# Grassmannian Product Codebooks for Limited Feedback Massive MIMO with Two-Tier Precoding

Stefan Schwarz *Member, IEEE*, Markus Rupp *Fellow, IEEE*, Stefan Wesemann

***Abstract*—In frequency-division duplex full-dimension massive MIMO systems, one of the main challenges is to obtain sufficiently accurate channel state information (CSI) at the transmitter to enable efficient multi-user MIMO transmission. In this paper, we propose a novel dual stage Grassmannian product quantization approach that is suitable for high-dimensional CSI quantization and feedback. We apply the proposed method for outer-tier CSI feedback in two-tier precoding architectures, which employ channel-subspace based outer-tier precoding strategies, such as, maximum eigenmode transmission. The proposed method is especially effective if the channel can be decomposed in the angular domain, such that DFT codebooks enable an efficient CSI compression. Our dual stage product codebook quantization approach mitigates the well-known inefficiency of oversampled DFT codebooks for growing codebook sizes, by varying the intermediate quantization dimension of the proposed quantizer.**

***Index Terms*—limited feedback, Grassmannian quantization, massive MIMO, product codebook, FDD, FD-MIMO**

## I. Introduction

Full-dimension massive multiple-input multiple-output (MIMO) is a key fifth generation (5G) technology [1]. When considered in time division duplex (TDD) systems, massive MIMO exploits the channel reciprocity to directly estimate the channel state information at the transmitter (CSIT) [2]. In this paper, however, our focus is on frequency division duplex (FDD) systems, where we assume that the system does not leverage any form of reciprocity for CSIT estimation, i.e., not even reciprocity of the angular scattering function [3]. That is, CSIT can only be obtained by dedicated feedback from the users over signaling channels of limited capacity.

Efficient limited channel state information (CSI) feedback methods that are suitable for small scale MIMO systems, with a number of antennas in the order of ten, have been investigated extensively over the last decade. Efficient codebook designs and quantization metrics for single-user MIMO have been proposed, e.g., in [4], the impact of imperfect CSIT on the achievable transmission rate has been investigated, e.g., in [5–8], and efficient differential/predictive manifold quantizers have been developed in [9–12]. These differential/predictive quantizers exploit temporal channel correlation to enable efficient multi-user MIMO transmission with a minimal feedback overhead. Even though these methods are highly efficient in terms of rate-distortion performance, they are hardly applicable in massive MIMO, because their computational complexity for large-dimensional MIMO systems is currently not feasible in practice. In [13–15] correlated codebook constructions are proposed that exploit spatial correlation of the MIMO channel to enhance the efficiency of CSI quantization. These methods allow to reduce the CSI feedback overhead whenever the MIMO channel exhibits spatial correlation. The approach is so effective that it was included in the dual codebook framework of LTE-Advanced [16, 17].

To reduce the complexity of CSI quantization in massive MIMO, a number of authors have successfully proposed codebook constructions that facilitate computationally efficient CSI quantization. In [18], the authors propose a trellis coded quantization approach for multiple-input single-output systems that achieves a performance close to random vector quantization and allows for efficient quantization by means of the Viterbi algorithm. In [19], the authors propose a compressed sensing based CSI feedback algorithm, that utilizes a $K$ singular value decomposition (SVD) dictionary that enables effective CSI compression of spatially correlated channels. In [20], the authors propose a deep learning based CSI sensing and recovery approach that improves the tradeoff between the achieved compression ratio and the complexity, by directly learning spatial structures combined with time correlation of the channel from training samples of time-varying massive MIMO channels. In [21], the authors exploit the reciprocity of long-term channel parameters, such as, the signal angles of arrival in multi-scattering channels, to substantially reduce the required amount of CSI feedback.

In this paper, we propose a novel approach for reducing the complexity of CSI quantization, by introducing a dual stage Grassmannian product codebook quantization approach. This method allows to split up a high-dimensional Grassmannian quantization problem into two lower-dimensional problems, enabling a significant complexity reduction. The two quantization stages can in principle apply any Grassmannian codebook constructions In our work, we exploit random subspace quantization (RSQ) codebooks to obtain analytic rate-distortion results, as well as, discrete Fourier transform (DFT) codebooks, which are known for their favorable CSI compression capabilities for limited scattering directional channels [22].

In order to limit the required quantization resolution of the high-dimensional Grassmannian quantization problem, we apply a two-tier precoding approach [23, 24]. Specifically, we apply a common user-group-specific outer-tier precoding that maximizes the received signal power of the served users and an inner-tier precoder that mitigates the residual multi-user interference. As we show in this paper, the outer-tier precoder can be based on relatively coarse CSIT without loosing too much signal power, whereas the inner-tier precoder requires highly accurate CSIT to avoid residual multi-user interference caused by CSIT imperfections. However, since inner-tier precoding is based on CSIT about the effective low-dimensional (outer-precoded) channel matrices, inner-tier CSI feedback can

S. Schwarz is with the Christian Doppler Laboratory for Dependable Wireless Connectivity for the Society in Motion, Technische Universitaet Wien, Austria; (e-mail: sschwarz@nt.tuwien.ac.at).

M. Rupp is with the Institute of Telecommunications, Technische Universitaet Wien, Austria; (email: mrupp@nt.tuwien.ac.at)

S. Wesemann is with Nokia Bell Labs, Stuttgart 70435, Germany; (e-mail: stefan.wesemann@nokia-bell-labs.com)

employ the above mentioned well-known efficient CSI feedback methods for small-scale MIMO systems. We therefore focus in this paper on the design of rate-distortion, as well as, computationally efficient CSI quantization methods to provide the required CSIT for outer-tier precoding.

This paper reveals the underlying Grassmannian manifold structure of the quantization problem, which was not utilized in [25], and provides an analytic performance investigation of our product codebook when combined with two-tier precoding.

*Contributions:* The following main contributions are developed throughout this paper:

1) We propose a novel dual stage Grassmannian product codebook quantization approach that enables computationally efficient limited feedback in FDD massive MIMO.
2) We provide an analytic performance investigation of the rate-distortion performance of the dual stage quantization approach, assuming RSQ codebooks are employed within the two quantization stages. The analysis exhibits a small distortion loss of dual stage quantization compared to single stage quantization, while providing a significant reduction of complexity.
3) The proposed dual stage quantization approach does per se not restrict which Grassmannian codebook constructions are utilized in the two quantization stages. For multi-scattering directional channels, we specialize the dual stage quantization approach to employ a DFT codebook in the first quantization stage. This allows to further reduce implementation complexity by a fast Fourier transform (FFT) realization of the DFT codebook.
4) We propose an analytic asymptotic approximation of the achievable transmission rate of two-tier precoding with imperfect outer-tier CSIT for massive MIMO scenarios.
5) We benchmark the proposed product codebook design on real-world channel measurement data, in order to demonstrate its value under realistic channel conditions.

*Organization:* Our paper is organized as follows:

1) In Sec. II, we introduce the considered single-cell downlink multi-user massive MIMO system model, the employed multi-scattering directional channel model, as well as, the applied two-tier precoding architecture.
2) In Sec. III, we provide an overview of the implemented two-tier limited CSI feedback approach, which is suitable for the considered two-tier precoding architecture. We also discuss the relationship to 3GPP 5G feedback methods.
3) In Sec. IV, we consider the outer-tier CSI feedback and propose our novel dual stage Grassmannian quantization approach. We furthermore provide an analytic rate-distortion performance characterization, assuming RSQ codebooks are applied within the two quantization stages.
4) In Sec. V, we propose quantization codebooks for the two quantization stages that enable even further complexity reduction. Specifically, we consider a DFT codebook for the first quantization stage, which can be efficiently implemented by an FFT. For the second quantization stage we consider RSQ, as well as, scalar quantization. We numerically exhibit the rate-distortion performance of these codebooks for multi-scattering directional channels.
5) In Sec. VI, we provide an asymptotic performance investigation of the considered two-tier precoding architecture with perfect inner-tier CSIT and imperfect outer-tier CSIT.
6) In Sec. VII, we apply the proposed methods on real-world channel measurement data for single-cell downlink multi-user MIMO transmission.

We would like to emphasize that our proposed dual stage quantization approach should not be viewed as an alternative/competitor to existing Grassmannian codebook constructions. It is rather a general concept for splitting a high-dimensional Grassmannian quantization problem into two lower-dimensional problems, without sacrificing too much performance while gaining significantly in terms of complexity. The actual quantization codebooks that are applied in the two quantization stages are not restricted; i.e., any of the existing Grassmannian codebook constructions can be employed within these two individual stages.

*Notation::* We denote vectors and matrices by boldface lower- and upper-case letters $\mathbf{x}$ and $\mathbf{X}$. The transpose and conjugate-transpose of matrix $\mathbf{X}$ are $\mathbf{X}^{\mathrm{T}}$ and $\mathbf{X}^{\mathrm{H}}$, the Frobenius norm is $\|\mathbf{X}\|$ and the subspace spanned by vector $\mathbf{x}$ is $\mathrm{span}\,(\mathbf{x})$. We employ the notation $[\mathbf{X}]_{k,\ell}$ to access the element in row $k$ and column $\ell$. We denote the complex Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\mathbf{C}$ as $\mathcal{CN}\,(\boldsymbol{\mu}, \mathbf{C})$, and the expected value of random variable $r$ as $\mathbb{E}\,(r)$. We utilize the symbol $\stackrel{!}{=}$, e.g., $a \stackrel{!}{=} b$, to highlight that the equality is enforced by construction.

## II. System Model

In this section, we describe the considered input-output relationship in Sec. II-A, the employed channel model in Sec. II-B and the applied two-tier precoding architecture in Sec. II-C.

### A. Single-Cell Downlink Multi-User MIMO Transmission

We consider downlink transmission from a single transmitter to $U$ receivers (users). The transmitter is equipped with $N_t$ transmit antennas and the users are equipped with $N_r \ll N_t$ receive antennas. We assume that the transmitter sends $N_s \leq N_r$ data streams in parallel to each user, such that a total of $UN_s \leq N_t$ data streams is spatially multiplexed.

We consider linear multi-user precoding; thus, the input-output relationship of user $u$ is

$$\mathbf{y}_u = \mathbf{H}_u^{\mathrm{H}}\mathbf{F}_u\mathbf{x}_u + \mathbf{H}_u^{\mathrm{H}} \sum_{j=1, j\neq u}^{U} \mathbf{F}_j\mathbf{x}_j + \mathbf{z}_u. \tag{1}$$

Here, $\mathbf{H}_u \in \mathbb{C}^{N_t \times N_r}$ and $\mathbf{z}_u \in \mathbb{C}^{N_r \times 1}$ denote the channel matrix and the noise of user $u$. $\mathbf{F}_j \in \mathbb{C}^{N_t \times N_s}$ is the precoding matrix and $\mathbf{x}_j \in \mathbb{C}^{N_s \times 1}$ are the transmit symbols of user $j$. We assume that $\mathbb{E}\left(\mathbf{x}_j\mathbf{x}_j^{\mathrm{H}}\right) = 1/N_s\,\mathbf{I}_{N_s}$ and $\mathbf{z}_u \sim \mathcal{CN}\left(\mathbf{0}, \sigma_z^2\mathbf{I}_{N_r}\right)$.

We will frequently utilize the compact form SVD of the channel matrix

$$\begin{aligned} \mathbf{H}_u &= \mathbf{U}_u\boldsymbol{\Sigma}_u\mathbf{V}_u^{\mathrm{H}}, \\ \mathbf{U}_u^{\mathrm{H}}\mathbf{U}_u &= \mathbf{I}_{N_r},\ \mathbf{V}_u^{\mathrm{H}}\mathbf{V}_u = \mathbf{I}_{N_r}, \\ \boldsymbol{\Sigma}_u &= \mathrm{Diag}\left(\sigma_u^{(1)}, \ldots, \sigma_u^{(N_r)}\right), \end{aligned} \tag{2}$$

where we assume that the singular values $\sigma_u^{(i)}$ on the main diagonal of $\boldsymbol{\Sigma}_u$ are sorted in decreasing order.

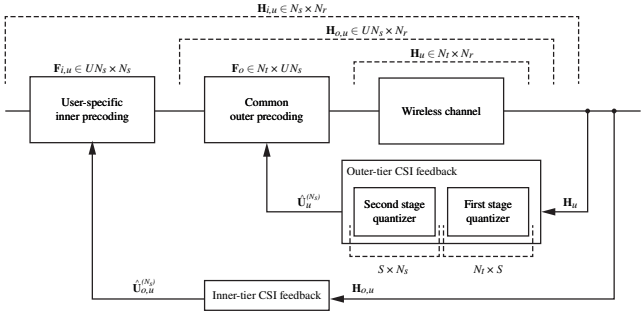



Fig. 1. Considered two-tier precoding and CSI feedback architecture.

### B. Channel Modeling

We consider a narrow-band multi-scattering directional channel model contributing with $N_p$ paths to obtain the channel matrices [26]

$$\mathbf{H}_u = \sqrt{\frac{N_t N_r}{N_p}} \sum_{p=1}^{N_p} \alpha_{p,u} \mathbf{a}_{t,u}\left(\phi_p^{(t,u)}, \theta_p^{(t,u)}\right) \mathbf{a}_{r,u}\left(\phi_p^{(r,u)}, \theta_p^{(r,u)}\right)^{\mathrm{T}}, \tag{3}$$

where $\alpha_{p,u} \in \mathbb{C}$ denotes the complex-valued amplitude of path $p$ and $\mathbf{a}_{t,u}\left(\phi_p^{(t,u)}, \theta_p^{(t,u)}\right)$, $\mathbf{a}_{r,u}\left(\phi_p^{(r,u)}, \theta_p^{(r,u)}\right)$ are the antenna array response vectors evaluated at the respective transmit, receive azimuth and elevation angles $\phi_p^{(t,u)}$, $\phi_p^{(r,u)}$ and $\theta_p^{(t,u)}$, $\theta_p^{(r,u)}$. We consider statistically independent propagation paths $\mathbb{E}\left(\alpha_{p,u}\alpha_{\ell,u}^*\right) = \rho_{p,u}\delta_{p,\ell}$ with power normalization $\sum_{p=1}^{N_p} \rho_{p,u} = 1$. In massive MIMO scenarios, the number of paths $N_p$ is often much smaller than the number of transmit antennas $N_t$, especially at higher carrier frequencies [27, 28]; yet, the considered clustered channel model can also represent Rayleigh fading situations when the number of paths grows large. The rank of the channel matrix obtained from the multi-scattering channel model, and thus the number of streams $N_s$, is upper-bounded by the minimum of $N_p$, $N_t$ and $N_r$.

When employing the multi-scattering directional channel model in our numerical simulations, we assume equally strong paths $\rho_{p,u} = \frac{1}{N_p}$. We furthermore consider a two-dimensional scenario employing a horizontally aligned uniform linear antenna array (ULA) with array response/manifold vector

$$\left[\mathbf{a}_{x,u}^{\mathrm{ULA}}(\phi)\right]_k = \frac{g_e(\phi)}{\sqrt{N_x}} \exp\left(j2\pi d_h (k-1)\sin(\phi)\right), \tag{4}$$

with $x \in \{r, t\}$, $d_h$ denoting the inter antenna-element spacing in multiples of the wavelength $\lambda$, and $g_e(\phi)$ representing the complex-valued antenna-element gain pattern. In our simulations we assume $g_e(\phi) = 1, \forall \phi$, $d_h = \lambda/2$ and consider uniformly distributed angles $\phi \in [-\pi, \pi]$.

### C. Two-Tier Precoding

We assume that the transmitter applies a two-tier precoding approach, as illustrated in Fig. 1, consisting of:

1) A user-group-specific common outer-tier precoder $\mathbf{F}_o \in \mathbb{C}^{N_t \times UN_s}$.
2) An inner-tier precoder $\mathbf{F}_{i,u} \in \mathbb{C}^{UN_s \times N_s}$ for each user.

The total precoder of a user $u$ is then

$$\mathbf{F}_u = \gamma_u \mathbf{F}_o \mathbf{F}_{i,u}. \tag{5}$$

Here, the power normalization factor $\gamma_u$ can be set such that either an instantaneous power constraint $\|\mathbf{F}_u\|^2 = P_u$ or an average power constraint $\mathbb{E}\left(\|\mathbf{F}_u\|^2\right) = P_u$ is satisfied.

