - |\langle \boldsymbol{\nu}_j,\boldsymbol{w}_i\rangle| \right|$ for all i, j. Figure 4 (top) shows the steps that need to be performed to compute the CDI based on the chordal distance. To compute the chordal distance each user must compute $|\langle \boldsymbol{h}_m,\boldsymbol{w}_i\rangle|$ for all $i = 1,\ldots,|\mathcal{C}|$. We observe that determining the CDI under the proposed efficient feedback protocol is slightly more complex than using the minimum chordal distance. But, as we show next the proposed protocol achieves a huge sum rate gain.

B. Simulations

In the simulations we consider 3 base stations located in 3 adjacent cells and 30 users uniformly distributed over the network area, i.e. a radius of 250 meter around the center of the base stations. The physical layer is configured according to LTE [15]. The base station are equipped with $n_t = 4$ transmit antennas and each user is equipped with $n_r = 1$ or $n_r = 2$ receive antenna (specified in the caption). The transmit codebook and feedback codebook is given by the LTE codebook defined in [15] which has $N = 16$ elements and, hence, we require $B = 4$ bit to feedback back the CDI. The channels are modeled by the spatial channel model extended (SCME) [16] using the urban macro scenario. The simulation parameters a summarized in Table I.

TABLE I

SIMULATION PARAMETERS

Parameter	Value/Assumption
Number of base stations	3
Frequency reuse	full
Number of users $\|\mathcal{U}\|$	30 (uniformly distributed)
Number of transmit antennas n_t	4 (uncorrelated)
Number of receive antennas n_r	1 or 2 (uncorrelated)
Receiver type	maximum ratio combining
Maximum number of scheduled users per scheduling block n_s	4
Equivalent SNR	153 dB
LTE carrier frequency / bandwidth	2 GHz / 10 MHz
Number of PRB	50
Scheduling block size	1 PRB $=$ 12 subcarrier
LTE channel model	SCME (urban macro)
Inter cell interference modeling	explicit

Each user reports a feedback message to the base station with maximal receive SNR. We will consider four different feedback strategies:

1) Perfect (average) CSIT: the base station knows the channel averaged over all subcarriers perfectly, i.e.
$$\bar{\boldsymbol{H}}_m = \frac{1}{F}\sum_{f=1}^{F}\boldsymbol{H}_{m,f},$$
where $\boldsymbol{H}_{m,f}$ is the channel of user m on subcarrier f and F is the number of subcarriers for which feedback is generated.
2) Minimum chordal distance: user m determines its CDI feedback by minimizing the chordal distance (6) to the channel $\bar{\boldsymbol{h}}_m = \langle \boldsymbol{u}, \bar{\boldsymbol{H}}_m\rangle$, where $\boldsymbol{u}$ is chosen to maximize $|\langle \boldsymbol{u}, \bar{\boldsymbol{H}}_m\rangle|$.
3) Rate Approximation as described in Section III with the rates
$$r_m\left(\pi, \lambda_m \boldsymbol{h}_m\right) = \frac{1}{F}\sum_{f=1}^{F} r_m\left(\pi, \lambda_m \boldsymbol{h}_{m,f}\right)$$
4) Efficient Rate Approximation as described in Section V-A, where $\boldsymbol{h}_m$ is given by the average channel $\bar{\boldsymbol{h}}_m$ as defined for minimum chordal distance above.

The base stations run a local and independent scheduler. In each transmission interval the scheduler can select $n_s = 2$ users for transmission on the same time and frequency resources. The scheduling is performed in a greedy fashion according to [8]. For simplicity we assume no delay in the CSI report, scheduling, transmission or performance evaluation.

Figure 5 depicts the CDF of the spectral efficiency for users with $n_r = 1$ receive antennas. The ZF scheme is implemented according to [8]. The PU2RC scheme is based on the same transmit codebook as RA and is implemented according to [17]. We observe that with perfect CSIT ZF outperforms greedy scheduling with a fixed codebook. With partial CSIT the RA scheme significantly outperforms ZF with a gain of approximately up to 70%. Remarkable is also the gain of about 35% of RA over PU2RC. Moreover, Figure 5 shows that RA with the efficient distance function (19) performs very close to the full RA scheme.

Figure 6 depicts the CDF of the spectral efficiency for users with $n_r = 2$ receive antennas. We observe that with perfect CSIT ZF outperforms greedy scheduling with a fixed codebook. With partial CSIT the RA scheme significantly outperforms all other schemes and achieves a gain of approximately 50% over ZF. Remarkable is also the 35% gain of RA over PU2RC.

VI. Conclusion

We analyzed the performance of RA and showed that it outperforms ZF for a large fraction of a practically relevant SNR range. Hence, a remarkable result is that it is often much better to reduce

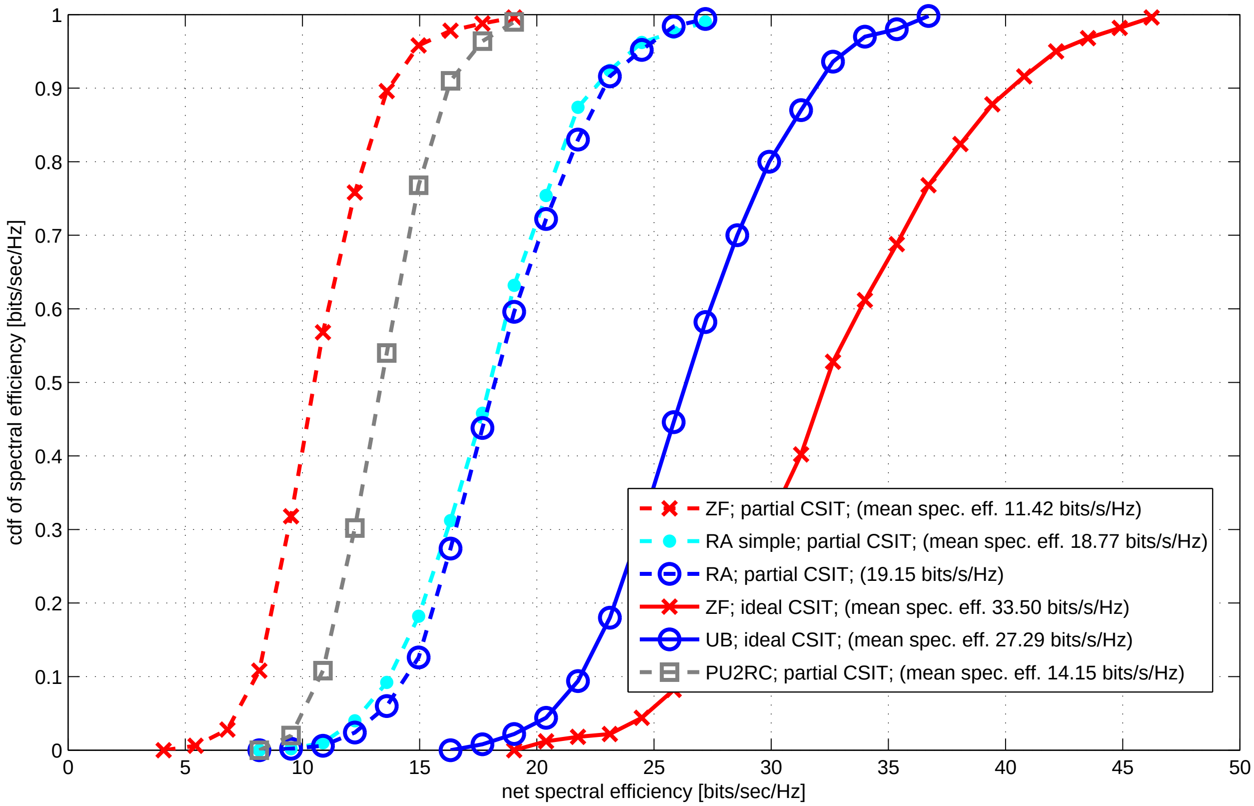

Fig. 5. CDF of spectral efficiency; Comparing PU2RC and ZF under perfect and partial CSIT; Setup: $n_r = 1$.