#### 1) Outer-tier precoding:

The task of this precoder is to maximize the received signal strength of the users and to reduce the dimensions of the effective channel matrix, in order to simplify the channel estimation, CSI feedback and symbol detection. To maximize the instantaneous received signal power of each user, we assume that the transmitter applies multi-user maximum eigenmode transmission (MET) as outer-tier precoder

$$\mathbf{F}_o = \left[\mathbf{U}_1^{(N_s)}, \mathbf{U}_2^{(N_s)}, \ldots, \mathbf{U}_U^{(N_s)}\right]. \tag{6}$$

Here, $\mathbf{U}_j^{(N_s)} \in \mathbb{C}^{N_t \times N_s}$ denotes the matrix consisting of the $N_s$ left singular vectors $\mathbf{u}_j^{(i)}$ of $\mathbf{U}_j$ that correspond to the largest singular values

$$\mathbf{U}_j^{(N_s)} = \left[\mathbf{u}_j^{(1)}, \ldots, \mathbf{u}_j^{(N_s)}\right]. \tag{7}$$

Alternatively, the MET precoder can also be calculated from statistical CSIT to maximize the long-term average received power. In this case, the matrices $\mathbf{U}_j^{(N_s)}$ are obtained from an eigendecomposition of an estimate of the channel correlation matrix $\mathbb{E}\left(\mathbf{H}_u \mathbf{H}_u^{\mathrm{H}}\right)$. For both cases, the CSI feedback approaches developed in Sec. IV are applicable.

We denote the product of the outer precoder and the channel matrix of user $u$ as the outer precoded channel $\mathbf{H}_{o,u} = \mathbf{F}_o^{\mathrm{H}} \mathbf{H}_u \in \mathbb{C}^{UN_s \times N_r}$.

In a mobile communications system, the scheduler would select a user group to ensure that the matrices $\mathbf{U}_i^{(N_s)}$ of different users are close to orthogonal. In this paper, we do, however, not consider user grouping and scheduling.

#### 2) Inner-tier precoding:

In general, there will be residual multi-user interference after multi-user MET outer-tier precoding; the goal of the inner-tier precoders is to deal with this residual multi-user interference. For this purpose different precoding strategies can be applied, such as zero forcing (ZF) beamforming, block diagonalization (BD) and regularized block diagonalization (RBD) precoding [29, 30]. Our main focus in this paper is on ZF beamforming when transmitting a single stream per user $N_s = 1$, and BD precoding when $N_s > 1$.

The goal of the ZF/BD filters is to keep the $N_s$-dimensional subspace of $\mathbf{H}_{o,j}$ corresponding to its maximum singular values free of interference. To achieve this goal, we apply an SVD $\mathbf{H}_{o,j} = \mathbf{U}_{o,j} \mathbf{\Sigma}_{o,j} \mathbf{V}_{o,j}^{\mathrm{H}}$ and define, similar to (7), the matrix $\mathbf{U}_{o,j}^{(N_s)} \in \mathbb{C}^{UN_s \times N_s}$, which consist of the $N_s$ left singular vectors $\mathbf{u}_{o,j}^{(i)}$ of $\mathbf{U}_{o,j} \in \mathbb{C}^{UN_s \times N_r}$ that correspond to the largest singular values. With this, the set of ZF/BD inner beamformers/precoders can be obtained from the following well-known conditions

$$\left(\mathbf{U}_{o,j}^{(N_s)}\right)^{\mathrm{H}} \mathbf{F}_{i,u} \stackrel{!}{=} \mathbf{0}, \ \forall j \neq u, \tag{8}$$

$$\mathrm{rank}\left(\left(\mathbf{U}_{o,u}^{(N_s)}\right)^{\mathrm{H}} \mathbf{F}_{i,u}\right) = N_s, \ \|\mathbf{F}_{i,u}\|^2 = 1, \ \forall u. \tag{9}$$

## III. Two-Tier Limited Feedback

In this section, we describe the basic concepts of two-tier limited feedback in combination with two-tier precoding. In general, there are a number of possibilities for providing the necessary CSIT for two-tier precoding. The users can, for example, directly quantize and feed back their full channel matrices $\mathbf{H}_u$, which enables calculation of the outer- and inner-tier precoders at the transmitter. This approach, however, is not efficient and implies excessive feedback overhead in massive MIMO systems, due to the large first dimension $N_t$ of $\mathbf{H}_u$.

### A. Outer- and Inner-Tier CSI Feedback

To reduce the required CSI feedback overhead, we utilize the structure of the applied two-tier precoding strategy to facilitate a more efficient limited feedback implementation. Specifically, we consider a two-tier limited feedback approach consisting of an outer- and inner-tier, as illustrated in Fig. 1:

1) Outer-tier CSI feedback about the $N_s$ maximum eigenmodes $\mathbf{U}_j^{(N_s)}$ of the channel matrix $\mathbf{H}_j$: The accuracy of this information determines the achievable signal power gain of the system. As we show further below, it is possible to apply relatively low rate and low resolution quantization on this tier without sacrificing too much signal to noise ratio (SNR).
2) Inner-tier CSI feedback about the $N_s$ maximum eigenmodes $\mathbf{U}_{o,j}^{(N_s)}$ of the outer precoded channel matrix $\mathbf{H}_{o,j}$: This information is critical for mitigating multi-user interference. Its accuracy determines the achievable multiplexing gain and it therefore requires much higher quantization resolution than the outer-tier. However, the dimensions of the quantization problem are much smaller than for the outer-tier CSI feedback and thus the feedback overhead stays manageable.

In this paper, we put the scope on efficient outer-tier CSI feedback design, to deal with the large dimensionality of this quantization problem, as well as, on the impact of the outer-tier CSI feedback accuracy on the achievable transmission rate of the multi-user system.

We do not specifically consider the inner-tier CSI feedback in this paper, since efficient quantization methods for this low-dimensional quantization problem are well known; see for example [4, 9–12, 31–34]. Furthermore, the impact of the CSI feedback accuracy on ZF/BD precoding strategies is well investigated [5–8, 35] and these results are directly applicable to inner-tier precoding. In our related conference paper [36], we demonstrate that the proposed outer-tier precoding can be combined with predictive inner-tier CSI feedback.

### B. Relationship to 3GPP 4G/5G Feedback Approaches

Two-tier precoding is currently also within the scope of 3GPP 4G/5G standardization of full-dimension and massive MIMO systems. In Rel. 13, beamformed CSI reference symbols (RS) (class B) have been introduced to reduce the pilot signaling overhead [37]. In this approach, a vendor specific beamset, a so-called grid-of-beams codebook, is utilized as outer-tier precoder to achieve a beamforming gain and to reduce the dimension of the effective MIMO channel [38]. Essentially, beamformed RS are sent consecutively over all beams of the codebook and the user selects the beam that maximizes the gain of the effective channel.

The grid-of-beams approach is in principle similar to the DFT-based codebook that we consider below for our first quantization stage of the outer-tier feedback; see Sec. V. However, as we show below, such a grid-of-beams is not efficient for the quantization of multiple eigenmodes $\mathbf{U}_j^{(N_s)}$, due to the unit-modulus structure imposed on the individual elements of the codebook. To mitigate this issue, we propose to add a second quantization stage that improves the resolution of the codebook and mitigates the unit-modulus structure.

Within 3GPP Rel. 14, hybrid CSI RS have been introduced, which combine non-beamformed and beamformed RS to enable effective two-tier precoding with limited reference symbol and feedback overhead [39]. In this approach, non-beamformed CSI RS are provided with a relatively low resolution in time and frequency (to keep the overhead low), in order to enable CSI feedback for outer-tier precoding. Additionally, beamformed RS are provided to estimate the effective outer-precoded MIMO channel and to support inner-tier CSI feedback with comparatively high resolution in time and frequency. Until Rel. 15, inner-tier CSI feedback is restricted to $N_s \leq 2$ eigenmodes per user; yet it will be extended to $N_s \leq 4$ within Rel. 16. 3GPP also proposes to apply time-frequency domain compression to the eigenmodes $\mathbf{U}_{o,j}^{(N_s)}$, by applying feedback clustering over multiple physical resource blocks. These methods are compatible to our outer-tier CSI feedback.

## IV. Outer-tier CSI Feedback Design

In this section, we describe the propose outer-tier CSI feedback approach in detail. In Sec. IV-A, we first of all show that Grassmannian quantization is applicable for providing outer-tier CSI feedback in the considered two-tier precoding architecture. We then review well-known results from single stage quantization in Sec. IV-B as a basis for the development of dual stage quantization in Sec. IV-C. Finally, in Sec. IV-D, we provide a rate-distortion performance comparison of single and dual stage quantization.

### A. Applicability of Grassmannian Quantization

The calculation of the outer-tier precoders according to Eq. (6) requires CSIT about the maximum eigenmodes of the users' channel matrices and thus calls for feed back of the matrices $\mathbf{U}_j^{(N_s)}$ by the users.

However, the achievable rate of our two-tier precoding approach is invariant w.r.t. right multiplication of $\mathbf{F}_o$ by any unitary matrix $\mathbf{Q}$. To see this, consider the outer precoded channel obtained after multiplying $\mathbf{F}_o$ by an arbitrary unitary matrix $\mathbf{Q}$, $\tilde{\mathbf{F}}_o = \mathbf{F}_o\mathbf{Q}$

$$\begin{aligned}\tilde{\mathbf{H}}_{o,j} &= \tilde{\mathbf{F}}_o^{\mathrm{H}}\mathbf{H}_j = \left(\mathbf{F}_o\mathbf{Q}\right)^{\mathrm{H}}\mathbf{H}_j \\ &= \mathbf{Q}^{\mathrm{H}}\mathbf{U}_{o,j}\mathbf{\Sigma}_{o,j}\mathbf{V}_{o,j}^{\mathrm{H}} = \tilde{\mathbf{U}}_{o,j}\mathbf{\Sigma}_{o,j}\mathbf{V}_{o,j}^{\mathrm{H}}. \end{aligned} \tag{10}$$

For this modified outer-precoded channel $\tilde{\mathbf{H}}_{o,j}$, the block-diagonalization conditions of the inner precoder $\tilde{\mathbf{F}}_{i,u}$ are

$$\left(\tilde{\mathbf{U}}_{o,j}^{(N_s)}\right)^{\mathrm{H}}\tilde{\mathbf{F}}_{i,u} = \left(\mathbf{U}_{o,j}^{(N_s)}\right)^{\mathrm{H}}\mathbf{Q}\tilde{\mathbf{F}}_{i,u} \stackrel{!}{=} \mathbf{0}, \tag{11}$$

$$\Longrightarrow \tilde{\mathbf{F}}_{i,u} \stackrel{!}{=} \mathbf{Q}^{\mathrm{H}}\mathbf{F}_{i,u}, \tag{12}$$

where $\mathbf{F}_{i,u}$ is the BD precoder of (8), corresponding to $\mathbf{F}_o$. The resulting inner precoder thus compensates for the unitary transformation of $\mathbf{F}_o$ by $\mathbf{Q}$ and thus the achievable rates of both systems (employing $\mathbf{F}_o$ and $\tilde{\mathbf{F}}_o$, resp.) are the same.

This observation holds specifically also for any block-diagonal unitary matrix $\mathbf{Q} = \mathrm{Blkdiag}\left(\mathbf{Q}_1, \ldots, \mathbf{Q}_U\right), \mathbf{Q}_j \in \mathbb{C}^{N_s \times N_s}$. In terms of CSI quantization, this implies that each user can feed back an arbitrarily rotated version of the maximum eigenmodes $\tilde{\mathbf{U}}_j^{(N_s)} = \mathbf{U}_j^{(N_s)}\mathbf{Q}_j$. In fact, each user can actually feed back an arbitrary matrix that spans the same subspace as $\mathbf{U}_j^{(N_s)}$ and the transmitter can determine an appropriate orthogonal basis for outer-tier precoding by applying an SVD. This renders well known Grassmannian quantization principles as applicable for CSI feedback calculation. We therefore put the scope on Grassmannian CSI quantization in the following.

### B. Single Stage Quantization

As shown above, MET outer-tier precoding steers the transmit signal into the subspace spanned by the maximum eigenmodes $\mathrm{span}\left(\mathbf{U}_j^{(N_s)}\right)$. This subspace information can efficiently be conveyed by an orthogonal basis that spans the subspace. With single stage Grassmannian quantization, a quantization codebook of orthogonal bases of $N_s$-dimensional subspaces of the $N_t$-dimensional complex Euclidean space

$$\mathcal{Q}_{N_s}^{(N_t)} = \left\{\mathbf{W}_\ell \in \mathbb{C}^{N_t \times N_s} | \mathbf{W}_\ell^{\mathrm{H}}\mathbf{W}_\ell = \mathbf{I}_{N_s}, \forall \ell\right\} \tag{13}$$

is utilized to quantize $\mathbf{U}_j^{(N_s)}$; see, for example, [4] for efficient codebook designs.

A suitable quantization metric, to measure the subspace distance between $\mathrm{span}\left(\mathbf{U}_j^{(N_s)}\right)$ and the subspaces spanned by the elements of the quantization codebook, is the Grassmannian chordal distance. We therefore apply minimum chordal distance quantization to select the CSI feedback from the given codebook

$$\begin{aligned}\hat{\mathbf{U}}_{j,\text{single}}^{(N_s)} &= \underset{\mathbf{W}_\ell \in \mathcal{Q}_{N_s}^{(N_t)}}{\arg\min}\ d_c^2\left(\mathbf{U}_j^{(N_s)}, \mathbf{W}_\ell\right) \\ &= \underset{\mathbf{W}_\ell \in \mathcal{Q}_{N_s}^{(N_t)}}{\arg\min}\ N_s - \mathrm{tr}\left(\left(\mathbf{U}_j^{(N_s)}\right)^{\mathrm{H}}\mathbf{W}_\ell\mathbf{W}_\ell^{\mathrm{H}}\mathbf{U}_j^{(N_s)}\right).\end{aligned} \tag{14}$$

As we show in Sec. VI, the achievable rate of our two-tier precoding approach with imperfect outer-tier CSIT is determined by the average chordal distance quantization error; hence, minimizing the chordal distance quantization error is equivalent to maximizing the achievable transmission rate.

Upper and lower bounds for the distortion-rate function of this kind of quantization problem are provided in [40]. The upper bound of [40] is achieved by RSQ, i.e., quantization codebooks consisting of matrices $\mathbf{W}_\ell$ that span random isotropically distributed $N_s$-dimensional subspaces.

From the distortion-rate bounds of [40] it is known that the average chordal distance distortion scales as:

$$\begin{aligned}\bar{\mathrm{d}}_{c,\text{single}}^2 &= \mathbb{E}\left(d_c^2\left(\mathbf{U}_j^{(N_s)}, \hat{\mathbf{U}}_{j,\text{single}}^{(N_s)}\right)\right) \\ &\approx k_{N_t,N_s,N_s} K^{-1/(N_s(N_t-N_s))}\end{aligned} \tag{15}$$

where $K = \left|\mathcal{Q}_{N_s}^{(N_t)}\right| = 2^b$ is the size of the quantization codebook and $b$ is the corresponding number of required CSI feedback bits. The dimension-dependent constant $k_{N_t,N_s,N_s}$ is provided in [40].[1]

Since in massive MIMO $N_t$ is very large and $N_s \ll N_t$, the exponent $1/(N_s(N_t - N_s))$ in (15) is very small and, thus, huge quantization codebooks are required to achieve a small quantization error. Yet, performing the quantization procedure in (14) is computationally feasible only for relatively small codebook sizes. Hence, the single stage quantization approach is practically not applicable in massive MIMO systems.

### C. Dual Stage Product Codebook Quantization

In order to reduce the computational complexity of the outer-tier CSI quantization problem, we propose to utilize a product codebook construction in the following. Specifically, we utilize two codebooks

$$\mathcal{Q}_{1,S}^{(N_t)} = \mathcal{Q}_S^{(N_t)}, \quad \mathcal{Q}_{2,N_s}^{(S)} = \mathcal{Q}_{N_s}^{(S)}, \tag{16}$$

where $S \geq N_s$ denotes the intermediate subspace dimension of the first stage quantization codebook and both codebooks follow the structure of (13) with proper matrix dimensions.