flexibility at the base station in favor of having more reliable CSIT. Several advantages of RA not been accounted for in this paper and might be part of future work. The RA scheme separates the feedback and transmit codebook, which allows to design both codebooks under different constraints. For example, the transmit codebook could be designed to minimize the inter cell interference where the feedback codebook could be designed to minimize the rate approximation error. Furthermore, each user could load different feedback codebooks if its environment changes. Finally, having approximations of the individual user rates available at the base station can be beneficial in multi cell systems with cooperating base stations.

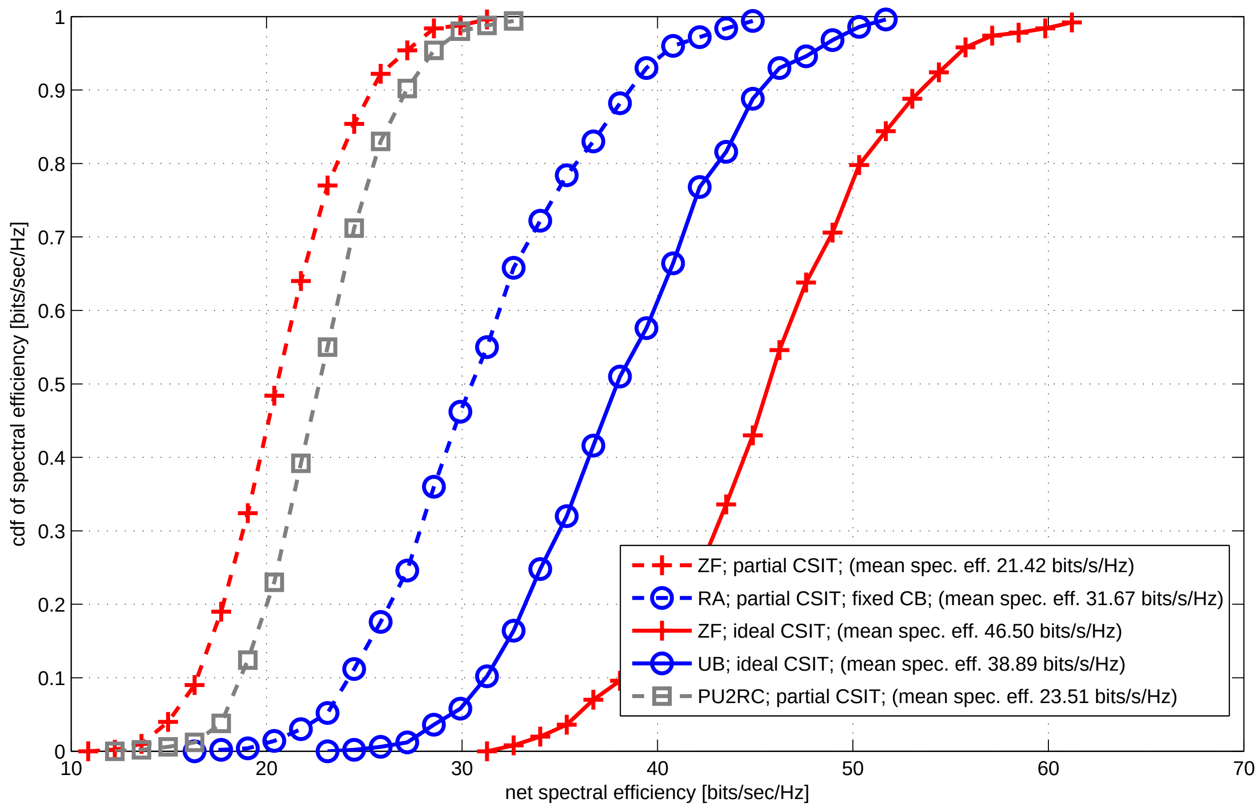

Fig. 6. CDF of spectral efficiency; Comparing RA, PU2RC and ZF under perfect and partial CSIT; Setup: $n_r = 2$.

APPENDIX A
PROOF OF THEOREM 1

Proof: For any mapping of beamforming vectors π we have

$$\begin{aligned}
\Delta R_{\text{CSIT}} &\leq \mathbb{E}_{\boldsymbol{H}}\left[R_{ZF}\left(\boldsymbol{H}\right) - R\left(\pi, \boldsymbol{H}\right)\right] \\
&= \mathbb{E}_{\boldsymbol{H}}\left[\sum_{m \in \mathcal{S}} \log\left(1 + \lambda_m^2 \phi_{m,ZF}^{\boldsymbol{h}}\right) - \log\left(1 + \frac{\lambda_m^2 \phi_{m,\pi(m)}^{\boldsymbol{h}}}{1 + \lambda_m^2\left(1 - \phi_{m,\pi(m)}^{\boldsymbol{h}}\right)}\right)\right] \\
&= \mathbb{E}_{\boldsymbol{H}}\left[\sum_{m \in \mathcal{S}} \log\left(1 + \lambda_m^2 \frac{\phi_{m,ZF}^{\boldsymbol{h}} - \phi_{m,\pi(m)}^{\boldsymbol{h}} + \lambda_m^2 \phi_{m,ZF}^{\boldsymbol{h}}(1 - \phi_{m,\pi(l)}^{\boldsymbol{h}})}{1 + \lambda_m^2}\right)\right] \\
&= \mathbb{E}_{\boldsymbol{H}}\left[\sum_{m \in \mathcal{S}} \log\left(1 + \lambda_m^2 \frac{\left(1 + \lambda_m^2\right)\phi_{m,ZF}^{\boldsymbol{h}} - \left(1 + \lambda_m^2 \phi_{m,ZF}^{\boldsymbol{h}}\right)\phi_{m,\pi(m)}^{\boldsymbol{h}}}{1 + \lambda_m^2}\right)\right].
\end{aligned}$$

Note that under UB $\pi : \mathcal{S} \to \mathcal{C}$ is simply a permutation on the beamforming vectors in $\mathcal{C}$ since $\mathcal{U} = \mathcal{S} = \{1, 2, \ldots, n_t\}$. Since, $\phi_{m,\pi(m)}^{\boldsymbol{h}}$ and $\phi_{m,ZF}^{\boldsymbol{h}}$ are independent random variables from

Jensen's inequality we get

$$\mathbb{E}_{\boldsymbol{H}} \sum_{m\in\mathcal{S}} \log\left(1+\lambda_m^2 \frac{\left(1+\lambda_m^2\right)\phi_{m,ZF}^{\boldsymbol{h}} - \left(1+\lambda_m^2 \phi_{m,ZF}^{\boldsymbol{h}}\right)\phi_{m,\pi(m)}^{\boldsymbol{h}}}{1+\lambda_m^2}\right)$$

$$\leq \sum_{m\in\mathcal{S}} \log\left(1+\lambda_m^2 \frac{\left(1+\lambda_m^2\right)\mathbb{E}_{\boldsymbol{H}}\left[\phi_{m,ZF}^{\boldsymbol{h}}\right] - \left(1+\lambda_m^2 \mathbb{E}_{\boldsymbol{H}}\left[\phi_{m,ZF}^{\boldsymbol{h}}\right]\right)\mathbb{E}_{\boldsymbol{H}}\left[\phi_{m,\pi(m)}^{\boldsymbol{h}}\right]}{1+\lambda_m^2}\right)$$

Due to the ZF criterion, for all i, $\boldsymbol{h}_i$ and $\boldsymbol{w}_{ZF,i}$ are independent isotropic vectors, see e.g. [3]. Hence, we have that $\phi_{m,ZF}^{\boldsymbol{h}} = |\langle \boldsymbol{h}_m, \boldsymbol{w}_{ZF,m}\rangle|^2$ is a beta distributed random variable with parameters 1 and n_t-2 and expectation

$$\mathbb{E}_{\boldsymbol{H}}\left[\phi_{m,ZF}^{\boldsymbol{h}}\right] = \frac{1}{n_t-1}.$$

The expectation $\mathbb{E}_{\boldsymbol{H}}[\phi_{m,\pi(m)}^{\boldsymbol{h}}] = \mathbb{E}_{\boldsymbol{H}}[|\langle \boldsymbol{h}_m, \boldsymbol{w}_{\pi(m)}\rangle|^2]$ is harder to obtain; we resort to bounding it. Suppose that without loss of generality $\mu_1 \geq \mu_2 \geq ... \geq \mu_{n_t}$. Since, $\boldsymbol{h}_m$ is isotropically distributed on $\mathbb{S}^{n_t-1}$ we can chose $\mathcal{C}$ as the columns the standard ONB in $\mathbb{C}^{n_t}$. Now a reasonable (but clearly suboptimal) choice of π can be obtained by processing the users in a greedy fashion, i.e. user m gets the beamformer

$$\pi(m) = \arg\max_{j\in\mathcal{J}(m)} \left|\langle \boldsymbol{h}_m, \boldsymbol{w}_j\rangle\right|^2,$$

where we defined the set of possible beamformer indeces as $\mathcal{J}(m) = \mathcal{S}\setminus\{\pi(1),\pi(2),\ldots,\pi(m-1)\}$. Consider now a fixed m. Then, in order to get the image of $\pi(m)$ we will have n_t-m+1 degrees of freedom due to our assumption of non-random channel gains.

In order to get an explicit expression for $\mathbb{E}_{\boldsymbol{H}}[|\langle \boldsymbol{h}_m, \boldsymbol{w}_{\pi(m)}\rangle|^2]$ we will require the order statistics of elements of vectors uniformly distributed on the unit sphere $\mathbb{C}^{n_t}$. More specifically we need expectation of $\max_{j\in\mathcal{J}} |\hat{\boldsymbol{h}}_{m,j}|^2$, where $\hat{\boldsymbol{h}}_m \in \mathbb{C}^n$ and $\mathcal{J} \subseteq \{1,2,\ldots,n_t\}$ with $k = |\mathcal{J}| \leq n_t$. From standard results we have the inequalities

$$\Pr\left(\max_{j\in\mathcal{J}(m)} |\hat{\boldsymbol{h}}_{m,j}|^2 \leq \sqrt{(1-\epsilon)\log\left(|\mathcal{J}(m)|\right)}\right) \leq e^{-|\mathcal{J}(m)|^\epsilon},$$

$$\Pr\left(\|\hat{\boldsymbol{h}}_m\|_2 \geq (1+\epsilon)\sqrt{n_t}\right) \leq \frac{1}{\epsilon n_t},$$

where $\epsilon > 0$, so that we arrive at the following expression for the expectation

$$\mathbb{E}_{\boldsymbol{H}}\left[\phi_{m,\pi(m)}^{\boldsymbol{h}}\right] \geq \frac{(1-\epsilon)\log(n_t-m+1)}{(1+\epsilon)^2 n_t}\left(1-e^{-(n_t-m+1)^\epsilon} - \frac{1}{\epsilon n_t}\right).$$

which proves the claim. □

APPENDIX B
PROOF OF LEMMA 1

Proof: Using $\mathcal{U} = \mathcal{S} = \{1, 2, \ldots, n_t\}$ we get

$$r_m(\pi, \lambda_m \boldsymbol{h}_m) - r_m(\pi, \vartheta_m \boldsymbol{\nu}_m) = \log\left(\frac{(\lambda_m^2 + 1)\left(1 + \vartheta_m^2\left(1 - \phi^{\boldsymbol{\nu}}_{m,\pi(m)}\right)\right)}{(\vartheta_m^2 + 1)\left(1 + \lambda_m^2\left(1 - \phi^{\boldsymbol{h}}_{m,\pi(m)}\right)\right)}\right)$$

$$= \log\left(\frac{1 - \tilde{\vartheta}_m \phi^{\boldsymbol{\nu}}_{m,\pi(m)}}{1 - \tilde{\lambda}_m \phi^{\boldsymbol{h}}_{m,\pi(m)}}\right) = \log\left(1 + \frac{\tilde{\lambda}_m \phi^{\boldsymbol{h}}_{m,\pi(m)} - \tilde{\vartheta}_m \phi^{\boldsymbol{\nu}}_{m,\pi(m)}}{1 - \tilde{\lambda}_m \phi^{\boldsymbol{h}}_{m,\pi(m)}}\right).$$

Here, we have set

$$\tilde{\vartheta}_m := \frac{\vartheta_m^2}{\vartheta_m^2 + 1}, \quad \tilde{\lambda}_m := \frac{\lambda_m^2}{\lambda_m^2 + 1}.$$

Similarly, the negative term can be bounded from above by

$$-\log\left(\frac{1 - \tilde{\vartheta}_m \phi^{\boldsymbol{\nu}}_{m,\pi(m)}}{1 - \tilde{\lambda}_m \phi^{\boldsymbol{h}}_{m,\pi(m)}}\right) \leq \log\left(1 + \frac{\tilde{\vartheta}_m \phi^{\boldsymbol{\nu}}_{m,\pi(m)} - \tilde{\lambda}_m \phi^{\boldsymbol{h}}_{m,\pi(m)}}{1 - \tilde{\vartheta}_m \phi^{\boldsymbol{\nu}}_{m,\pi(m)}}\right).$$