*1) First stage quantizer:*

In the first quantization stage, we apply minimum chordal distance quantization to $\mathbf{U}_j^{(N_s)}$ utilizing the codebook $\mathcal{Q}_{1,S}^{(N_t)}$ as defined in (16)

$$\begin{aligned}\hat{\mathbf{U}}_j^{(S)} &= \underset{\mathbf{W}_\ell \in \mathcal{Q}_{1,S}^{(N_t)}}{\arg\min}\ d_c^2\left(\mathbf{U}_j^{(N_s)}, \mathbf{W}_\ell\right) = \\ &= \underset{\mathbf{W}_\ell \in \mathcal{Q}_{1,S}^{(N_t)}}{\arg\min}\ N_s - \mathrm{tr}\left(\left(\mathbf{U}_j^{(N_s)}\right)^{\mathrm{H}}\mathbf{W}_\ell\mathbf{W}_\ell^{\mathrm{H}}\mathbf{U}_j^{(N_s)}\right).\end{aligned} \tag{17}$$

Notice that $\hat{\mathbf{U}}_j^{(S)}$ is of size $N_t \times S$; it therefore represents the $S$-dimensional subspace of the quantization codebook $\mathcal{Q}_{1,S}^{(N_t)}$ that is closest to the $N_s$-dimensional subspace spanned by $\mathbf{U}_j^{(N_s)}$. In general, for a given quantization codebook size, quantizing into a higher-dimensional subspace gives a lower quantization error. However, for outer-tier precoding we require an $N_s$-dimensional subspace. We determine this $N_s$-dimensional subspace by applying our subspace quantization based combining (SQBC) method proposed in [8, 41].

Given the quantized $S$-dimensional subspace $\mathrm{span}\left(\hat{\mathbf{U}}_j^{(S)}\right)$, the goal of SQBC is to determine the $N_s$-dimensional subspace of $\mathrm{span}\left(\hat{\mathbf{U}}_j^{(S)}\right)$ that achieves the same chordal distance to $\mathbf{U}_j^{(N_s)}$ as $\hat{\mathbf{U}}_j^{(S)}$. The defining equation for the semi-unitary SQBC matrix $\mathbf{B}_j \in \mathbb{C}^{S \times N_s}$, $\mathbf{B}_j^{\mathrm{H}}\mathbf{B}_j = \mathbf{I}_{N_s}$ thus is

$$d_c^2\left(\hat{\mathbf{U}}_j^{(S)}, \mathbf{U}_j^{(N_s)}\right) \stackrel{!}{=} d_c^2\left(\hat{\mathbf{U}}_j^{(S)}\mathbf{B}_j, \mathbf{U}_j^{(N_s)}\right). \tag{18}$$

[1] We combine the multiple separately provided constants of [40] into the single constant $k_{n,p,q}$; here $n$ denotes the dimension of the complex Euclidean embedding space, $p$ is the dimension of the source subspace and $q$ is the dimension of the subspaces spanned by the codebook entries.

As shown in [8], this condition can be satisfied by the following construction

$$\mathbf{B}_j = \left(\hat{\mathbf{U}}_j^{(S)}\right)^{\mathrm{H}} \mathbf{U}_j^{(N_s)} \left( \left(\mathbf{U}_j^{(N_s)}\right)^{\mathrm{H}} \hat{\mathbf{U}}_j^{(S)} \left(\hat{\mathbf{U}}_j^{(S)}\right)^{\mathrm{H}} \mathbf{U}_j^{(N_s)} \right)^{-\frac{1}{2}}. \tag{19}$$

The matrix product $\hat{\mathbf{U}}_j^{(S)}\mathbf{B}_j$ spans the $N_s$-dimensional subspace that corresponds to the projection of $\mathbf{U}_j^{(N_s)}$ onto $\hat{\mathbf{U}}_j^{(S)}$. Notice, any other matrix $\mathbf{B}_j\mathbf{Q}$ with $\mathbf{Q}$ unitary achieves the same result, establishing a Grassmannian equivalence relationship that can be exploited for the quantization of $\mathbf{B}_j$.

*2) Second stage quantizer:*

In the second quantization stage, we apply minimum chordal distance quantization to $\mathbf{B}_j$ utilizing the codebook $\mathcal{Q}_{2,N_s}^{(S)}$

$$\hat{\mathbf{B}}_j = \underset{\mathbf{W}_\ell \in \mathcal{Q}_{2,N_s}^{(S)}}{\arg\min}\ d_c^2\left(\mathbf{B}_j, \mathbf{W}_\ell\right). \tag{20}$$

The quantized CSI is then obtained as the following product

$$\hat{\mathbf{U}}_{j,\text{dual}}^{(N_s)} = \hat{\mathbf{U}}_j^{(S)}\hat{\mathbf{B}}_j. \tag{21}$$

*3) Performance of the dual stage quantizer:*

The average quantization distortion of such a dual stage product codebook construction is governed by the following theorem:

**Theorem 1.** *The average chordal distance quantization distortion of the dual stage product codebook of Sec. IV-C, considering RSQ codebooks $\mathcal{Q}_{1,S}^{(N_t)}$, $\mathcal{Q}_{2,N_s}^{(S)}$ and/or isotropically distributed quantization source samples $\mathbf{U}_j^{(N_s)} \in \mathbb{C}^{N_t \times N_s}$, is given by*

$$\begin{aligned}\bar{\mathrm{d}}_{c,dual}^2 &= \mathbb{E}\left(d_c^2\left(\mathbf{U}_j^{(N_s)}, \hat{\mathbf{U}}_{j,dual}^{(N_s)}\right)\right) \\ &= N_s - \frac{1}{N_s}\left(N_s - \bar{\mathrm{d}}_{c,1}^2\right)\left(N_s - \bar{\mathrm{d}}_{c,2}^2\right), \end{aligned} \tag{22}$$

$$\bar{\mathrm{d}}_{c,1}^2 = \mathbb{E}\left(d_c^2\left(\mathbf{U}_j^{(N_s)}, \hat{\mathbf{U}}_j^{(S)}\right)\right), \tag{23}$$

$$\bar{\mathrm{d}}_{c,2}^2 = \mathbb{E}\left(d_c^2\left(\mathbf{B}_j, \hat{\mathbf{B}}_j\right)\right). \tag{24}$$

The proof of Th. 1 is provided in App. A.

When employing RSQ, the matrix $\mathbf{B}_j$ spans an isotropically distributed $N_s$-dimensional subspace in the $S$-dimensional complex Euclidean space. This implies that the two average distortions $\bar{\mathrm{d}}_{c,1}^2, \bar{\mathrm{d}}_{c,2}^2$ are both governed by the distortion-rate bounds provided in [40]. Specifically, we have

$$\bar{\mathrm{d}}_{c,1}^2 \approx k_{N_t,N_s,S}\, K_1^{-\frac{1}{N_s(N_t-S)}}, \tag{25}$$

$$\bar{\mathrm{d}}_{c,2}^2 \approx k_{S,N_s,N_s}\, K_2^{-\frac{1}{N_s(S-N_s)}}, \tag{26}$$

where $K_1 = \left|\mathcal{Q}_{1,S}^{(N_t)}\right| = 2^{b_1}, K_2 = \left|\mathcal{Q}_{2,N_s}^{(S)}\right| = 2^{b_2}$ and the dimension-dependent constants $k_{n,p,q}$ are defined in [40] (see footnote[1]).

Naturally, $\bar{\mathrm{d}}_{c,\text{dual}}^2$ is lower bounded by the average quantization distortion of the single stage codebook $\bar{\mathrm{d}}_{c,\text{single}}^2$ for a given total CSI feedback overhead $b = b_1 + b_2$. The advantage of the dual stage codebook is that the search over a single huge codebook is replaced by two searches over smaller codebooks since $2^{b_1} + 2^{b_2} \leq 2^{b_1+b_2} = 2^b$, thus providing a reduction of computational complexity.

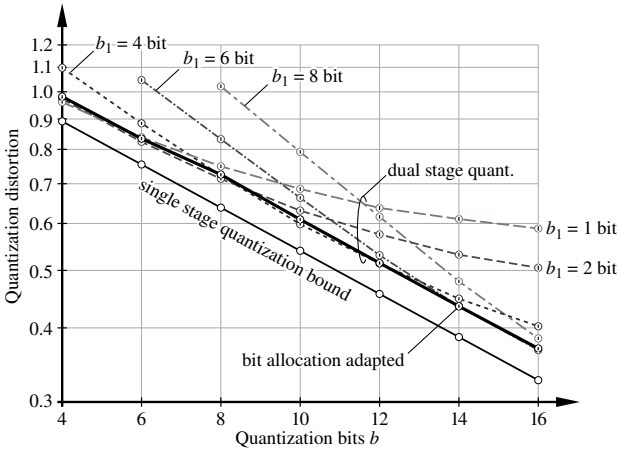



Fig. 2. Comparison of the single and dual stage quantization codebooks as a function of the bit budget and the bit-partitioning.

The achieved distortion of the dual stage quantization codebook depends on the bit allocation $b_1$, $b_2$ between the two stages, as well as, on the selection of the intermediate subspace dimensions $S$. Minimizing $\bar{\mathrm{d}}_{c,\text{dual}}^2$ over the bit allocation and the subspace dimension leads to the trivial result that all bits are allocated to either the first or the second stage and $S$ is set equal to $N_s$ or $N_t$, respectively; that is, the dual stage quantizer is effectively reduced to single stage quantization. Hence, in order to exploit the complexity reduction promised by dual stage quantization, we either have to fix the bit allocation, for example to $b_1 = b_2 = b/2$ to achieve minimum codebook sizes, and optimize $\bar{\mathrm{d}}_{c,\text{dual}}^2$ w.r.t. $S$; or we fix $S$ and optimize $\bar{\mathrm{d}}_{c,\text{dual}}^2$ w.r.t. the bit allocation. Our numerical experiments below demonstrate such optimizations.

### D. Comparison of Single and Dual Stage Quantization

*a) Scaling of the quantization distortion with fixed $S$ and growing $b$:* The behavior of the quantization distortion of the single stage codebook and the dual stage product codebook as a function of $b$ and the bit-partitioning between $b_1$ and $b_2 = b - b_1$ is shown in Fig. 2 for $N_t = 6$, $N_s = N_r = 2$, $S = 4$ and Rayleigh fading. Notice, we consider here a small-scale MIMO system, since for the single stage codebook it is hardly feasible to simulate codebook sizes larger than $2^{16}$ in terms of computational complexity. We observe in Fig. 2 that the optimal bit-partitioning between $b_1$ and $b_2$ depends on the bit budget $b$; the various dashed lines show the performance with fixed $b_1$. The black-solid line shows the performance with optimized bit-partitioning, where we calculated the optimal bit allocation by minimizing Eq. (22) employing the approximations (25), (26). With this optimized bit allocation, the performance of the dual stage codebook is reasonably close to the single stage codebook; the loss lies in the order of $1.5\,$bit.

*b) Impact of the bit allocation on the optimal intermediate subspace dimension $S$:* The dual stage quantization distortion $\bar{\mathrm{d}}_{c,\text{dual}}^2$ is lower bounded by the maximum of $\bar{\mathrm{d}}_{c,1}^2$ and $\bar{\mathrm{d}}_{c,2}^2$, with equality if $\bar{\mathrm{d}}_{c,2}^2 = 0$ or $\bar{\mathrm{d}}_{c,1}^2 = 0$, respectively. This implies that for given $b_1, b_2$, the choice of the subspace dimension $S$ governs the achievable quantization distortion. This behavior

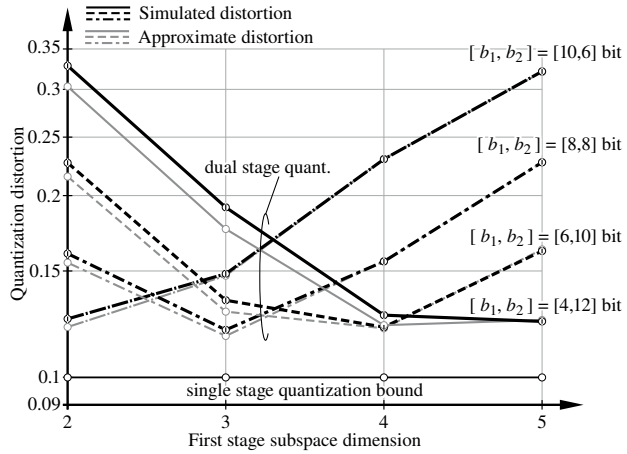



Fig. 3. Comparison of the single and dual stage quantization codebooks as a function of the first stage subspace dimension $S$ for various choices of $b_1$.

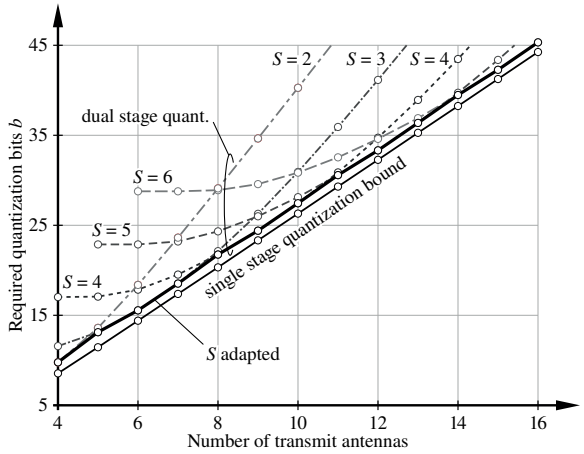



Fig. 4. Comparison of the required number of quantization bits for single and dual stage quantization as a function of the number of transmit antennas $N_t$ to achieve a fixed distortion of $\bar{\mathrm{d}}^2_{c,\text{single}} = \bar{\mathrm{d}}^2_{c,\text{dual}} = 0.125$.

is investigated in Fig. 3, where we consider $N_t = 6$, $N_s = N_r = 1$ and vary $S \in \{N_s + 1, \ldots, N_t - 1\}$ for the fixed bit-partitioning $[b_1, b_2] \in \{[4, 12], [6, 10], [8, 8], [10, 6]\}$ bits. In the figure, we show the performance of single stage quantization, the simulated performance of the dual stage product codebook and the approximate performance according to Eq. (22) employing the approximations (25), (26). We observe that the approximation is not perfect and it leads to a wrong selection $S = 4$ rather than the optimal $S = 5$ for $[b_1, b_2] = [4, 12]$; yet, the performance degradation is minimal.[2] We furthermore observe that the optimal subspace dimension $S$ gets smaller with growing $b_1$: this means that optimal selection effectively shifts the larger quantization burden to the quantization stage that provides better resolution. Finally, we observe that the minimal distortion over $S$ for different $[b_1, b_2]$ does not vary significantly. Thus, with an appropriate selection of $S$, we can reasonably employ an equal bit-partitioning which provides the lowest computational complexity. In our example, we achieve the lowest complexity with $S = 3$ and $[b_1, b_2] = [8, 8]$, which requires two searches over codebooks of size 256. For comparison, to achieve the same distortion of approximately 0.12, single stage quantization requires a single search over a codebook of size about $27\,000 \approx 2^{14.72}$.