Hence, we have

$$r_m(\pi, \lambda_m \boldsymbol{h}_m) - r_m(\pi, \vartheta_m \boldsymbol{\nu}_m) \leq \max_\pi \log\left(1 + \frac{\tilde{\lambda}_m \phi^{\boldsymbol{h}}_{m,\pi(m)} - \tilde{\vartheta}_m \phi^{\boldsymbol{\nu}}_{m,\pi(m)}}{1 - \tilde{\lambda}_m \phi^{\boldsymbol{h}}_{m,\pi(m)}}\right) \tag{20}$$

$$+ \max_\pi \log\left(1 + \frac{\tilde{\vartheta}_m \phi^{\boldsymbol{\nu}}_{m,\pi(m)} - \tilde{\lambda}_m \phi^{\boldsymbol{h}}_{m,\pi(m)}}{1 - \tilde{\vartheta}_m \phi^{\boldsymbol{\nu}}_{m,\pi(m)}}\right) \tag{21}$$

for any $\boldsymbol{\nu} \in \mathcal{V}$. Assume the following strategy: to minimize (20) take $\boldsymbol{\nu}_m \in \mathcal{V}$ such that, if $|\langle \boldsymbol{h}_m, \boldsymbol{w}^* \rangle|^2$ is maximized for $\boldsymbol{w}^*$, then $|\langle \boldsymbol{\nu}_m, \boldsymbol{w}^* \rangle|^2 \geq |\langle \boldsymbol{h}_m, \boldsymbol{w}^* \rangle|^2$. Since $\mathcal{C} \subseteq \mathcal{V}$ such vector always exists. By this strategy we can set

$$\frac{\lambda_m^2 |\langle \boldsymbol{h}_m, \boldsymbol{w}^* \rangle|^2}{1 + \lambda_m^2 \left(1 - |\langle \boldsymbol{h}_m, \boldsymbol{w}^* \rangle|^2\right)} = \frac{\vartheta_m^2 |\langle \boldsymbol{\nu}_m, \boldsymbol{w}^* \rangle|^2}{1 + \vartheta_m^2 \left(1 - |\langle \boldsymbol{\nu}_m, \boldsymbol{w}^* \rangle|^2\right)},$$

and we get after some calculations

$$\tilde{\vartheta}_m = \frac{\lambda_m^2}{\lambda_m^2 + 1} \frac{|\langle \boldsymbol{h}_m, \boldsymbol{w}^* \rangle|^2}{|\langle \boldsymbol{\nu}_m, \boldsymbol{w}^* \rangle|^2} = \tilde{\lambda}_m \frac{|\langle \boldsymbol{h}_m, \boldsymbol{w}^* \rangle|^2}{|\langle \boldsymbol{\nu}_m, \boldsymbol{w}^* \rangle|^2}$$

and finally

$$\max_\pi \log\left(1 + \frac{\tilde{\lambda}_m \phi^{\boldsymbol{h}}_{m,\pi(m)} - \tilde{\vartheta}_m \phi^{\boldsymbol{\nu}}_{m,\pi(m)}}{1 - \tilde{\lambda}_m \phi^{\boldsymbol{h}}_{m,\pi(m)}}\right) = \max_{\boldsymbol{w} \neq \boldsymbol{w}^*} \log\left(1 + \frac{\tilde{\lambda}_m |\langle \boldsymbol{h}_m, \boldsymbol{w} \rangle|^2 - \tilde{\lambda}_m \frac{|\langle \boldsymbol{h}_m, \boldsymbol{w}^* \rangle|^2}{|\langle \boldsymbol{\nu}_m, \boldsymbol{w}^* \rangle|^2} |\langle \boldsymbol{\nu}_m, \boldsymbol{w} \rangle|^2}{1 - \tilde{\lambda}_m |\langle \boldsymbol{h}_m, \boldsymbol{w} \rangle|^2}\right)$$

$$\leq \max_{\boldsymbol{w} \neq \boldsymbol{w}^*} \log\left(1 + \tilde{\lambda}_m\left(\frac{\left||\langle \boldsymbol{h}_m, \boldsymbol{w} \rangle|^2 - |\langle \boldsymbol{\nu}_m, \boldsymbol{w} \rangle|^2\right|}{1 - \tilde{\lambda}_m |\langle \boldsymbol{h}_m, \boldsymbol{w} \rangle|^2} + \frac{\langle \boldsymbol{\nu}_m, \boldsymbol{w} \rangle}{\langle \boldsymbol{\nu}_m, \boldsymbol{w}^* \rangle} \frac{\left||\langle \boldsymbol{\nu}, \boldsymbol{w}^* \rangle|^2 - |\langle \boldsymbol{h}_m, \boldsymbol{w}^* \rangle|^2\right|}{1 - \tilde{\lambda}_m \langle \boldsymbol{h}_m, \boldsymbol{w} \rangle}\right)\right).$$

Similar, for (21) we obtain

$$\max_{\pi}\log\left(1+\frac{\tilde{\vartheta}_m\phi^{\boldsymbol{\nu}}_{m,\pi(m)}-\tilde{\lambda}_m\phi^{\boldsymbol{h}}_{m,\pi(m)}}{1-\tilde{\vartheta}_m\phi^{\boldsymbol{\nu}}_{m,\pi(m)}}\right)\leq\max_{\boldsymbol{w}\neq\boldsymbol{w}^*}\log\left(1+\frac{\tilde{\lambda}_m\frac{|\langle\boldsymbol{h}_m,\boldsymbol{w}^*\rangle|^2}{|\langle\boldsymbol{\nu}_m,\boldsymbol{w}^*\rangle|^2}|\langle\boldsymbol{\nu}_m,\boldsymbol{w}\rangle|^2-\tilde{\lambda}_m|\langle\boldsymbol{h}_m,\boldsymbol{w}\rangle|^2}{1-\tilde{\lambda}_m\frac{|\langle\boldsymbol{h}_m,\boldsymbol{w}^*\rangle|^2}{|\langle\boldsymbol{\nu}_m,\boldsymbol{w}^*\rangle|^2}|\langle\boldsymbol{\nu}_m,\boldsymbol{w}\rangle|^2}\right)$$

$$\leq\max_{\boldsymbol{w}\neq\boldsymbol{w}^*}\log\left(1+\tilde{\lambda}_m\left(\frac{||\langle\boldsymbol{h}_m,\boldsymbol{w}\rangle|^2-|\langle\boldsymbol{\nu}_m,\boldsymbol{w}\rangle|^2|}{1-\tilde{\lambda}_m|\langle\boldsymbol{\nu}_m,\boldsymbol{w}\rangle|^2}+\frac{\langle\boldsymbol{\nu}_m,\boldsymbol{w}\rangle}{\langle\boldsymbol{\nu}_m,\boldsymbol{w}^*\rangle}\frac{||\langle\boldsymbol{\nu},\boldsymbol{w}^*\rangle|^2-|\langle\boldsymbol{h}_m,\boldsymbol{w}^*\rangle|^2|}{1-\tilde{\lambda}_m|\langle\boldsymbol{\nu}_m,\boldsymbol{w}\rangle|^2}\right)\right)$$

and since $|\langle\boldsymbol{h}_m,\boldsymbol{w}^*\rangle|^2\leq|\langle\boldsymbol{\nu},\boldsymbol{w}^*\rangle|^2$ we have $|\langle\boldsymbol{\nu}_m,\boldsymbol{w}\rangle|^2<1\ \forall\boldsymbol{w}\in\mathcal{C},\boldsymbol{w}\neq\boldsymbol{w}^*$ which proves the claim. □