*c) Scaling of the required number of quantization bits with growing $N_t$:* We next investigate how the required number of quantization bits scales as a function of the number of transmit antennas $N_t$, when the achieved quantization distortion is fixed. Specifically, we consider $N_t \in [4, 16]$, $N_r = N_s = 1$ and $\bar{\mathrm{d}}^2_{c,\text{single}} = \bar{\mathrm{d}}^2_{c,\text{dual}} = 0.125$ in our example, and we partition the total quantization bits of the dual stage quantizer equally amongst the two stages $b_1 = b_2 = b/2$. We can determine the number of required feedback bits from the analytic bounds (15), respectively, (22), (25) and (26). Furthermore, we utilize the bounds to determine the optimal intermediate quantization subspace dimension $S$ for a given $N_t$. The corresponding results are shown in Fig. 4. We observe that the dual stage quantizer with adapted $S$ achieves the same scaling behavior as the single stage quantizer, but it exhibits a small offset of approximately 1.5 bits. With fixed $S$, the scaling behavior is worse due to the equal bit partitioning $b_1 = b_2 = b/2$. Notice, the results shown in Fig. 4 are obtained from the quantization bounds; however, we verified by Monte-Carlo simulations that the intended distortion of $\bar{\mathrm{d}}^2_{c,\text{single}} = \bar{\mathrm{d}}^2_{c,\text{dual}} = 0.125$ is achieved up to $N_t \leq 7$ ($b \leq 17.4$ bits) for single stage quantization and up to $N_t \leq 11$ ($b_1 = b_2 = 30.6/2$ bits) for dual stage quantization; for larger values of $b, b_1, b_2$ the computational complexity (and thus the simulation time) of Monte-Carlo simulations becomes unreasonable.

[2] To obtain an even better approximation, one might have to consider the $o(1)$-terms provided in [40, Theorem 4] which we neglected here.

## V. Low Complexity Dual Stage Quantization

The dual stage quantization approach described above already provides substantial complexity gains compared to single stage quantization. Yet, for massive MIMO, the necessary codebook sizes of the first and second stage quantizers $\mathcal{Q}^{(N_t)}_{1,S}$, $\mathcal{Q}^{(S)}_{2,N_s}$ to achieve a sufficiently small quantization error are still very large and hardly feasible when employing an exhaustive search according to (17), (20) utilizing unstructured codebooks, such as RSQ. To further reduce the complexity, it is therefore necessary to employ quantization codebooks and metrics that can be computed efficiently.

In this section, we thus propose low-complexity quantizers for the first and second quantization stages of our dual stage construction. In Sec. V-A, we propose to utilize a DFT codebook for the first quantization stage. It is known that DFT codebooks can perform well for the structure imposed on $\mathbf{U}^{(N_s)}_j$ by multi-scattering directional channels, thus rendering them useful for the first quantization stage. However, this structure is not preserved by the SQBC matrix $\mathbf{B}_j$ and hence DFT codebooks are not effective for the second quantization stage. We therefore propose to utilize a scalar quantization for the second stage in Sec. V-B, which generally requires $S$ to be small to keep the quantization overhead reasonable. In Sec. V-C, we provide a numerical performance investigation of the low complexity implementation for multi-scattering directional channels.

### A. DFT-based first stage quantizer

To realize the first quantization stage with low computational complexity, we consider (oversampled) DFT-based quantization codebooks [42]

$$\mathcal{Q}_{S,\text{DFT}}^{(N_t,N_{\text{DFT}})} = \left\{ \frac{1}{\sqrt{N_t}} \left[\mathbf{D}_{N_{\text{DFT}}}\right]_{1:N_t,\mathcal{S}_q} \middle| \forall \mathcal{S}_q \in \mathcal{C}^{(S)} \right\}, \quad (27)$$

$$\mathcal{C}^{(S)} = \Bigg\{ \mathcal{S}_q \subseteq \{1,\ldots,N_{\text{DFT}}\}, |\mathcal{S}_q| = S, \mathcal{S}_q[i] \neq \mathcal{S}_q[j] \ \forall i \neq j, \ q \in \left\{1,\ldots,\binom{N_{\text{DFT}}}{S}\right\} \Bigg\}, \quad (28)$$

$$\left[\mathbf{D}_{N_{\text{DFT}}}\right]_{\ell,k} = e^{-j\frac{2\pi(\ell-1)(k-1)}{N_{\text{DFT}}}}.$$

Here, $\left[\mathbf{D}_{N_{\text{DFT}}}\right]_{1:N_t,\mathcal{S}_q}$ denotes the matrix consisting of the first $N_t$ rows of the DFT matrix $\mathbf{D}_{N_{\text{DFT}}}$ of size $N_{\text{DFT}}$ and the columns indexed by the set $\mathcal{S}_q$. Valid sets $\mathcal{S}_q$ are taken from the collection of sets $\mathcal{C}^{(S)}$; they contain $S$ unique indices $k \in \{1,\ldots,N_{\text{DFT}}\}$. Since our proposed quantizer only conveys subspace information, the order of the columns indexed by $\mathcal{S}_q$ is irrelevant; w.l.o.g., we can assume that the indices in $\mathcal{S}_q$ are sorted in increasing order. This implies that the number of feasible unordered sets $\mathcal{S}_q$ and, hence, the size of the collection $\mathcal{C}^{(S)}$ is $\binom{N_{\text{DFT}}}{S}$. Correspondingly, the size of the DFT quantization codebook $\mathcal{Q}_{S,\text{DFT}}^{(N_t,N_{\text{DFT}})}$ in bits, representing the CSI feedback overhead, is $b_1 = \log_2\left(\binom{N_{\text{DFT}}}{S}\right)$.

It is well-known that DFT-based quantization codebooks perform well for multi-scattering directional channels, as described in Sec. II-B, and small oversampling $N_{\text{DFT}} \gtrapprox N_t$. Yet, these DFT-based codebooks fail to achieve the distortion scaling performance of RSQ (25) with growing codebook size; in fact, the unit-modulus constraint of the elements of $\mathbf{D}_{N_{\text{DFT}}}$ leads to a saturation of the achieved chordal distance quantization error with growing codebook size and fixed dimensions $N_t, S$ [43]. This is the main limiting factor when utilizing DFT-based codebooks in single stage quantization structures as described in Sec. IV-B. However, this is not a significant problem in the proposed dual stage product codebook quantization structure, since it allows to reduce the quantization error floor of the DFT codebook by increasing the subspace dimension $S$ of the first quantization stage, as we demonstrate further below.

*1) FFT implementation of the first stage for $N_{DFT} = N_t$:*

If $N_{\text{DFT}} = N_t$, i.e., no oversampling is employed, the chordal distance quantization according to (18) for DFT codebooks can efficiently be realized by means of an inverse FFT

$$\tilde{\mathbf{U}}_j^{(N_s)} = \text{IFFT}_{N_{\text{DFT}}}\left(\mathbf{U}_j^{(N_s)}\right) \in \mathbb{C}^{N_{\text{DFT}}\times N_s}, \quad (29)$$

$$\mathcal{S}_j^{(S)} = \underset{\mathcal{S}_q\in\mathcal{C}^{(S)}}{\arg\max} \left\| \left[\tilde{\mathbf{U}}_j^{(N_s)}\right]_{\mathcal{S}_q,:} \right\|^2, \quad (30)$$

$$\hat{\mathbf{U}}_j^{(S)} = \left[\mathbf{D}_{N_{\text{DFT}}}\right]_{1:N_t,\mathcal{S}_j^{(S)}}, \quad (31)$$

where the index set $\mathcal{S}_q$ satisfies the conditions specified in (28). Notice, implementing (30) does not require an exhaustive search over all feasible sets $\mathcal{S}_q$; one can simply select the $S$ rows of $\tilde{\mathbf{U}}_j^{(N_s)}$ that have the largest norm. The corresponding CSI feedback of the first stage quantizer is the index of the set $\mathcal{S}_j^{(S)}$ within the collection $\mathcal{C}^{(S)}$.

*2) FFT implementation of the first stage for $N_{DFT} > N_t$:*

For $N_{\text{DFT}} > N_t$, one could apply the same approach to the zero-padded matrix

$$\mathbf{U}_{j,\text{pad}}^{(N_s)} = \begin{bmatrix} \mathbf{U}_j^{(N_s)} \\ \mathbf{0}_{N_{\text{DFT}}-N_t,N_s} \end{bmatrix} \quad (32)$$

to realize an oversampled FFT. However, in this case maximizing the norm as in (30) is not equivalent to minimizing the chordal distance, as the columns of $\left[\mathbf{D}_{N_{\text{DFT}}}\right]_{1:N_t,\mathcal{S}_j^{(S)}}$ are in general not orthogonal for $N_{\text{DFT}} > N_t$. A chordal distance minimization utilizing an oversampled DFT codebook would therefore require an exhaustive search over all feasible index sets $\mathcal{S}_q$ in the collection $\mathcal{C}^{(S)}$. This search over $\binom{N_{\text{DFT}}}{S}$ possibilities is computationally not feasible for large $N_{\text{DFT}}$.

To mitigate this issue, we propose a low-complexity greedy quantization method employing a variant of the orthogonal matching pursuit (OMP) algorithm; the pseudo-code of the algorithm is provided in Alg. 1.

The algorithm iteratively selects columns of the DFT-matrix to quantize $\mathbf{U}_j^{(N_s)}$; the index-set of selected columns up to iteration $s$ is denoted by $\mathcal{S}_j^{(s)}$. The variables $\hat{\mathbf{U}}_j^{(s)}$ in (38) and $\mathbf{B}_j^{(s)}$ in (39) denote intermediate quantization results of iteration $s$. Given the intermediate results from the previous iteration $s-1$, the algorithm first calculates a null-space projection of $\mathbf{U}_j^{(N_s)}$ onto the null-space of $\hat{\mathbf{U}}_j^{(s-1)}\mathbf{B}_j^{(s-1)}$ in (34), in order to determine the part of $\mathbf{U}_j^{(N_s)}$ that is not yet well-represented by $\hat{\mathbf{U}}_j^{(s-1)}\mathbf{B}_j^{(s-1)}$. This null-space projection $\mathbf{U}_{\text{null}}$ is then utilized to determine the next column index $k^*$ of the DFT matrix to be added to the quantization index set $\mathcal{S}_j^{(s)}$ in (37). This is achieved by maximizing the inner-product with the columns of the DFT matrix that have not yet been selected in prior iterations, i.e. are not part of $\mathcal{S}_j^{(s-1)}$. This can be efficiently realized by the inverse FFT. Since the order of the selected columns is irrelevant for a subspace representation, we sort them in increasing order. As we consider an oversampled DFT with $N_{\text{DFT}} > N_t$, the matrix $\left[\mathbf{D}_{N_{\text{DFT}}}\right]_{1:N_t,\mathcal{S}_j^{(s)}}$ does in general not contain orthogonal columns and we therefore orthogonalize it by applying an SVD in (38). The algorithm terminates with an index-set $\mathcal{S}_j^{(S)}$ of the DFT-matrix that attempts to minimize the subspace distance to $\mathbf{U}_j^{(N_s)}$. However, since this is a greedy algorithm, it cannot be guaranteed to provide the globally optimal index-set.

Regarding computational complexity, the main complexity comes form the SVDs required in (34) and (38), as well as, for the calculation of the inverse matrix square-root in (39); hence, the complexity is mainly dictated by three SVDs per iteration, i.e. a total of $3\,S$ SVDs. The quantization itself is efficiently achieved by the inverse FFT operation in (35). In contrast, the exhaustive search over the collection $\mathcal{C}^{(S)}$ requires in each iteration an SVD of $\left[\mathbf{D}_{N_{\text{DFT}}}\right]_{1:N_t,\mathcal{S}_j^{(S)}}$ for the calculation of the subspace chordal distance. This is necessary since the calculation of the chordal distance requires an orthogonal basis for the subspace spanned by the (generally non-orthogonal)

columns index by $\mathcal{S}_j^{(S)}$. It therefore requires a total of $\binom{N_{\text{DFT}}}{S}$ SVDs, which is substantially larger than $3\,S$ for large $N_{\text{DFT}}$.

**Algorithm 1** First stage quantization based on OMP.

Initialize the quantization index set $\mathcal{S}_j^{(0)} = \{\,\}$
**for** $s = 1$ to $S$ **do**
  **if** $s \neq 1$ **then**
    Apply a null-space projection to calculate the FFT input

$$\hat{\mathbf{U}} = \hat{\mathbf{U}}_j^{(s-1)} \mathbf{B}_j^{(s-1)}, \tag{33}$$

$$\mathbf{U}_{\text{null}} \mathbf{\Sigma} \mathbf{V}^{\text{H}} = \left( \mathbf{I}_{N_t} - \hat{\mathbf{U}} \hat{\mathbf{U}}^{\text{H}} \right) \mathbf{U}_j^{(N_s)} \tag{34}$$

  **else**
    Set the FFT input $\mathbf{U}_{\text{null}} = \left[ \mathbf{U}_j^{(N_s)} \right]_{:,1}$
  **end if**
  Initialize the quantization metric vector $\mathbf{m} = \mathbf{0}_{N_{\text{DFT}}}$
  Calculate the oversampled inverse FFT of $\mathbf{U}_{\text{null}}$

$$\tilde{\mathbf{U}}_{\text{null}} = \text{IFFT}_{N_{\text{DFT}}} \left( \begin{bmatrix} \mathbf{U}_{\text{null}} \\ \mathbf{0}_{N_{\text{DFT}}-N_t, N_s} \end{bmatrix} \right) \tag{35}$$

  Calculate the quantization metric
  **for** $k \in \{1, \ldots, N_{\text{DFT}}\}, k \notin \mathcal{S}_j^{(s-1)}$ **do**

$$\mathbf{m}[k] = \left\| \left[ \tilde{\mathbf{U}}_{\text{null}} \right]_{k,:} \right\| \tag{36}$$

  **end for**
  Find the best quantization index $k^* = \arg\max_k \mathbf{m}[k]$
  Update the quantization index set

$$\mathcal{S}_j^{(s)} = \text{sort} \left( \left\{ \mathcal{S}_j^{(s-1)}, k^* \right\}, \text{increasing} \right) \tag{37}$$

  Update the quantized subspace by applying an SVD

$$\hat{\mathbf{U}}_j^{(s)} \mathbf{\Sigma} \mathbf{V}^{\text{H}} = \left[ \mathbf{D}_{N_{\text{DFT}}} \right]_{1:N_t, \mathcal{S}_j^{(s)}} \tag{38}$$

  Calculate the temporary SQBC matrix

$$\mathbf{B}_j^{(s)} = \left( \hat{\mathbf{U}}_j^{(s)} \right)^{\text{H}} \mathbf{U}_t \left( \mathbf{U}_t^{\text{H}} \hat{\mathbf{U}}_j^{(s)} \left( \hat{\mathbf{U}}_j^{(s)} \right)^{\text{H}} \mathbf{U}_t \right)^{-\frac{1}{2}}, \tag{39}$$

$$\mathbf{U}_t = \left[ \mathbf{U}_j^{(N_s)} \right]_{:,1:\min(s,N_s)}$$

**end for**
Output $\hat{\mathbf{U}}_j^{(S)}$, $\mathbf{B}_j^{(S)}$ and $\mathcal{S}_j^{(S)}$.

### *B. Scalar Second Stage Quantizer*

To realize the second stage quantizer in low-complexity, we propose to employ a scalar quantization of the individual elements of $\mathbf{B}_j$

$$\hat{b}_{n,m} = \underset{b_i \in \mathcal{B}^{(S)}}{\arg\min} \left| [\mathbf{B}_j]_{n,m} - b_i \right|^2, \tag{40}$$

$$\forall n \in \{1, \ldots, S\},\ m \in \{1, \ldots, N_s\},$$

where $\mathcal{B}^{(S)}$, $\left|\mathcal{B}^{(S)}\right| = 2^{b_s}$ with $b_s = b_2/(N_s S)$, denotes the scalar complex-valued quantization codebook. The scalars $\hat{b}_{n,m}$ are individually fed back to the transmitter. The reconstructed CSIT is obtained by an SVD

$$\hat{\mathbf{B}}_j \mathbf{\Sigma} \mathbf{V}^{\text{H}} = \begin{bmatrix} \cdots & & \\ \vdots & \hat{b}_{n,m} & \vdots \\ & \cdots & \end{bmatrix}, \tag{41}$$

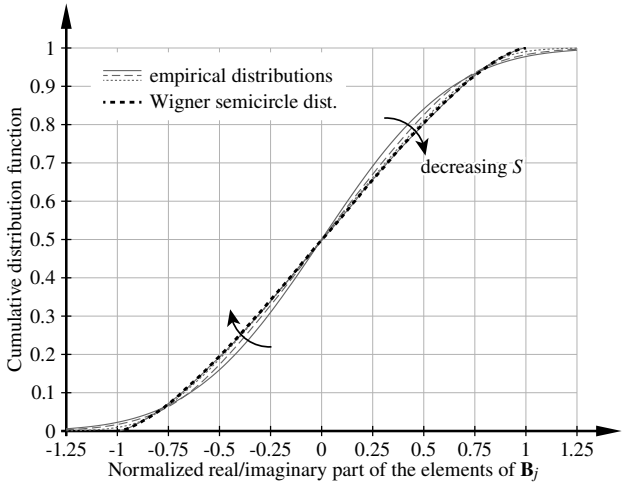



Fig. 5. Distribution of the normalized real/imaginary parts of the coefficients of $\mathbf{B}_j$ for isotropically distributed full rank channel matrices ($N_t = 64$, $N_s = N_r = 2$) with varying subspace dimension $S \in \{2, 8, 32, 64\}$.

where the right-hand side is the matrix reconstructed from the scalars $\hat{b}_{n,m}\ \forall n, m$ and the left-hand side is a compact-size SVD of this matrix (the matrices $\mathbf{\Sigma}$ and $\mathbf{V}$ are discarded).