APPENDIX C
PROOF OF LEMMA 2

Proof: According to (14), RA aims on minimizing $\max_{\pi}|r_m(\pi,\lambda_m\boldsymbol{h}_m)-r_m(\pi,\vartheta\boldsymbol{\nu})|$ over the elements of $\mathcal{V}$. Therefore, from (20) and (21) and Jensen's inequality we have

$$\Delta R_{\text{RA}}\leq\sum_{m=1}^{n_t}\log\left(1+\mathbb{E}_{\boldsymbol{H}}\left[\min_{\substack{1>\tilde{\vartheta}_m>0\\ \boldsymbol{\nu}\in\mathcal{V}}}\max_{\pi}\frac{\left|\tilde{\lambda}_m\phi^{\boldsymbol{h}}_{m,\pi(m)}-\tilde{\vartheta}_m\phi^{\boldsymbol{\nu}}_{m,\pi(m)}\right|}{1-\tilde{\lambda}_m\phi^{\boldsymbol{h}}_{m,\pi(m)}}\right]\right)$$
$$+\sum_{m\in\mathcal{S}}\log\left(1+\mathbb{E}_{\boldsymbol{H}}\left[\min_{\substack{1>\tilde{\vartheta}_m>0\\ \boldsymbol{\nu}\in\mathcal{V}}}\max_{\pi}\frac{\left|\tilde{\lambda}_m\phi^{\boldsymbol{h}}_{m,\pi(m)}-\tilde{\vartheta}_m\phi^{\boldsymbol{\nu}}_{m,\pi(m)}\right|}{1-\tilde{\vartheta}_m\phi^{\boldsymbol{\nu}}_{m,\pi(m)}}\right]\right).\qquad(22)$$

Let us re-write the first term on the right side of (22). We first exploit that whenever $\max_{\pi}|\phi^{\boldsymbol{h}}_{m,\pi(m)}|\geq 1-\epsilon,\epsilon\leq 0.5$, then by Lemma 1 the error can be uniformly bounded from above by

$$\frac{\tilde{\lambda}_m\epsilon}{1-\tilde{\lambda}_m\epsilon}=\frac{\lambda_m^2\epsilon}{1+\lambda_m^2(1-\epsilon)}\leq\frac{\lambda_m^2\epsilon}{1+\lambda_m^2\epsilon},$$

and since clearly $\max_{\pi}|\phi^{\boldsymbol{h}}_{m,\pi(m)}|\geq\frac{1}{n_t}$ and $(1-\epsilon)\leq\frac{\epsilon}{n_t-1}$ for $\epsilon\leq 1-\frac{1}{n_t}$ we have for $\max_{\pi}|\phi^{\boldsymbol{h}}_{m,\pi(m)}|\geq\max(0,1-\epsilon)$

$$\frac{\tilde{\lambda}_m\epsilon}{1-\tilde{\lambda}_m\epsilon}\leq\frac{\lambda_m^2\epsilon}{1+\lambda_m^2\frac{\epsilon}{n_t-1}}=\frac{\tilde{\lambda}_m\epsilon}{1+\tilde{\lambda}_m\left(\frac{\epsilon}{n_t-1}-1\right)},$$

for any $\epsilon>0$ (even that for $\epsilon>1$). On the other hand, we have for $\max_{\pi}|\phi^{\boldsymbol{h}}_{m,\pi(m)}|<\max(0,1-\epsilon)$

$$\frac{\left|\tilde{\lambda}_m\phi^{\boldsymbol{h}}_{m,\pi(m)}-\tilde{\vartheta}_m\phi^{\boldsymbol{\nu}}_{m,\pi(m)}\right|}{1-\tilde{\lambda}_m\phi^{\boldsymbol{h}}_{m,\pi(m)}}\leq\frac{\left|\tilde{\lambda}_m\phi^{\boldsymbol{h}}_{m,\pi(m)}-\tilde{\vartheta}_m\phi^{\boldsymbol{\nu}}_{m,\pi(m)}\right|}{1-\tilde{\lambda}_m+\tilde{\lambda}_m\epsilon}$$
$$\leq\frac{\left|\tilde{\lambda}_m\phi^{\boldsymbol{h}}_{m,\pi(m)}-\tilde{\vartheta}_m\phi^{\boldsymbol{\nu}}_{m,\pi(m)}\right|}{1+\tilde{\lambda}_m\left(\frac{\epsilon}{n_t-1}-1\right)},$$

Hence, we can write for some pair $\boldsymbol{h}_m, \boldsymbol{\nu}_m$

$$\frac{\left|\tilde{\lambda}_m \phi^{\boldsymbol{h}}_{m,\pi(m)} - \tilde{\vartheta}_m \phi^{\boldsymbol{\nu}}_{m,\pi(m)}\right|}{1 - \tilde{\lambda}_m \phi^{\boldsymbol{h}}_{m,\pi(m)}} \leq \frac{\max\left\{\tilde{\lambda}_m \epsilon, \left|\tilde{\lambda}_m \phi^{\boldsymbol{h}}_{m,\pi(m)} - \tilde{\vartheta}_m \phi^{\boldsymbol{\nu}}_{m,\pi(m)}\right|\right\}}{1 + \tilde{\lambda}_m \left(\frac{\epsilon}{n_t - 1} - 1\right)}$$

and setting

$$\tilde{\lambda}_m \epsilon = \min_{\substack{1 > \tilde{\vartheta}_m > 0 \\ \boldsymbol{\nu} \in \mathcal{V}}} \max_{\pi} \left|\tilde{\lambda}_m \phi^{\boldsymbol{h}}_{m,\pi(m)} - \tilde{\vartheta}_m \phi^{\boldsymbol{\nu}}_{m,\pi(m)}\right| = \left|\tilde{\lambda}_m \phi^{\boldsymbol{h}}_{m,\pi^*(m)} - \tilde{\vartheta}^*_m \phi^{\boldsymbol{\nu}^*}_{m,\pi^*(m)}\right|$$

yields

$$\min_{\substack{1 > \tilde{\vartheta}_m > 0 \\ \boldsymbol{\nu} \in \mathcal{V}}} \max_{\pi} \frac{\left|\tilde{\lambda}_m \phi^{\boldsymbol{h}}_{m,\pi(m)} - \tilde{\vartheta}_m \phi^{\boldsymbol{\nu}}_{m,\pi(m)}\right|}{1 - \tilde{\lambda}_m \phi^{\boldsymbol{h}}_{m,\pi(m)}} \leq \frac{\left|\tilde{\lambda}_m \phi^{\boldsymbol{h}}_{m,\pi^*(m)} - \tilde{\vartheta}^*_m \phi^{\boldsymbol{\nu}^*}_{m,\pi^*(m)}\right|}{1 - \tilde{\lambda}_m + \frac{1}{n_t - 1}\left|\tilde{\lambda}_m \phi^{\boldsymbol{h}}_{m,\pi^*(m)} - \tilde{\vartheta}^*_m \phi^{\boldsymbol{\nu}^*}_{m,\pi^*(m)}\right|}.$$