The quantization codebook $\mathcal{B}^{(S)}$ can, e.g., be optimized by means of the Lloyd-Max algorithm [44]. As mentioned above, for isotropically distributed channel matrices, matrix $\mathbf{B}_j$ is also isotropically distributed and satisfies $\text{tr}\left(\mathbf{B}_j^{\text{H}} \mathbf{B}_j\right) = N_s$. Assuming that the elements of $\mathbf{B}_j$ are thus identically distributed, we conclude

$$\mathbb{E}\left(\text{tr}\left(\mathbf{B}_j^{\text{H}} \mathbf{B}_j\right)\right) = \sum_{n=1}^{S} \sum_{m=1}^{N_s} \mathbb{E}\left( \left| [\mathbf{B}_j]_{n,m} \right|^2 \right)$$

$$= \mathbb{E}\left( \left| [\mathbf{B}_j]_{n,m} \right|^2 \right) N_s S \ \Rightarrow\ \mathbb{E}\left( \left| [\mathbf{B}_j]_{n,m} \right|^2 \right) = \frac{1}{S}. \tag{42}$$

This condition can be satisfied, e.g., if we assume that the elements $[\mathbf{B}_j]_{n,m}$ are uniformly distributed in the complex Euclidean space within a disc of radius $R = 2/S$. It then follows that the real and imaginary part of $[\mathbf{B}_j]_{n,m}$ are distributed according to a Wigner semicircle distribution of radius $R$. Notice, for $S = N_s = 2$, $\mathbf{B}_j$ is a random unitary matrix and its elements follow exactly the Wigner semicircle distribution.

In Fig. 5, we investigate the empirical cumulative distribution function of the real/imaginary parts of $\mathbf{B}_j$ for full rank isotropically distributed channel matrices (e.g., Rayleigh fading or a multi-scattering environment where the number of multipath scatterers $N_p$ is at least $N_r$) and varying subspace dimension $S$. The simulation result is shown for $N_t = 64$, $N_s = N_r = 2$; yet, we observed that the empirical distribution of the real and imaginary part of $[\mathbf{B}_j]_{n,m}$ is independent of $N_t, N_r$ and $N_s$, as long as the channel matrices are full rank and isotropically distributed, implying an isotropic distribution of $\mathbf{B}_j$. In Fig. 5, the real/imaginary parts are normalized w.r.t. $R$. As expected, for $S = N_s = 2$, the real and imaginary part of the elements of $\mathbf{B}_j$ follow exactly the Wigner semicircle distribution. With growing $S > N_s$, the empirical distributions deviate increasingly from the Wigner semicircle distribution. Yet, the agreement is still good enough, such that we employ in all our remaining simulations a quantization codebook

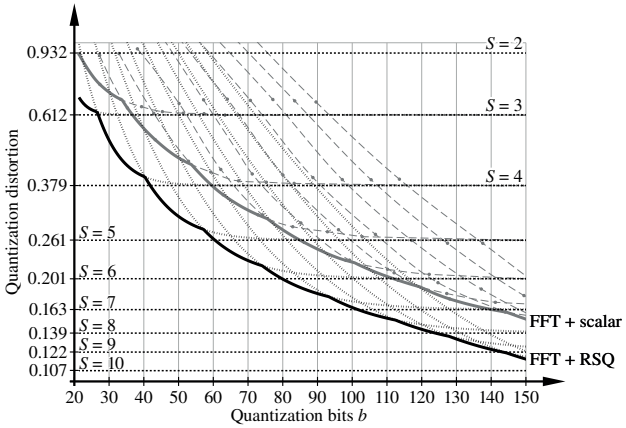



Fig. 6. Detailed performance investigation of dual stage quantization considering $N_t = 64, N_r = 2, N_s = 2$, a multi-scattering channel with $N_p = 4$, a DFT codebook size of $N_{\text{DFT}} = N_t = 64$ and varying subspace dimension $S \in \{2, \ldots, 10\}$.

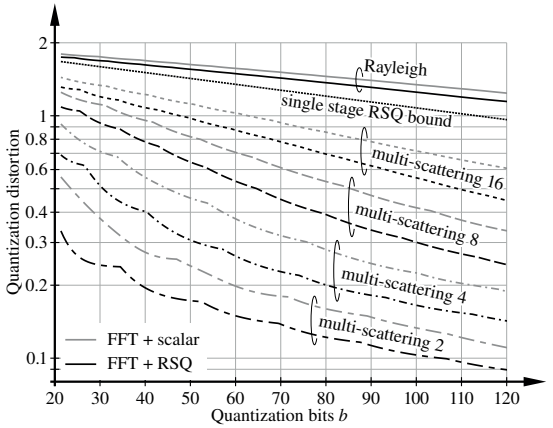



Fig. 7. Comparison of dual stage quantization utilizing a DFT codebook of size $N_{\text{DFT}} = 64$ for multi-scattering channels with a varying number of multipath components $N_p \in \{2, 4, 8, 16\}$ and for Rayleigh fading.

$\mathcal{B}^{(S)}$ that is optimized for complex-valued elements $[\mathbf{B}_j]_{n,m}$ uniformly distributed within a disc of radius $R = 2/S$.

If the channel matrix is not full rank, $\mathbf{B}_j$ contains elements of small magnitude and the empirical distribution deviates more strongly from the Wigner semicircle distribution.

### C. Investigation of Low Complexity Dual Stage Quantization

*a) Impact of the subspace dimension $S$ on the quantization performance:* In Fig. 6, we provide a detailed performance investigation of the proposed dual stage product codebook as a function of the subspace dimension $S$ and the total number of quantization bits $b$. We consider a system with $N_t = 64, N_r = 2, N_s = 2$, a multi-scattering channel with $N_p = 4$ equally strong multipath components and a first stage quantizer employing a DFT codebook of size $N_{\text{DFT}} = N_t = 64$. The feedback overhead of the first stage quantizer ranges from $b_1 = \log_2\left(\binom{N_{\text{DFT}}}{S}\right) \approx 11$ bits for $S = 2$ to $b_1 \approx 37$ bits for $S = 10$; the corresponding number of bits of the second stage quantizer follows from $b_2 = b - b_1$. The horizontal lines Fig. 6 show the error floor achieved with perfect second stage quantization, which ranges from $\bar{\text{d}}^2_{c,\text{dual}} = 0.932$ for $S = 2$ to $\bar{\text{d}}^2_{c,\text{dual}} = 0.107$ for $S = 10$. This error-floor is caused by the unit-modulus constraint of the elements of the DFT codebook. Hence, to achieve a diminishing quantization distortion with growing number of bits $b$, the subspace dimension $S$ has to increase with $b$.

With imperfect second stage quantization, we can observe how the optimal subspace dimension $S$ increases with growing number of quantization bits $b$. The thin dashed and dotted lines show the performance for fixed $S$, whereas the solid thick lines represent the minimum quantization error achieved by adapting $S$. In the figure, we compare the performance of scalar second stage quantization to the performance utilizing RSQ as second stage quantizer.[3] The scalar quantization exhibits a significant loss w.r.t. to RSQ. However, RSQ is in practice computationally not feasible for $b_2 \geq 16$.

Even though not shown in the figure, we also investigated the performance of single stage quantization utilizing an oversampled DFT codebook of size $N_{\text{DFT}} \in 2^{\{6,\ldots,16\}}$, which achieves $\bar{\text{d}}^2_{c,\text{single}} = 0.932$ for $N_{\text{DFT}} = 2^6$ and exhibits an error-floor of $\bar{\text{d}}^2_{c,\text{single}} \approx 0.7$ for $N_{\text{DFT}} \geq 2^9$. Hence, when employing the low-complexity DFT codebook, the dual stage quantization can strictly outperform single stage quantization, because it can omit the error-floor of the DFT codebook by increasing the intermediate subspace dimension $S$, which offloads the quantization burden to the second quantization stage.

*b) Performance of DFT based quantization for multi-scattering channels with varying number of paths:* We next investigate the performance for different multi-scattering channels with varying number of equally strong multipath components $N_p \in \{2, 4, 8, 16\}$, as well as, for Rayleigh fading in Fig. 7. Notice, the results for $N_p = 4$ denoted as "multi-scattering 4" in Fig. 7 coincide with those of Fig. 6. For Rayleigh fading, the dual stage codebook performs worse than single stage quantization, whose performance follows closely the "single stage RSQ bound" shown in the figure. Yet, for the multi-scattering channels, the dual stage quantization utilizing the DFT codebook outperforms single stage quantization with the RSQ codebook substantially.[4] As mentioned already above in the discussion of Fig. 6, single stage quantization can also gain from the more efficient subspace representation of the DFT codebook for multi-scattering channels; yet, it exhibits an error floor that lies significantly above the performance achievable by the dual stage quantizer.

## VI. Asymptotic Rate Approximation of Two-Tier Precoding with Imperfect CSIT

In this section, we restrict ourselves to single-antenna users $N_r = 1$ and, correspondingly, single stream transmission per user $N_s = 1$ employing ZF beamforming. For this situation, we approximately calculate the achievable transmission rate of massive MIMO systems assuming perfect inner-tier CSIT and imperfect outer-tier CSIT. In Sec. VI-A, we specialize our system model to single-stream transmission per user and particularize the employed power constraint. In Sec. VI-B, we provide our analytic approximation of the achievable rate of

[3]The RSQ results are obtained from the approximation (26); yet, we verified their accuracy up to $b_2 \leq 12$ bits by means of simulations.

[4]Notice, the single stage RSQ bound is valid for all considered channel models; it only requires the channels to be isotropically distributed.

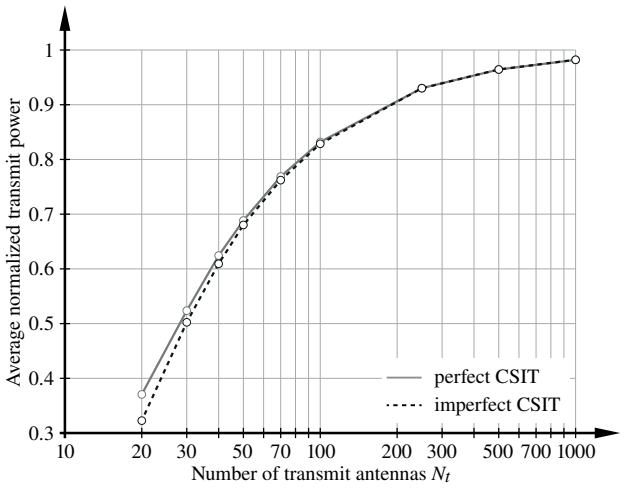



Fig. 8. Average per-user transmit power of inner-tier ZF beamforming with the power constraint (47) as a function of $N_t$ for $P_u = 1$ and $U = 10$.

the considered system and we investigate the SNR loss caused by imperfect outer-tier CSIT. We evaluate our analysis by means of Monte-Carlo simulations in Sec. VI-C.

### A. Outer-Tier MET and Inner-Tier ZF Beamforming

With ZF beamforming and single-antenna users, the input-output relationship (1) reduces to

$$y_u = \sigma_u \gamma_u \mathbf{u}_u^{\mathrm{H}} \mathbf{F}_o \left( \mathbf{f}_{i,u} x_u + \sum_{j=1, j\neq u}^{U} \mathbf{f}_{i,j} x_j \right) + z_u, \quad (43)$$

where we omit the superscript $^{(1)}$ as used in (2), since the channel reduces to just a single singular value $\sigma_u$ and its corresponding left singular vector $\mathbf{u}_u$. The ZF precoder of user $u$ can be obtained from

$$\mathbf{f}_{i,u} = \mathbf{H}_o^{\mathrm{H}} \left( \mathbf{H}_o \mathbf{H}_o^{\mathrm{H}} \right)^{-1} \mathbf{e}_u, \quad (44)$$

$$\mathbf{H}_o = \left[ \mathbf{u}_1, \ldots, \mathbf{u}_U \right]^{\mathrm{H}} \mathbf{F}_o \in \mathbb{C}^{U \times U}, \quad (45)$$

where $\mathbf{e}_u \in \mathbb{C}^{U \times 1}$ denotes the $u$-th canonical basis vector.

There are several meaningful ways for the selection of the power normalization factor $\gamma_u$. If we consider an instantaneous power constraint, we would select $\gamma_u$ according to

$$\gamma_u^2 = \frac{P_u}{\left\| \mathbf{F}_o \mathbf{f}_{i,u} \right\|^2} = \frac{P_u}{\mathbf{e}_u^{\mathrm{H}} (\mathbf{H}_o \mathbf{H}_o^{\mathrm{H}})^{-1} \mathbf{H}_o \mathbf{F}_o^{\mathrm{H}} \mathbf{F}_o \mathbf{H}_o^{\mathrm{H}} (\mathbf{H}_o \mathbf{H}_o^{\mathrm{H}})^{-1} \mathbf{e}_u}. \quad (46)$$

For this expression, however, we were not able to calculate a good closed-form approximation of the average SNR $\beta_u = \frac{\mathbb{E}\left( (\sigma_u \gamma_u)^2 \right)}{\sigma_z^2}$. If we consider a hybrid precoding architecture, where $\mathbf{f}_{i,u}$ is implemented in base band and $\mathbf{F}_o$ is realized by passive/active radio frequency (RF) components after the power amplifier (PA), we would rather consider a power constraint on the PA-output leading to

$$\gamma_u^2 = \frac{P_u}{\left\| \mathbf{f}_{i,u} \right\|^2} = \frac{P_u}{\mathbf{e}_u^{\mathrm{H}} \left( \mathbf{H}_o \mathbf{H}_o^{\mathrm{H}} \right)^{-1} \mathbf{e}_u}. \quad (47)$$

We consider this power constraint in our analysis below.

Notice, in the massive MIMO limit of $N_t \to \infty$, Eq. (46) reduces to (47), since $\mathbf{F}_o^{\mathrm{H}} \mathbf{F}_o \to \mathbf{I}_U$ due to asymptotic orthogonality. However, this limiting case becomes relevant only slowly, as we show in Fig. 8, where we plot the per-user transmit power for $P_u = 1$ and $U = 10$, when using the power constraint (47), as a function of the number of transmit antennas $N_t$. We observe that the transmit power for large $N_t$ converges to one, corresponding to the power constraint (46); yet, for small $N_t$ the transmit power is far below one. The reason for this behavior is that with small number of antennas the inner product between the channel vectors of different users is generally not small and, therefore, the orthogonalization by the ZF beamformer effectively causes a power reduction of the intended signal. Nevertheless, Fig. 8 also shows that the average transmit power with perfect and imperfect outer-tier CSIT is very similar, even when considering a large chordal distance error of $\bar{\mathrm{d}}_c^2 = 0.5$, not only in the massive MIMO limit but also for relatively small numbers of antennas. Hence, the statement obtained below for the SNR loss caused by imperfect outer-tier CSIT is approximately valid for both power constraints, since the SNR loss only depends on the relative receive power with perfect and imperfect CSIT for equal transmit power, irrespective of whether the transmit power is one or below one.