Equivalently, for the second term on the right side of (22) we have

$$\begin{aligned}\frac{\left|\tilde{\lambda}_m \phi^{\boldsymbol{h}}_{m,\pi(m)} - \tilde{\vartheta}_m \phi^{\boldsymbol{\nu}}_{m,\pi(m)}\right|}{1 - \tilde{\vartheta}_m \phi^{\boldsymbol{\nu}}_{m,\pi(m)}} &\leq \frac{\max\left\{\tilde{\lambda}_m \epsilon, \left|\tilde{\lambda}_m \phi^{\boldsymbol{h}}_{m,\pi(m)} - \tilde{\vartheta}_m \phi^{\boldsymbol{\nu}}_{m,\pi(m)}\right|\right\}}{1 - \tilde{\lambda}_m \phi^{\boldsymbol{h}}_{m,\pi(m)} + \tilde{\lambda}_m \phi^{\boldsymbol{h}}_{m,\pi(m)} - \tilde{\vartheta}_m \phi^{\boldsymbol{\nu}}_{m,\pi(m)}} \\ &\leq \frac{\max\left\{\tilde{\lambda}_m \epsilon, \left|\tilde{\lambda}_m \phi^{\boldsymbol{h}}_{m,\pi(m)} - \tilde{\vartheta}_m \phi^{\boldsymbol{\nu}}_{m,\pi(m)}\right|\right\}}{1 + \tilde{\lambda}_m \left(\frac{\max\left\{\epsilon - \left|\tilde{\lambda}_m \phi^{\boldsymbol{h}}_{m,\pi(m)} - \tilde{\vartheta}_m \phi^{\boldsymbol{\nu}}_{m,\pi(m)}\right|, 0\right\}}{n_t - 1} - 1\right)}.\end{aligned}$$

Setting $\epsilon = (1 + \epsilon') \left|\tilde{\lambda}_m \phi^{\boldsymbol{h}}_{m,\pi^*(m)} - \tilde{\vartheta}^*_m \phi^{\boldsymbol{\nu}^*}_{m,\pi^*(m)}\right|$, $\epsilon' > 0$, since the error term is still increasing in ϵ, yields

$$\min_{\substack{1 > \tilde{\vartheta}_m > 0 \\ \boldsymbol{\nu} \in \mathcal{V}}} \max_{\pi} \frac{\left|\tilde{\lambda}_m \phi^{\boldsymbol{h}}_{m,\pi(m)} - \tilde{\vartheta}_m \phi^{\boldsymbol{\nu}}_{m,\pi(m)}\right|}{1 - \tilde{\vartheta}_m \phi^{\boldsymbol{\nu}}_{m,\pi(m)}} \leq \frac{(1 + \epsilon')\left|\tilde{\lambda}_m \phi^{\boldsymbol{h}}_{m,\pi^*(m)} - \tilde{\vartheta}^*_m \phi^{\boldsymbol{\nu}^*}_{m,\pi^*(m)}\right|}{1 - \tilde{\lambda}_m + \frac{\epsilon'}{n_t - 1}\left|\tilde{\lambda}_m \phi^{\boldsymbol{h}}_{m,\pi^*(m)} - \tilde{\vartheta}^*_m \phi^{\boldsymbol{\nu}^*}_{m,\pi^*(m)}\right|},$$

Finally, expanding the fraction with $(1 - \tilde{\lambda}_m)$ and applying Jensen's inequality again proves the claim. □

Appendix D
Proof of Lemma 3

Proof: Without loss of generality, we assume the unitary transmit codebook $\mathcal{C}$ is given by the standard ONB and define the vectors

$$\boldsymbol{\phi}^{\boldsymbol{h}}_m = \left(\phi^{\boldsymbol{h}}_{m,\pi(1)}, \ldots, \phi^{\boldsymbol{h}}_{m,\pi(n_t)}\right)^T,$$

that define points on the (n_t-1)–simplex $\mathcal{K}_{n_t-1} = \{\boldsymbol{x} \in \mathbb{C}^{n_t-1} : x > 0 \text{ and } \|\boldsymbol{x}\|_1 = 1\}$ with edge length $\sqrt{2}$. In a similar manner for each element of the feedback codebook $\boldsymbol{\nu}_i \in \mathcal{V}$ we can define

$$\boldsymbol{q}_i = \boldsymbol{\phi}_m^{\boldsymbol{\nu}_i} = \left(\phi_{m,\pi(1)}^{\boldsymbol{\nu}_i}, \dots, \phi_{m,\pi(n_t)}^{\boldsymbol{\nu}_i}\right)^T,$$

which gives $N = 2^B = |\mathcal{V}|$ points $\mathcal{Q} = \{\boldsymbol{q}_1, \boldsymbol{q}_2, \dots, \boldsymbol{q}_N\}$ on the (n_t-1)–simplex. Using this notation and $\tilde{\vartheta}_m = \tilde{\lambda}_m$, $D_m(B)$ can be written as

$$D_m(B) = \tilde{\lambda}_m \min_{\boldsymbol{q}\in\mathcal{Q}} \|\boldsymbol{\phi}_m^{\boldsymbol{h}} - \boldsymbol{q}\|_\infty,$$

which can be upperbounded by

$$D_m(B) \le \max_{\boldsymbol{x}\in\mathcal{K}_{n_t-1}} \min_{\boldsymbol{q}\in\mathcal{Q}} \|\boldsymbol{x} - \boldsymbol{q}\|_\infty =: \delta.$$

Let $d := n_t - 1$. The idea of the proof is the following. The intersection of the balls $\mathcal{B}_\infty^{n_t}$ with radius δ and the center in $\mathcal{K}^d$ is a polytope with $2d$ facets. The spheres $\delta\mathcal{B}_2^d$ of radius δ are inscribed in this polytope. Next we bound the number of spheres $\delta\mathcal{B}_2^d$ required to cover the simplex $\mathcal{K}^d$. This number can be given by the covering number $N(\mathcal{K}, \delta\mathcal{B}_2^d)$, which can be bounded from above as follows. Using the Rogers-Zong Lemma [18], which states that the covering number $N(\mathcal{A}, \mathcal{B})$, that is, the number of convex bodies $\mathcal{B}$ required to cover a convex body $\mathcal{A}$, can be upper bounded by

$$N(\mathcal{A}, \mathcal{B}) \le \Theta(\mathcal{B}) \frac{\text{vol}(\mathcal{A} - \mathcal{B})}{\text{vol}(\mathcal{B})}, \tag{23}$$

where $\Theta(\mathcal{B}) \ge 1$ is the covering density of $\mathcal{B}$; if $\mathbb{R}^d$ can be tiled by translates of $\mathcal{B}$ then $\Theta(\mathcal{B}) = 1$; if the covering has some overlap then $\Theta(\mathcal{B}) > 1$. Further, we require the Rogers-Shephard inequality [19], which states that

$$\text{vol}(\mathcal{A} - \mathcal{B})\text{vol}(\mathcal{A} \cap \mathcal{B}) \le \binom{2d}{d} \text{vol}(\mathcal{A})\text{vol}(\mathcal{B}). \tag{24}$$