### B. Asymptotic Achievable Rate and SNR Loss

The achievable transmission rate of the considered scenario is in the massive MIMO limit governed by the following theorem for perfect and imperfect outer-tier CSIT:

**Theorem 2.** *With perfect inner-tier CSIT and imperfect outer-tier CSIT with average chordal distance error $\bar{\mathrm{d}}_c^2$, the per-user achievable transmission rate of (imperfect) outer MET precoding and inner ZF beamforming, with the power constraint (47), $N_t \to \infty$ and isotropically distributed channel vectors, is well approximated by*

$$\mathbb{E}\left( \log_2 \left( 1 + \frac{P_u \sigma_u^2}{\sigma_z^2 \left[ \left( \mathbf{H}_o \mathbf{H}_o^{\mathrm{H}} \right)^{-1} \right]_{u,u}} \right) \right) \to \log_2 \left( 1 + \frac{P_u N_t}{\sigma_z^2 H_{\bar{\mathrm{d}}_c^2}^{-1}} \right), \quad (48)$$

$$H_{\bar{\mathrm{d}}_c^2}^{-1} = \frac{1}{4 m_{\bar{\mathrm{d}}_c^2}} \left( \frac{\lambda_{\bar{\mathrm{d}}_c^2}^{-}}{\lambda_{\bar{\mathrm{d}}_c^2}^{+}} + \frac{\lambda_{\bar{\mathrm{d}}_c^2}^{+}}{\lambda_{\bar{\mathrm{d}}_c^2}^{-}} + 2 \right), \quad (49)$$

$$\lambda_{\bar{\mathrm{d}}_c^2}^{+} = m_{\bar{\mathrm{d}}_c^2} + s_{\bar{\mathrm{d}}_c^2}, \quad \lambda_{\bar{\mathrm{d}}_c^2}^{-} = m_{\bar{\mathrm{d}}_c^2} - s_{\bar{\mathrm{d}}_c^2}, \quad (50)$$

$$m_{\bar{\mathrm{d}}_c^2} = (1 - \bar{\mathrm{d}}_c^2) + \frac{U-1}{N_t}, \quad (51)$$

$$s_{\bar{\mathrm{d}}_c^2}^2 = \frac{U-1}{N_t} \Bigg( 4 (1 - \bar{\mathrm{d}}_c^2)^2 + 2 (1 - \bar{\mathrm{d}}_c^2) \bar{\mathrm{d}}_c^2 + 4 (1 - \bar{\mathrm{d}}_c^2) \frac{U-2}{N_t} + \frac{U-2}{N_t} \Bigg), \quad (52)$$

$$\mathbf{H}_o = \left[ \mathbf{u}_1, \ldots, \mathbf{u}_U \right]^{\mathrm{H}} \mathbf{F}_o, \ \mathbf{F}_o = \left[ \hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_U \right]. \quad (53)$$

Here, $\hat{\mathbf{u}}_i$ denotes the imperfect estimate of $\mathbf{u}_i$ available at the transmitter, such that $\mathbb{E}\left( \left| \mathbf{u}_i^{\mathrm{H}} \hat{\mathbf{u}}_i \right|^2 \right) = 1 - \bar{\mathrm{d}}_c^2$. The factor $m_{\bar{\mathrm{d}}_c^2}$ denotes the expected value of the mean of the eigenvalues of the matrix $\mathbf{H}_o \mathbf{H}_o^{\mathrm{H}}$ and similarly $s_{\bar{\mathrm{d}}_c^2}$ denotes the expected value of their standard-deviation. Correspondingly, $\lambda_{\bar{\mathrm{d}}_c^2}^{+}$, $\lambda_{\bar{\mathrm{d}}_c^2}^{-}$ denote "typical" eigenvalues that lie one standard deviation

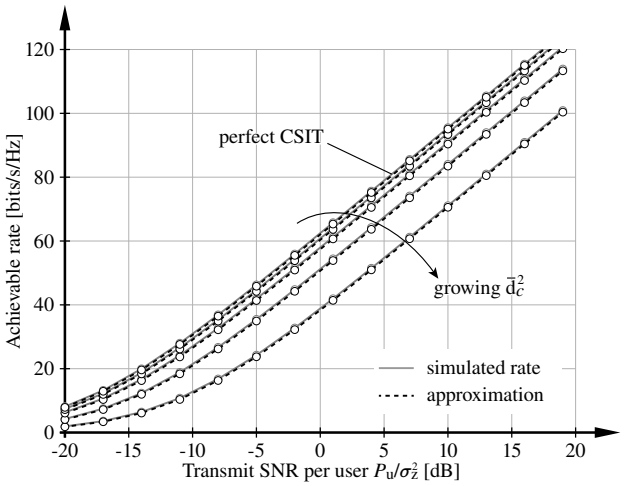



Fig. 9. Comparison of the rate approximation according to Th. 2 and the simulated achievable transmission rate of outer-tier MET precoding with imperfect CSIT and inner-tier ZF beamforming with perfect CSIT.

above and below the mean. The proof of Th. 2 is provided in App. B. It utilizes the bounds for the trace of the inverse of symmetric positive definite matrices developed in [45].

From Th. 2 we immediately deduce the following corollary:

**Corollary 2.1.** *The SNR loss in [dB] of outer-tier MET precoding with imperfect CSIT combined with inner-tier ZF beamforming with perfect CSIT, with respect to outer-tier MET precoding and inner-tier ZF beamforming with perfect outer- and inner-tier CSIT, is in the massive MIMO limit $N_t \to \infty$ for isotropically distributed channel vectors and either power constraint (46) or (47) well approximated by*

$$\Delta SNR_{\bar{d}_c^2} = 10 \log_{10}\left(\frac{H_{\bar{d}_c^2}^{-1}}{H_0^{-1}}\right), \tag{54}$$

*where $H_0^{-1}$ is obtained from (49) for $\bar{d}_c^2 = 0$.*

The receive SNR with perfect and imperfect CSIT is determined by two factors: 1) the inner-product between the normalized channel vector $\mathbf{u}_i$ and the ZF beamformer of user $i$; 2) the transmit power with perfect and imperfect CSIT. The first factor is independent of the power normalization and is thus the same for both power constraints (46), (47). The transmit power, however, is different. For the power constraint (46) it is equal to $P_u$ with perfect and imperfect CSIT by construction. For the power constraint (47), the transmit power depends on $N_t$ as shown in Fig. 8, but tends to the same value for perfect and imperfect CSIT. Hence, the SNR loss of both power constraints tends to the same value, as it only depends on the receive power ratio with perfect and imperfect CSIT for equal transmit power.

### C. Evaluation of the Approximations

In Fig. 9, we evaluate the asymptotic rate approximation of Th. 2 for $N_t = 100$, $U = 10$, $P_u = 1$ and $\bar{d}_c^2 \in \{0, 0.1, 0.25, 0.5, 0.75\}$ assuming Rayleigh fading channels. We observe that the rate approximation (48) provides a tight fit to the simulated rate. As expected, with growing CSIT error, the SNR loss w.r.t. perfect CSIT increases. In Tab. I, we provide a comparison between the simulated SNR loss (measured at high SNR) and the approximation provided in Cor. 2.1. Again, we obtain a tight fit between the simulated values and the approximation; hence, Eq. (54) can be utilized to determine the necessary CSIT accuracy to achieve a certain SNR loss, which in turn can be employed to determine the required codebook sizes using, e.g., Eqs (22), (25), and (26) in case of RSQ and dual-stage CSI quantization.

TABLE I
COMPARISON OF SNR LOSS VALUES OBTAINED FROM SIMULATIONS AND FROM (54) FOR THE SCENARIO SHOWN IN FIG. 9.

| **CSIT error $\bar{d}_c^2$** | **Simulated SNR loss** | **SNR loss (54)** |
|---|---|---|
| 0.1 | 0.51 dB | 0.51 dB |
| 0.25 | 1.40 dB | 1.41 dB |
| 0.5 | 3.45 dB | 3.5 dB |
| 0.75 | 7.38 dB | 7.40 dB |

## VII. APPLICATION EXAMPLE

In this section, we apply the proposed two-tier CSI feedback and precoding methods on real-world measured channel traces. The channel traces have been measured on the Nokia Bell Labs campus in Stuttgart, Germany, as reported in [46]. In Sections VII-A and VII-B, we briefly describe the measurement and simulation setup, resp., and in Sec. VII-C we provide the simulation results based on measured channel traces.

### A. Measurement Setup

The base station was equipped with a uniform planar antenna array (UPA) of size $N_t = N_v \times N_h = 4 \times 16 = 64$ patch antenna elements, where $N_v = 4, N_h = 16$ are the number of rows and columns of the UPA, and was mounted on a roof top at a height of 20 m with a mechanical down-tilt of $10^\circ$. The vertical distance of the patch antenna elements was $d_v = \lambda$ and the horizontal distance was $d_h = \lambda/2$, where $\lambda$ is the wavelength at the carrier frequency of $f_c = 2.18$ GHz.

At the user side, two monopole receive antennas were mounted on the roof of a car with a distance of 15 cm. The channel was measured along two different routes in line of sight (LOS) and non line of sight (NLOS) conditions. The car was driving with a speed between 15 km/h and 25 km/h. The channel has been estimated once per LTE resource block (RB), i.e. once every 0.5 ms and 180 kHz. The measurement data covers a total bandwidth of 10 MHz. For our simulations, we utilized six measured channel traces, each with a length of 250 ms. We normalized the channel traces to an average power of $N_t \cdot N_r = 64 \cdot 2$, in order to be able to vary the SNR of the users. A few more details to the measurement setup can be found in [36], where we already utilized the same measurement data for ZF beamforming with $N_r = N_s = 1$.

### B. Simulation Setup

We utilize the measured channel traces for the simulation of multi-user MIMO transmission to $U = 6$ users with $N_s = 2$ streams per user, including explicit CSI feedback from the users. The measured channel traces provide the time-frequency selective channel transfer functions of an LTE compliant OFDM system. In our system model in Sec. II-A, we assume a frequency-flat channel corresponding to a single

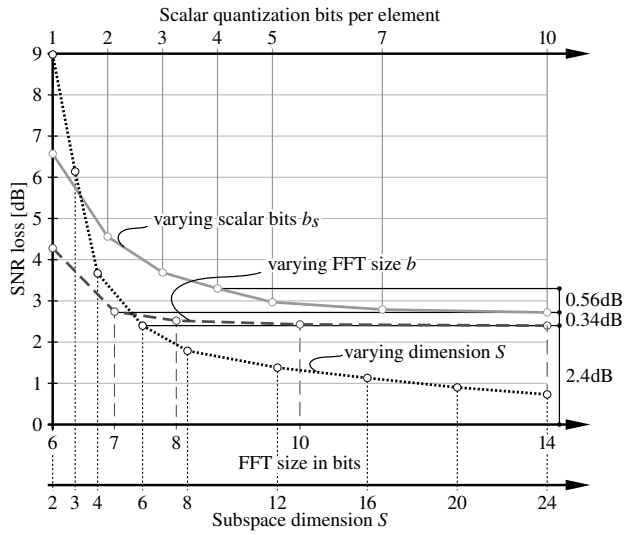



Fig. 10. Impact of the individual quantization parameters of the dual stage product codebook on the SNR loss w.r.t. unquantized CSIT.

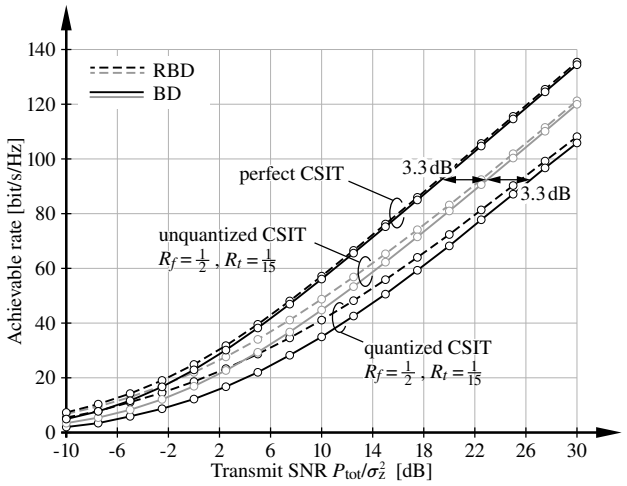



Fig. 11. Achievable transmission rate of the considered two-tier precoding architecture with perfect CSIT, unquantized outer-tier CSIT with feedback rates $R_f$ and $R_t$, and quantized outer-tier CSIT.

OFDM subcarrier. However, providing CSI feedback for each OFDM subcarrier imposes too much overhead. We therefore consider wideband CSI feedback in this section, calculating the channel eigenmodes $\mathbf{U}_j^{(N_s)}$ from an eigendecomposition of the channel correlation matrix, which is estimated by averaging over a time-window of 15 transmission time intervals (TTIs), i.e., $15 \cdot 0.5\,\text{ms} = 7.5\,\text{ms}$, and a frequency window of 2 MHz. We thus provide outer-tier CSI feedback with a feedback rate in time and frequency of $R_t = \frac{1}{7.5\,\text{ms}}$ and $R_f = \frac{1}{2\,\text{MHz}}$. These feedback intervals correspond approximately to the 92.5 % coherence time and bandwidth of the measured LOS channel traces, whereas for the NLOS traces the coherence already drops to approximately 85 %. As we will see below, the chosen time-frequency feedback granularity causes an SNR loss of 3.3 dB compared to perfect CSIT. We apply dual-stage CSI quantization as described in Sec. V, utilizing a DFT codebook for the first stage and scalar quantization as the second stage.

### C. Simulation Results

In our first simulation, we investigate the impact of the individual quantization parameters on the SNR loss w.r.t. unquantized CSIT. In Fig. 10, we exhibit the SNR loss w.r.t. unquantized CSIT as a function of different quantization parameters. Let us first consider the impact of the subspace dimension $S$ of the first quantization stage (the dotted line), assuming unquantized second stage CSIT and a very large oversampled FFT codebook for the first stage. The SNR loss is here caused by the error floor exhibited by the FFT codebook; as explained in Sec. V-C, this error floor can be reduced by increasing the intermediate subspace dimension $S$. We observe that the SNR loss drops very fast at the beginning for small $S$, but the gain then starts to diminish. Since we will apply the comparatively inefficient scalar quantization as the second quantization stage, it is preferable to select a relatively small $S$, to keep the quantization overhead of the second quantization stage within reasonable limits. For our remaining simulations we set $S = 6$, providing an SNR loss of approximately 2.4 dB.

With this fixed $S = 6$, we next vary the codebook size of the first quantization stage, as determined by the FFT size (the dashed line). We observe that there is a significant gain when going from the critically sampled FFT size of 64 to the oversampled FFT size of 128; however, beyond that value we hardly gain in SNR. We therefore fix the FFT size to 128 for the remaining simulations, giving an additional SNR loss of approximately 0.34 dB.

Finally, we investigate the impact of the scalar quantization bits $b_s$ (the solid line). As expected, the SNR loss diminishes with growing number of bits. For our remaining simulations, we employ $b_s = 4$, causing an additional SNR loss of approximately 0.56 dB and thus a total SNR loss of 3.3 dB w.r.t. unquantized CSIT.

With these parameter choices, we end up with a total feedback overhead of

$$\left(\log_2\left(\binom{N_{\text{DFT}}}{S}\right) + S\,N_s\,b_s\right) R_f R_t \approx 5.36\,\text{bit/ms/MHz}.$$

for the outer-tier CSI feedback. In Fig. 11, we show the actual achievable rate performance of the system as a function of the transmit SNR $P_{\text{tot}}/\sigma_z^2$. We apply inner-tier BD precoding with equal power allocation $P_u = P_{\text{tot}}/U$ (solid) as well as RBD precoding with water-filling power allocation amongst users and streams (dashed). We compare the performance with perfect CSIT, unquantized CSIT with the feedback rates $R_t$, $R_f$ as before, and imperfect CSIT with the quantization parameters as determined in our previous simulation. At high SNR, we observe the respective SNR losses according to our choices of feedback rates and quantization parameters. Since RBD and BD provide the same performance at high SNR, they both also exhibit the same SNR loss. Of course, to apply the inner-tier BD and RBD precoders, the system additionally has to supply inner-tier CSI feedback. As we demonstrate in our related conference paper [36], already established feedback methods, such as the differential manifold quantizers in [11, 12], can be utilized for this purpose.

## VIII. Conclusion

We proposed a novel dual stage Grassmannian quantization approach that enables efficient CSI quantization in massive

MIMO scenarios with low computational complexity. We provided an analytic characterization of the proposed dual stage product codebook design and evaluated its performance, revealing a relatively small degradation compared to single stage quantization. We furthermore applied the proposed product codebook for CSI quantization in a two-tier precoding architecture and calculated closed-form analytic expressions for the achievable transmission rate with imperfect outer-tier CSIT in the asymptotic regime $N_t \to \infty$. These results facilitate the selection of the quantization parameters to achieve a certain SNR loss compared to perfect CSIT. Our prospective future work includes the combination of the proposed methods with some form of frequency-independent channel reciprocity, such as, the reciprocity of the angular scattering function, to further reduce the required feedback overhead.

## Acknowledgements

The financial support by the Austrian Federal Ministry for Digital and Economic Affairs and the National Foundation for Research, Technology and Development is gratefully acknowledged.

## Appendix A
## Proof of Theorem 1

To prove Th. 1, we first consider the following decomposition of $\hat{\mathbf{B}}_j$ defined in (20)

$$\begin{aligned}\hat{\mathbf{B}}_j &= \mathbf{B}_j\mathbf{B}_j^{\mathrm{H}}\hat{\mathbf{B}}_j + \left(\mathbf{I}_S - \mathbf{B}_j\mathbf{B}_j^{\mathrm{H}}\right)\hat{\mathbf{B}}_j,\\ &= \mathbf{B}_j\mathbf{B}_j^{\mathrm{H}}\hat{\mathbf{B}}_j + \mathbf{B}_j^{\perp}\left(\mathbf{B}_j^{\perp}\right)^{\mathrm{H}}\hat{\mathbf{B}}_j, \end{aligned} \tag{55}$$

where $\mathbf{B}_j^{\perp}$ denotes a basis for the orthogonal complement of $\mathrm{span}\left(\mathbf{B}_j\right)$. Due to the construction of $\mathbf{B}_j$, we have

$$\begin{aligned}&\mathrm{tr}\left(\left(\mathbf{U}_j^{(N_s)}\right)^{\mathrm{H}}\hat{\mathbf{U}}_j^{(S)}\left(\hat{\mathbf{U}}_j^{(S)}\right)^{\mathrm{H}}\mathbf{U}_j^{(N_s)}\right) =\\ &\mathrm{tr}\left(\left(\mathbf{U}_j^{(N_s)}\right)^{\mathrm{H}}\hat{\mathbf{U}}_j^{(S)}\left(\mathbf{B}_j\mathbf{B}_j^{\mathrm{H}} + \mathbf{B}_j^{\perp}\left(\mathbf{B}_j^{\perp}\right)^{\mathrm{H}}\right)\left(\hat{\mathbf{U}}_j^{(S)}\right)^{\mathrm{H}}\mathbf{U}_j^{(N_s)}\right)\\ &= \mathrm{tr}\left(\left(\mathbf{U}_j^{(N_s)}\right)^{\mathrm{H}}\hat{\mathbf{U}}_j^{(S)}\mathbf{B}_j\mathbf{B}_j^{\mathrm{H}}\left(\hat{\mathbf{U}}_j^{(S)}\right)^{\mathrm{H}}\mathbf{U}_j^{(N_s)}\right), \end{aligned} \tag{56}$$

where the second equality follows from (18). We conclude

$$\left\|\left(\mathbf{U}_j^{(N_s)}\right)^{\mathrm{H}}\hat{\mathbf{U}}_j^{(S)}\mathbf{B}_j^{\perp}\right\| = 0 \Leftrightarrow \left(\mathbf{U}_j^{(N_s)}\right)^{\mathrm{H}}\hat{\mathbf{U}}_j^{(S)}\mathbf{B}_j^{\perp} = \mathbf{0}. \tag{57}$$

Now consider the distortion of the dual stage codebook

$$\begin{aligned}&N_s - d_c^2\left(\mathbf{U}_j^{(N_s)}, \hat{\mathbf{U}}_{j,\text{dual}}^{(N_s)}\right) =\\ &\mathrm{tr}\left(\left(\mathbf{U}_j^{(N_s)}\right)^{\mathrm{H}}\hat{\mathbf{U}}_j^{(S)}\hat{\mathbf{B}}_j\hat{\mathbf{B}}_j^{\mathrm{H}}\left(\hat{\mathbf{U}}_j^{(S)}\right)^{\mathrm{H}}\mathbf{U}_j^{(N_s)}\right) =\\ &\mathrm{tr}\left(\left(\mathbf{U}_j^{(N_s)}\right)^{\mathrm{H}}\hat{\mathbf{U}}_j^{(S)}(\mathbf{B}_j\mathbf{B}_j^{\mathrm{H}}\hat{\mathbf{B}}_j)(\hat{\mathbf{B}}_j^{\mathrm{H}}\mathbf{B}_j\mathbf{B}_j^{\mathrm{H}})\left(\hat{\mathbf{U}}_j^{(S)}\right)^{\mathrm{H}}\mathbf{U}_j^{(N_s)}\right), \end{aligned} \tag{58}$$

where we utilized the decomposition (55) and exploited (57) to eliminate the terms involving $\mathbf{B}_j^{\perp}$. Next, consider the product of the four matrices in the center

$$\begin{aligned}&\mathbb{E}\left(\mathrm{tr}\left(\mathbf{B}_j^{\mathrm{H}}\hat{\mathbf{B}}_j\hat{\mathbf{B}}_j^{\mathrm{H}}\mathbf{B}_j\right)\right) =\\ &N_s - \mathbb{E}\left(d_c^2\left(\mathbf{B}_j, \hat{\mathbf{B}}_j\right)\right) = N_s - \bar{\mathrm{d}}_{c,2}^2 \end{aligned} \tag{59}$$

For isotropically distributed $\hat{\mathbf{B}}_j$, i.e. RSQ, and/or $\mathbf{B}_j$ it furthermore holds that

$$\mathbb{E}\left(\mathbf{B}_j^{\mathrm{H}}\hat{\mathbf{B}}_j\hat{\mathbf{B}}_j^{\mathrm{H}}\mathbf{B}_j\right) = \left(1 - \frac{\bar{\mathrm{d}}_{c,2}^2}{N_s}\right)\mathbf{I}_{N_s}. \tag{60}$$

Plugging (60) into (58) and taking the expectation, we get

$$\begin{aligned}&N_s - \bar{\mathrm{d}}_{c,\text{dual}}^2 = \left(1 - \frac{\bar{\mathrm{d}}_{c,2}^2}{N_s}\right)\cdot\\ &\mathbb{E}\left(\mathrm{tr}\left(\left(\mathbf{U}_j^{(N_s)}\right)^{\mathrm{H}}\hat{\mathbf{U}}_j^{(S)}\mathbf{B}_j\mathbf{B}_j^{\mathrm{H}}\left(\hat{\mathbf{U}}_j^{(S)}\right)^{\mathrm{H}}\mathbf{U}_j^{(N_s)}\right)\right) \end{aligned} \tag{61}$$

$$\Rightarrow \bar{\mathrm{d}}_{c,\text{dual}}^2 = N_s - \left(1 - \frac{\bar{\mathrm{d}}_{c,2}^2}{N_s}\right)\left(N_s - \bar{\mathrm{d}}_{c,1}^2\right). \quad \square \tag{62}$$

## Appendix B
## Proof of Theorem 2

Consider the achievable transmission rate of the considered ZF transmission (44) with the power constraint (47)

$$R_u = \mathbb{E}\left(\log_2\left(1 + \frac{P_u\sigma_u^2}{\sigma_z^2\left[\left(\mathbf{H}_o\mathbf{H}_o^{\mathrm{H}}\right)^{-1}\right]_{u,u}}\right)\right). \tag{63}$$

For massive MIMO with $N_t \to \infty$, the values inside the logarithm tend towards their expected values and, hence, the rate is determined by the average output SNR

$$\frac{P_u\,\mathbb{E}\left(\sigma_u^2\right)}{\sigma_z^2\,\mathbb{E}\left(\left[\left(\mathbf{H}_o\mathbf{H}_o^{\mathrm{H}}\right)^{-1}\right]_{u,u}\right)}. \tag{64}$$

For massive MIMO and the channel normalization considered in (3), the squared-singular value $\sigma_u^2$ goes to one, due to channel hardening [47]

$$\mathbb{E}\left(\sigma_u^2\right) \overset{N_t\to\infty}{\longrightarrow} \sigma_u^2 \overset{N_t\to\infty}{\longrightarrow} 1. \tag{65}$$

The SNR is thus determined by the $u$-th diagonal element of the positive-definite matrix $\left(\mathbf{H}_o\mathbf{H}_o^{\mathrm{H}}\right)^{-1}$.

When considering the limiting case $N_t \to \infty$, it is common to invoke a mutual orthogonality condition, such as, $|\mathbf{u}_i^{\mathrm{H}}\hat{\mathbf{u}}_j|^2 \to 0$, and to simply assume that $\mathbf{H}_o\mathbf{H}_o^{\mathrm{H}} \to \mathbf{I}_U$ and, hence, $\left(\mathbf{H}_o\mathbf{H}_o^{\mathrm{H}}\right)^{-1} \to \mathbf{I}_U$. This approach is valid if the number of users $U$ is a constant; yet, this assumption does not provide an accurate result if $U$ scales linearly with $N_t$ and thus the dimensions of $\mathbf{H}_o\mathbf{H}_o^{\mathrm{H}}$ also grow unbounded. In this case, it is still true that the off-diagonal elements of $\mathbf{H}_o\mathbf{H}_o^{\mathrm{H}}$ go to zero; however, since their number goes to infinity at the same rate, the off-diagonal elements of $\left(\mathbf{H}_o\mathbf{H}_o^{\mathrm{H}}\right)^{-1}$ are not negligible.

According to [45], $\left[\left(\mathbf{H}_o\mathbf{H}_o^{\mathrm{H}}\right)^{-1}\right]_{u,u}$ can be upper bounded as follows

$$\left[\left(\mathbf{H}_o\mathbf{H}_o^{\mathrm{H}}\right)^{-1}\right]_{u,u} \le \frac{1}{4\left[\mathbf{H}_o\mathbf{H}_o^{\mathrm{H}}\right]_{u,u}}\left(\frac{\alpha}{\beta} + \frac{\beta}{\alpha} + 2\right), \tag{66}$$

where $\alpha$ denotes a lower bound on the smallest eigenvalue of $\left(\mathbf{H}_o\mathbf{H}_o^{\mathrm{H}}\right)$ and $\beta$ is an upper bound on the largest eigenvalue. Such upper and lower bounds on the smallest and largest eigenvalues have been published in [48]

$$\lambda_{\min}\left(\mathbf{H}_o\mathbf{H}_o^{\mathrm{H}}\right) \ge m - s\sqrt{U-1}, \tag{67}$$

$$\lambda_{\max}\left(\mathbf{H}_o\mathbf{H}_o^{\mathrm{H}}\right) \leq m + s\sqrt{U-1}, \tag{68}$$

$$m = \frac{\mathrm{tr}\left(\mathbf{H}_o\mathbf{H}_o^{\mathrm{H}}\right)}{U}, \quad s^2 = \frac{\mathrm{tr}\left((\mathbf{H}_o\mathbf{H}_o^{\mathrm{H}})^2\right)}{U} - m^2, \tag{69}$$

where $m$ is the mean of the eigenvalues and $s^2$ is their variance. To calculate these bounds for massive MIMO, we have to determine the diagonal elements of $\mathbf{H}_o\mathbf{H}_o^{\mathrm{H}}$ and $(\mathbf{H}_o\mathbf{H}_o^{\mathrm{H}})^2$ for $N_t \to \infty$, which we perform further below. However, it turns out that these bounds are far from tight in our situation; in fact, depending on $\bar{\mathrm{d}}_c^2$, (67) can equate to a negative lower bound for the strictly positive eigenvalues of $\mathbf{H}_o\mathbf{H}_o^{\mathrm{H}}$.

We thus propose to evaluate (66) employing "typical" eigenvalues to acquire an estimate of $\left[(\mathbf{H}_o\mathbf{H}_o^{\mathrm{H}})^{-1}\right]_{u,u}$

$$\left[(\mathbf{H}_o\mathbf{H}_o^{\mathrm{H}})^{-1}\right]_{u,u} \approx \frac{1}{4\left[\mathbf{H}_o\mathbf{H}_o^{\mathrm{H}}\right]_{u,u}}\left(\frac{\lambda^-}{\lambda^+} + \frac{\lambda^+}{\lambda^-} + 2\right), \tag{70}$$

where $\lambda^+, \lambda^-$ denote eigenvalues that lie one standard deviation above and below the mean[5], i.e.

$$\lambda^+ = m + s, \quad \lambda^- = m - s. \tag{71}$$

We now proceed with the calculation of the elements $\left[\mathbf{H}_o\mathbf{H}_o^{\mathrm{H}}\right]_{k,\ell}$ for $k \neq \ell$

$$\left[\mathbf{H}_o\mathbf{H}_o^{\mathrm{H}}\right]_{k,\ell} = \sum_{i=1}^{U} \mathbf{u}_k^{\mathrm{H}}\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i^{\mathrm{H}}\mathbf{u}_\ell = \mathbf{u}_k^{\mathrm{H}}\hat{\mathbf{u}}_k\hat{\mathbf{u}}_k^{\mathrm{H}}\mathbf{u}_\ell + \mathbf{u}_k^{\mathrm{H}}\hat{\mathbf{u}}_\ell\hat{\mathbf{u}}_\ell^{\mathrm{H}}\mathbf{u}_\ell + \sum_{i\neq\{k,\ell\}} \mathbf{u}_k^{\mathrm{H}}\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i^{\mathrm{H}}\mathbf{u}_\ell. \tag{72}$$

For $N_t \to \infty$, we have $d_c^2\left(\mathbf{u}_k, \hat{\mathbf{u}}_k\right) \to \bar{\mathrm{d}}_c^2$. With this, we decompose the CSIT estimate according to: $\hat{\mathbf{u}}_k = \mathbf{u}_k\sqrt{1-\bar{\mathrm{d}}_c^2} + \mathbf{u}_k^\perp\bar{\mathrm{d}}_c$, where $\mathbf{u}_k^{\mathrm{H}}\mathbf{u}_k^\perp = 0$. This leads to

$$\mathbf{u}_k^{\mathrm{H}}\hat{\mathbf{u}}_k\hat{\mathbf{u}}_k^{\mathrm{H}}\mathbf{u}_\ell = (1-\bar{\mathrm{d}}_c^2)\mathbf{u}_k^{\mathrm{H}}\mathbf{u}_\ell + \sqrt{1-\bar{\mathrm{d}}_c^2}\bar{\mathrm{d}}_c\left(\mathbf{u}_k^\perp\right)^{\mathrm{H}}\mathbf{u}_\ell.$$

Applying a similar decomposition to $\hat{\mathbf{u}}_\ell$, we can further develop (72) as

$$\left[\mathbf{H}_o\mathbf{H}_o^{\mathrm{H}}\right]_{k,\ell} \to \mathbf{u}_k^{\mathrm{H}}\left(2(1-\bar{\mathrm{d}}_c^2) + \sum_{i\neq\{k,\ell\}}\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i^{\mathrm{H}}\right)\mathbf{u}_\ell + \sqrt{1-\bar{\mathrm{d}}_c^2}\bar{\mathrm{d}}_c\left(\left(\mathbf{u}_k^\perp\right)^{\mathrm{H}}\mathbf{u}_\ell + \mathbf{u}_k^{\mathrm{H}}\mathbf{u}_\ell^\perp\right). \tag{73}$$

For the diagonal elements $\left[\mathbf{H}_o\mathbf{H}_o^{\mathrm{H}}\right]_{k,k}$, with $k = \ell$, the calculation is very similar, with the only difference that in (72) there is then only one term in front of the summation

$$\left[\mathbf{H}_o\mathbf{H}_o^{\mathrm{H}}\right]_{k,k} \to (1-\bar{\mathrm{d}}_c^2) + \frac{U-1}{N_t}, \tag{74}$$

where the terms $\mathbf{u}_k^{\mathrm{H}}\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i^{\mathrm{H}}\mathbf{u}_k, \forall i \neq k$ follow a beta-distribution $\beta\left(1, N_t - 1\right)$ and thus tend towards $1/N_t$. From this result and the mean of the eigenvalues according to (69), we get (51).