Using the assumption $\text{vol}(\mathcal{A} \cap \mathcal{B}) = \text{vol}(\mathcal{B})$ we get from (23) and (24) that the covering number $N(\mathcal{A}, \mathcal{B})$ is upper bounded by

$$N(\mathcal{A}, \mathcal{B}) \le \Theta(\mathcal{B}) \binom{2d}{d} \frac{\text{vol}(\mathcal{A})}{\text{vol}(\mathcal{B})}.$$

The volumes of the d–simplex $\mathcal{K}^d$ and the scaled ℓ_2–ball $\delta\mathcal{B}_2^d$ are

$$\text{vol}(\mathcal{K}^d) = \frac{\sqrt{d+1}}{d!} \text{ and } \text{vol}(\delta\mathcal{B}_2^d) = \frac{\pi^{d/2}}{\Gamma(1+n/2)} \delta^d,$$

where $\Gamma(\cdot)$ is the gamma function. Wrapping up, the covering number can be upperbounded by

$$
\begin{aligned}
N(\mathcal{K}^d, \delta\mathcal{B}_\infty^d) &\leq N(\mathcal{K}^d, \delta\mathcal{B}_2^d) = N(\frac{1}{\delta}\mathcal{K}^d, \mathcal{B}_2^d) \\
&\leq \Theta(\mathcal{B}_2^d)\binom{2d}{d}\frac{\Gamma(1+d/2)\sqrt{d+1}}{d!\pi^{d/2}} \cdot \frac{1}{\delta^d}.
\end{aligned}
$$

Solving for δ, i.e.

$$
\delta \leq \left(\Theta(\mathcal{B}_2^d)\binom{2n_t-2}{n_t-1}\frac{\Gamma(1+\frac{n_t-1}{2})\sqrt{n_t}}{(n_t-1)!\pi^{\frac{n_t-1}{2}}}\right)^{\frac{1}{n_t-1}} 2^{-\frac{B}{n_t-1}},
$$

proves the inequality. The inequality is valid provided δ is smaller than the inradius of the inscribed circle of the simplex. According to Klamkin [20] for a regular simplex the inradius equals the circumradius divided by n_t-1. The circumradius is easily shown by the volume ratio and Stirlings formula to be greater than $\sqrt{n_t-1}$. This together with the first inequality yields the lower bound on B.

□

Appendix E
Proof of Theorem 2

Proof: The terms of the sum in (12) can be bounded from above as follows.

$$
\begin{aligned}
&r_m(\pi, \lambda_m \boldsymbol{h}_m) - r_m(\pi, \vartheta_m \boldsymbol{\nu}_m) \\
&= \log\left(\frac{\frac{|\mathcal{S}|}{n_t} + \lambda_m^2 \sum_{l\in\mathcal{S}} \phi_{m,\pi(l)}^{\boldsymbol{h}}}{\frac{|\mathcal{S}|}{n_t} + \lambda_m^2 \sum_{l\in\mathcal{S}\setminus\{m\}} \phi_{m,\pi(l)}^{\boldsymbol{h}}}\right) - \log\left(\frac{\frac{|\mathcal{S}|}{n_t} + \vartheta_m^2 \sum_{l\in\mathcal{S}} \phi_{m,\pi(l)}^{\boldsymbol{\nu}}}{\frac{|\mathcal{S}|}{n_t} + \vartheta_m^2 \sum_{l\in\mathcal{S}\setminus\{m\}} \phi_{m,\pi(l)}^{\boldsymbol{\nu}}}\right) \\
&= \log\left(\frac{\frac{|\mathcal{S}|}{n_t} + \lambda_m^2 \sum_{l\in\mathcal{S}} \phi_{m,\pi(l)}^{\boldsymbol{h}}}{\frac{|\mathcal{S}|}{n_t} + \vartheta_m^2 \sum_{l\in\mathcal{S}} \phi_{m,\pi(l)}^{\boldsymbol{\nu}}}\right) + \log\left(\frac{\frac{|\mathcal{S}|}{n_t} + \vartheta_m^2 \sum_{l\in\mathcal{S}\setminus\{m\}} \phi_{m,\pi(l)}^{\boldsymbol{\nu}}}{\frac{|\mathcal{S}|}{n_t} + \lambda_m^2 \sum_{l\in\mathcal{S}\setminus\{m\}} \phi_{m,\pi(l)}^{\boldsymbol{h}}}\right).
\end{aligned}
$$

Setting $\vartheta_m^2 = \lambda_m^2$ we get

$$
\begin{aligned}
& r_m(\pi, \lambda_m \boldsymbol{h}_m) - r_m(\pi, \lambda_m \boldsymbol{\nu}_m) \\
&= \log\left(1 + \frac{n_t \lambda_m^2}{|\mathcal{S}|} \frac{\sum_{l \in \mathcal{S}} \phi^{\boldsymbol{h}}_{m,\pi(l)} - \phi^{\boldsymbol{\nu}}_{m,\pi(l)}}{1 + \lambda_m^2 \sum_{l \in \mathcal{S}} \phi^{\boldsymbol{\nu}}_{m,\pi(l)}}\right) + \log\left(1 + \frac{n_t \lambda_m^2}{|\mathcal{S}|} \frac{\sum_{l \in \mathcal{S} \setminus \{m\}} \phi^{\boldsymbol{\nu}}_{m,\pi(l)} - \phi^{\boldsymbol{h}}_{m,\pi(l)}}{1 + \lambda_m^2 \sum_{l \in \mathcal{S} \setminus \{m\}} \phi^{\boldsymbol{h}}_{m,\pi(l)}}\right) \\
&\leq \log\left(1 + \frac{n_t \lambda_m^2}{|\mathcal{S}|} \left| \sum_{l \in \mathcal{S}} \phi^{\boldsymbol{h}}_{m,\pi(l)} - \phi^{\boldsymbol{\nu}}_{m,\pi(l)} \right| \right) + \log\left(1 + \frac{n_t \lambda_m^2}{|\mathcal{S}|} \left| \sum_{l \in \mathcal{S} \setminus \{m\}} \phi^{\boldsymbol{h}}_{m,\pi(l)} - \phi^{\boldsymbol{\nu}}_{m,\pi(l)} \right| \right) \\
&\leq \log\left(1 + \frac{n_t \lambda_m^2}{|\mathcal{S}|} |\mathcal{S}| \max_{\pi} \left| \phi^{\boldsymbol{h}}_{m,\pi(m)} - \phi^{\boldsymbol{\nu}}_{m,\pi(m)} \right| \right) + \log\left(1 + \frac{n_t \lambda_m^2}{|\mathcal{S}|} |\mathcal{S} \setminus \{m\}| \max_{\pi} \left| \phi^{\boldsymbol{h}}_{m,\pi(m)} - \phi^{\boldsymbol{\nu}}_{m,\pi(m)} \right| \right) \\
&\leq 2 \log\left(1 + n_t \lambda_m^2 \max_{\pi} \left| \phi^{\boldsymbol{h}}_{m,\pi(m)} - \phi^{\boldsymbol{\nu}}_{m,\pi(m)} \right| \right) \\
&= 2 \log\left(1 + \frac{P \mu_m^2}{\sigma^2} \max_{\pi} \left| \phi^{\boldsymbol{h}}_{m,\pi(m)} - \phi^{\boldsymbol{\nu}}_{m,\pi(m)} \right| \right).
\end{aligned}
$$

The lower bound on $-(r_m(\pi, \lambda_m \boldsymbol{h}_m) - r_m(\pi, \vartheta_m \boldsymbol{\nu}_m))$ can be obtained in a similar manner. Taking expectations and using Jensen's inequality we obtain

$$
\mathbb{E}_{\boldsymbol{H}}\left[r_m(\pi, \lambda_m \boldsymbol{h}_m) - r_m(\pi, \lambda_m \boldsymbol{\nu}_m)\right] \leq 2 \log\left(1 + \mathbb{E}_{\boldsymbol{H}}\left[\frac{P \mu_m^2}{\sigma^2} \max_{\pi} \left| \phi^{\boldsymbol{h}}_{m,\pi(m)} - \phi^{\boldsymbol{\nu}}_{m,\pi(m)} \right| \right]\right).
$$

Since $\max_{\pi} |\phi^{\boldsymbol{h}}_{m,\pi(m)} - \phi^{\boldsymbol{\nu}}_{m,\pi(m)}|$ depends only on the channel directions $\boldsymbol{h}_m$ it is independent of the channel magnitude μ_m.

$$
\mathbb{E}_{\boldsymbol{H}}\left[r_m(\pi, \lambda_m \boldsymbol{h}_m) - r_m(\pi, \lambda_m \boldsymbol{\nu}_m)\right] \leq 2 \log\left(1 + \frac{P n_t}{\sigma^2} \mathbb{E}_{\boldsymbol{H}}\left[\max_{\pi} \left| \phi^{\boldsymbol{h}}_{m,\pi(m)} - \phi^{\boldsymbol{\nu}}_{m,\pi(m)} \right| \right]\right).
$$

Using the RA scheme and (12) yields the result. □

REFERENCES

[1] H. Weingarten, Y. Steinberg, and S. Shamai, "The Capacity Region of the Gaussian Multiple-Input Multiple-Output Broadcast Channel," *IEEE Transactions on Information Theory*, vol. 52, no. 9, pp. 3936–3964, Sep. 2006.

[2] D. Love, R. Heath, V. N. Lau, D. Gesbert, B. Rao, and M. Andrews, "An Overview of Limited Feedback in Wireless Communication Systems," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 8, pp. 1341–1365, Oct. 2008.

[3] N. Jindal, "MIMO Broadcast Channels With Finite-Rate Feedback," *IEEE Transactions on Information Theory*, vol. 52, no. 11, pp. 5045–5060, Nov. 2006.

[4] K. Kiran Mukkavilli, A. Sabharwal, E. Erkip, and B. Aazhang, "On Beamforming with Finite Rate Feedback in Multiple-Antenna Systems," *IEEE Transactions on Information Theory*, vol. 49, no. 10, pp. 2562–2579, Oct. 2003.

[5] M. Sharif and B. Hassibi, "On the Capacity of MIMO Broadcast Channels With Partial Side Information," *IEEE Transactions on Information Theory*, vol. 51, no. 2, pp. 506–522, Feb. 2005.

[6] C. Au-Yeung and D. Love, "On the Performance of Random Vector Quantization Limited Feedback Beamforming in a MISO System," *IEEE Transactions on Wireless Communications*, vol. 6, no. 2, pp. 458–462, Feb. 2007.

[7] G. Dimic and N. Sidiropoulos, "On Downlink Beamforming with Greedy User Selection: Performance Analysis and a Simple New Algorithm," *IEEE Transactions on Signal Processing*, vol. 53, no. 10, pp. 3857–3868, Oct. 2005.

[8] M. Trivellato, F. Boccardi, and H. Huang, "On Transceiver Design and Channel Quantization for Downlink Multiuser MIMO Systems with Limited Feedback," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 8, pp. 1494–1504, Oct. 2008.

[9] J. Schreck, P. Jung, G. Wunder, M. Ohm, and H.-P. Mayer, "Limited Feedback in Multiuser MIMO OFDM Systems Based on Rate Approximation," in *IEEE Global Telecommunications Conference*, Nov. 2009, pp. 1–6.

[10] T. Yoo, N. Jindal, and A. Goldsmith, "Multi-Antenna Downlink Channels with Limited Feedback and User Selection," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 7, pp. 1478–1491, Sep. 2007.

[11] K. Huang, J. G. Andrews, and R. Heath, "Performance of Orthogonal Beamforming for SDMA With Limited Feedback," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 1, pp. 152–164, 2009.

[12] R. de Francisco, M. Kountouris, D. T. M. Slock, and D. Gesbert, "Orthogonal Linear Beamforming in MIMO Broadcast Channels," in *2007 IEEE Wireless Communications and Networking Conference*, 2007, pp. 1210–1215.

[13] S. A. Jafar and A. J. Goldsmith, "Isotropic Fading Vector Broadcast Channels: The Scalar Upper Bound and Loss in Degrees of Freedom," *IEEE Transactions on Information Theory*, vol. 51, no. 3, pp. 848–857, 2005.

[14] K. Böröczky, *Finite packing and covering*. Cambridge University Press, 2004.

[15] 3GPP TS 36.201, "LTE Physical Layer - General Description," 2009.

[16] D. S. Baum, J. Salo, M. Milojevic, P. Kyösti, and J. Hansen, "MATLAB implementation of the interim channel model for beyond-3G systems (SCME)," 2005.

[17] R. Samsung, "Downlink MIMO for EUTRA," *3GPP TSG RAN WG1 Meeting*, 2006.

[18] C. A. Rogers and C. Zong, "Covering convex bodies by translates of convex bodies," *Mathematika*, vol. 44, no. 01, pp. 215–218, Jun. 1997.

[19] C. Rogers and G. Shephard, "The difference body of a convex body," *Archiv der Mathematik*, vol. 8, no. 3, pp. 220–233, 1957.

[20] M. S. Klamkin and G. A. Tsintsifas, "The Circumradius-Inradius Inequality for a Simplex," *Mathematics Magazine*, vol. 52, no. 1, Jan. 1979.