The diagonal elements $\left[(\mathbf{H}_o\mathbf{H}_o^{\mathrm{H}})^2\right]_{k,k}$ are obtained from

$$\left[(\mathbf{H}_o\mathbf{H}_o^{\mathrm{H}})^2\right]_{k,k} = \sum_{\ell=1}^{U}\left[\mathbf{H}_o\mathbf{H}_o^{\mathrm{H}}\right]_{k,\ell}\left[\mathbf{H}_o\mathbf{H}_o^{\mathrm{H}}\right]_{\ell,k}$$

[5]Notice, if we assume $\lambda^+ \approx m$ and $\lambda^- \approx m$ we again end-up with the inaccurate result $\left(\mathbf{H}_o\mathbf{H}_o^{\mathrm{H}}\right)^{-1} \to \mathbf{I}_U$.

$$= \sum_{\ell=1}^{U}\left|\left[\mathbf{H}_o\mathbf{H}_o^{\mathrm{H}}\right]_{k,\ell}\right|^2. \tag{75}$$

since $\left[\mathbf{H}_o\mathbf{H}_o^{\mathrm{H}}\right]_{\ell,k}$ is the complex-conjugate of $\left[\mathbf{H}_o\mathbf{H}_o^{\mathrm{H}}\right]_{k,\ell}$. When evaluating the squared absolute value of (73), the mix-terms involving $\mathbf{u}_k, \mathbf{u}_k^\perp$ and $\mathbf{u}_\ell, \mathbf{u}_\ell^\perp$ are negligible in the massive MIMO limit. Only the square-terms are relevant

$$\left|\mathbf{u}_k^{\mathrm{H}}\left(2(1-\bar{\mathrm{d}}_c^2) + \sum_{i\neq\{k,\ell\}}\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i^{\mathrm{H}}\right)\mathbf{u}_\ell\right|^2 \longrightarrow \frac{1}{N_t}\left(4(1-\bar{\mathrm{d}}_c^2)^2 + 4(1-\bar{\mathrm{d}}_c^2)\frac{U-2}{N_t} + \frac{U-2}{N_t}\right),$$

$$\left|\sqrt{1-\bar{\mathrm{d}}_c^2}\bar{\mathrm{d}}_c\left(\left(\mathbf{u}_k^\perp\right)^{\mathrm{H}}\mathbf{u}_\ell + \mathbf{u}_k^{\mathrm{H}}\mathbf{u}_\ell^\perp\right)\right|^2 \to \frac{2}{N_t}(1-\bar{\mathrm{d}}_c^2)\bar{\mathrm{d}}_c^2, \ \ell \neq k. \tag{76}$$

Plugging (74) and (76) into (75), we get

$$\left[(\mathbf{H}_o\mathbf{H}_o^{\mathrm{H}})^2\right]_{k,k} \to \left((1-\bar{\mathrm{d}}_c^2) + \frac{U-1}{N_t}\right)^2 + \frac{U-1}{N_t}\cdot\left(4(1-\bar{\mathrm{d}}_c^2)^2 + 2(1-\bar{\mathrm{d}}_c^2)\bar{\mathrm{d}}_c^2 + 4(1-\bar{\mathrm{d}}_c^2)\frac{U-2}{N_t} + \frac{U-2}{N_t}\right). \tag{77}$$

With this, we can finally calculate the variance of the eigenvalues (69) as given in (52). □

## REFERENCES


[1] F. Rusek, D. Persson, B. K. Lau, E. Larsson, T. Marzetta, O. Edfors, and F. Tufvesson, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Processing Magazine*, vol. 30, no. 1, pp. 40–60, Jan 2013.

[2] O. Elijah, C. Y. Leow, T. A. Rahman, S. Nunoo, and S. Z. Iliya, "A comprehensive survey of pilot contamination in massive MIMO – 5G system," *IEEE Communications Surveys Tutorials*, vol. 18, no. 2, pp. 905–923, 2016.

[3] M. B. Khalilsarai, S. Haghighatshoar, X. Yi, and G. Caire, "FDD massive MIMO via UL/DL channel covariance extrapolation and active channel sparsification," *CoRR*, vol. abs/1803.05754, 2018.

[4] D. Love and R. Heath, Jr., "Limited feedback unitary precoding for spatial multiplexing systems," *IEEE Transactions on Information Theory*, vol. 51, no. 8, pp. 2967–2976, 2005.

[5] N. Jindal, "MIMO broadcast channels with finite-rate feedback," *IEEE Transactions on Information Theory*, vol. 52, no. 11, p. 5, Nov. 2006.

[6] N. Jindal, "Antenna combining for the MIMO downlink channel," *IEEE Trans. on Wireless Comm.*, vol. 7, no. 10, pp. 3834 –3844, Oct. 2008.

[7] N. Ravindran and N. Jindal, "Limited feedback-based block diagonalization for the MIMO broadcast channel," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 8, pp. 1473 –1482, Oct. 2008.

[8] S. Schwarz and M. Rupp, "Subspace quantization based combining for limited feedback block-diagonalization," *IEEE Transactions on Wireless Communications*, vol. 12, no. 11, pp. 5868–5879, 2013.

[9] D. Sacristan-Murga, M. Payaro, and A. Pascual-Iserte, "Transceiver design framework for multiuser MIMO-OFDM broadcast systems with channel Gram matrix feedback," *IEEE Transactions on Wireless Communications*, vol. 11, no. 5, pp. 1774–1787, May 2012.

[10] O. El Ayach and R. Heath Jr., "Grassmannian differential limited feedback for interference alignment," *IEEE Transactions on Signal Processing*, vol. 60, no. 12, pp. 6481–6494, Dec 2012.

[11] S. Schwarz, R. Heath, Jr., and M. Rupp, "Adaptive quantization on a Grassmann-manifold for limited feedback beamforming systems," *IEEE Trans. on Signal Processing*, vol. 61, no. 18, pp. 4450–4462, 2013.

[12] S. Schwarz and M. Rupp, "Predictive quantization on the Stiefel manifold," *IEEE Signal Proc. Letters*, vol. 22, no. 2, pp. 234–238, 2015.

[13] D. J. Love and R. W. Heath, "Limited feedback diversity techniques for correlated channels," *IEEE Transactions on Vehicular Technology*, vol. 55, no. 2, pp. 718–722, March 2006.

[14] P. Xia and G. B. Giannakis, "Design and analysis of transmit-beamforming based on limited-rate feedback," *IEEE Transactions on Signal Processing*, vol. 54, no. 5, pp. 1853–1863, May 2006.

[15] V. Raghavan, R. W. Heath, and A. M. Sayeed, "Systematic codebook designs for quantized beamforming in correlated MIMO channels," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 7, pp. 1298–1310, Sep. 2007.

[16] T. Shuang, T. Koivisto, H. Maattanen, K. Pietikainen, T. Roman, and M. Enescu, "Design and evaluation of LTE-advanced double codebook," in *IEEE 73rd Vehicular Technology Conference*, pp. 1–5, May 2011.

[17] 3GPP, "Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures (Release 14)," June 2019, [Online]. Available: http://www.3gpp.org/ftp/Specs/html-info/36213.htm.

[18] J. Choi, Z. Chance, D. J. Love, and U. Madhow, "Noncoherent trellis coded quantization: A practical limited feedback technique for massive MIMO systems," *IEEE Transactions on Communications*, vol. 61, no. 12, pp. 5016–5029, December 2013.

[19] Z. Lv and Y. Li, "A channel state information feedback algorithm for massive MIMO systems," *IEEE Communications Letters*, vol. 20, no. 7, pp. 1461–1464, July 2016.

[20] T. Wang, C. Wen, S. Jin, and G. Y. Li, "Deep learning-based CSI feedback approach for time-varying massive MIMO channels," *IEEE Wireless Communications Letters*, vol. 8, no. 2, pp. 416–419, April 2019.

[21] X. Luo, P. Cai, X. Zhang, D. Hu, and C. Shen, "A scalable framework for CSI feedback in FDD massive MIMO via DL path aligning," *IEEE Trans. on Signal Processing*, vol. 65, no. 18, pp. 4702–4716, Sep. 2017.

[22] H. Xie, F. Gao, S. Zhang, and S. Jin, "A unified transmission strategy for TDD/FDD massive MIMO systems with spatial basis expansion model," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 4, pp. 3170–3184, April 2017.

[23] J. Chen and V. K. N. Lau, "Two-tier precoding for FDD multi-cell massive MIMO time-varying interference networks," *IEEE Journal on Sel. Areas in Communications*, vol. 32, no. 6, pp. 1230–1238, June 2014.

[24] A. Alkhateeb, G. Leus, and R. W. Heath, "Multi-layer precoding: A potential solution for full-dimensional massive MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 16, no. 9, pp. 5810–5824, Sep. 2017.

[25] S. Schwarz, "Robust full-dimension MIMO transmission based on limited feedback angular-domain CSIT," *EURASIP Journal on Wireless Communications and Networking*, vol. 2018, no. 1, pp. 1–20, Mar 2018.

[26] P. Almers, E. Bonek, A. Burr, N. Czink, M. Debbah, V. Degli-Esposti, H. Hofstetter, P. Kyösti, D. Laurenson, G. Matz, A. Molisch, C. Oestges, and H. Özcelik, "Survey of channel and radio propagation models for wireless MIMO systems," *EURASIP Journal on Wireless Communications and Networking*, vol. 2007, p. 19, 2007.

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

[28] E. Zöchmann, S. Caban, C. F. Mecklenbräuker, S. Pratschner, M. Lerch, S. Schwarz, and M. Rupp, "Better than Rician: Modelling millimetre wave channels as Two-Wave with Diffuse Power," *ArXiv e-prints*, Apr. 2018, under review at EURASIP JWCN.

[29] Q. Spencer, A. Swindlehurst, and M. Haardt, "Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels," *IEEE Trans. on Signal Processing*, vol. 52, no. 2, pp. 461 – 471, Feb. 2004.

[30] V. Stankovic and M. Haardt, "Generalized design of multi-user MIMO precoding matrices," *IEEE Transactions on Wireless Communications*, vol. 7, no. 3, pp. 953–961, 2008.

[31] T. Inoue and R. Heath, Jr., "Grassmannian predictive coding for delayed limited feedback MIMO systems," in *47th Annual Allerton Conference on Communication, Control, and Computing*, Oct. 2009.

[32] D. Zhu, Y. Zhang, G. Wang, and M. Lei, "Grassmannian subspace prediction for precoded spatial multiplexing MIMO with delayed feedback," *IEEE Sig. Proc. Letters*, vol. 18, no. 10, pp. 555 – 558, Oct. 2011.

[33] S. Schwarz, R. Heath, Jr., and M. Rupp, "Adaptive quantization on the Grassmann-manifold for limited feedback multi-user MIMO systems," in *38th International Conference on Acoustics, Speech and Signal Processing*, pp. 5021 – 5025, Vancouver, Canada, May 2013.

[34] S. Schwarz and M. Rupp, "Evaluation of distributed multi-user MIMO-OFDM with limited feedback," *IEEE Transactions on Wireless Communications*, vol. 13, no. 11, pp. 6081–6094, Aug. 2014.

[35] G. Caire, N. Jindal, M. Kobayashi, and N. Ravindran, "Multiuser MIMO achievable rates with downlink training and channel state feedback," *IEEE Transactions on Information Theory*, vol. 56, no. 6, pp. 2845–2866, June 2010.

[36] S. Schwarz, M. Rupp, and S. Wesemann, "Two-tier Grassmannian limited feedback for FD-MIMO with concatenated precoding," 2019, under review at IEEE International Symposium on Personal, Indoor and Mobile Radio Communications.

[37] 3GPP, "Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Channels and Modulation (Release 15)," June 2019, [Online]. Available: http://www.3gpp.org/ftp/Specs/html-info/36211.htm.

[38] H. Ji, Y. Kim, J. Lee, E. Onggosanusi, Y. Nam, J. Zhang, B. Lee, and B. Shim, "Overview of full-dimension MIMO in LTE-advanced pro," *IEEE Communications Mag.*, vol. 55, no. 2, pp. 176–184, February 2017.

[39] F. Zhang, S. Sun, Q. Gao, and H. Li, "Hybrid CSI-RS transmission mechanism-based 3D beamforming scheme for FDD massive MIMO system," *China Communications*, vol. 13, pp. 109–119, N 2016.

[40] W. Dai, Y. Liu, and B. Rider, "Quantization bounds on Grassmann manifolds and applications to MIMO communications," *IEEE Trans. on Information Theory*, vol. 54, no. 3, pp. 1108 –1123, March 2008.

[41] S. Schwarz and M. Rupp, "Antenna combiners for block-diagonalization based multi-user MIMO with limited feedback," in *IEEE International Conf. on Communications*, pp. 127–132, Budapest, June 2013.

[42] F. Boccardi, H. Huang, and A. Alexiou, "Hierarchical quantization and its application to multiuser eigenmode transmissions for MIMO broadcast channels with limited feedback," in *IEEE 18th International Symposium on Personal, Indoor and Mobile Radio Communications*, pp. 1–5, Sep. 2007.

[43] A. Decurninge and M. Guillaud, "Cube-split: Structured quantizers on the Grassmannian of lines," in *IEEE Wireless Communications and Networking Conference*, pp. 1–6, March 2017.

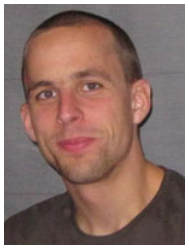
[44] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[45] Z. Bai and G. H. Golub, "Bounds for the trace of the inverse and the determinant of symmetric positive definite matrices," *Baltzer Journals*, pp. 1–9, Apr. 1996.

[46] D. Phan-Huy, S. Wesemann, J. Bjoersell, and M. Sternad, "Adaptive massive MIMO for fast moving connected vehicles: It will work with predictor antennas!" in *WSA 2018; 22nd International ITG Workshop on Smart Antennas*, pp. 1–8, March 2018.

[47] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Aspects of favorable propagation in massive MIMO," in *22nd European Signal Processing Conference (EUSIPCO)*, pp. 76–80, Sept 2014.

[48] H. Wolkowicz and G. P. Styan, "Bounds for eigenvalues using traces," *Linear Algebra and its Applications*, vol. 29, pp. 471 – 506, 1980.


**Stefan Schwarz** received his Dipl.-Ing. degree in electrical engineering in 2009, his Dr. techn. degree in telecommunications engineering in 2013, and his habilitation in the field of mobile communications in 2019, all from Vienna University of Technology – TU Wien. He currently holds a tenure track position at the Institute of Telecommunications of TU Wien, where he heads the Christian Doppler Laboratory for Dependable Wireless Connectivity for the Society in Motion. Email: sschwarz@nt.tuwien.ac.at



**Markus Rupp** received his Dipl.-Ing. degree in 1988 at the University of Saarbrücken and his Dr.-Ing. degree in 1993 at Technische Universität Darmstadt. Until 1995 he had a postdoctoral position at the University of Santa Barbara, California. From 1995 to 2001 he was with the Wireless Technology Research Department of Bell-Labs, NJ. Since October 2001 he is a full professor for Digital Signal Processing in Mobile Communications at TU Wien. Email: mrupp@nt.tuwien.ac.at



**Stefan Wesemann** received his Dipl.-Ing. degree in 2006 in Information Systems Engineering, and the Dr.-Ing. in 2014 in Telecommunications, both from Dresden University of Technology. From 2007 to 2010, he was a Senior Engineer at Signalion GmbH (now National Instruments). In 2014, he joined Nokia Bell Labs as a core team member of the FutureCell (F-Cell) project. Since 2016 he is part of the Next Generation Wireless Systems department. Email: stefan.wesemann@nokia-bell-labs.com