NEAR CAPACITY TOMLINSON-HARASHIMA PRECODING WITH UNIFORM STREAM PROPERTIES

Raphael Hunger, Maximilian Riemensberger, and Wolfgang Utschick

Associate Institute for Signal Processing
Munich University of Technology, 80290 Munich, Germany
hunger@tum.de

ABSTRACT

We address the nonlinear transceiver design in a point-to-point MIMO system with *Tomlinson-Harashima precoding* (THP). By jointly optimizing both transmitter and receiver, capacity can be achieved up to the shaping loss and each stream can be decoded separately. In contrast to linear filtering, THP allows for uniform stream properties rendering bit-loading unnecessary and allows to span an arbitrarily high number of streams for the sake of a reduced cardinality modulation alphabet. Existing work studied either the decision feedback equalizer version with the nonlinearity located at the receiver or the perfect dirty paper precoding where the *geometric mean decomposition* (GMD) can be applied. We explicitly take the modulo operator into account leading to the fact that the *generalized triangular decomposition* has to be applied instead of the GMD.

1. INTRODUCTION

Recently, decision feedback aided nonlinear filtering was shown to achieve capacity in point-to-point systems with the nice characteristic of uniform stream properties and minimum sum mean square error [1, 2]. Thus, undesirable bit-loading and different coding on each stream, which is necessary for linear filtering to achieve capacity, become obsolete. The nonlinear structure can also be shifted to the transmitter: in [2], the authors also consider perfect *dirty paper* (DP) precoding for which stream-coding must be done jointly leading to a highly complex precoder. We deploy *Tomlinson Harashima Precoding* (THP) as a low-complexity practical implementation of DP. For the THP system, we show that capacity is achieved except for the shaping loss [3]. Moreover, a uniform stream property is achieved in the relevant SNR region and an arbitrarily high number of streams can be transmitted for the sake of a reduced cardinality modulation alphabet. However, the main difference to the work [1, 2] is that for THP, the so called *generalized triangular decomposition* (GTD) [4] has to be applied instead of the *geometric mean decomposition* (GMD) [5] which in contrast to the GMD not necessarily exists. The GTD must be used due to the fact that the modulo operator does not affect the last stream which therefore has a different variance. Both GTD and GMD are matrix factorizations where a matrix is decomposed into a the product of a unitary matrix, an upper/lower triangular matrix with prescribed main diagonal, and another unitary matrix.

In Section 2, we discuss the system model and briefly review the *linear* capacity achieving filtering in Section 3. Afterwards, we show in Section 4 that the capacity achieving precoder in THP has enough degrees of freedom left to let all individual streams have uniform properties in the relevant SNR region. To this end, the GTD is utilized, the properties and construction of which are explained in Section 5. Finally, simulation results in Section 6 show that THP can clearly outperform the decision feedback aided system despite the power loss and the modulo loss since THP does not suffer from error propagation.

Notation: Matrices and vectors are denoted by upper and lower case bold italic letters, respectively. The operators $\mathrm{E}[\cdot]$, $(\cdot)^{\mathrm{H}}$, $(\cdot)^{\mathrm{T}}$, $\mathrm{tr}(\cdot)$, $\det(\cdot)$, and $[\cdot]_{i,j}$ denote expectation with respect to the noise $\boldsymbol{\eta}$ and the signal $\boldsymbol{v}$, conjugate transposition, transposition, trace, determinant, and the matrix element in the ith row and jth column, respectively. The set $\mathbb{T}^{K\times K} \subset \mathbb{C}^{K\times K}$ contains all $K \times K$ strictly upper triangular matrices, $\mathbb{A}$ represents the discrete or continuous modulation alphabet, $\mathbb{S}_{+}^{K\times K} \subset \mathbb{C}^{K\times K}$ denotes the cone of the $K \times K$ positive semidefinite matrices, $\mathbb{D}^{K\times r} \subset \mathbb{R}_{+,0}^{K\times r}$ is the set of all (not necessarily square) diagonal matrices, and the set of all $K \times K$ unitary matrices is denoted by $\mathbb{U}^{K\times K} \subset \mathbb{C}^{K\times K}$. $\Re\{\cdot\}$, $\Im\{\cdot\}$, and j denote real and imaginary part of the complex argument, and the imaginary unit $\sqrt{-1}$, respectively. $\succcurlyeq$ is a partial order on the proper cone of positive semidefinite matrices.

2. SYSTEM MODEL

We consider a point-to-point MIMO communication link as depicted in Figure 1 where the zero-mean data vector $\boldsymbol{s} \in \mathbb{A}^K$ with identity covariance matrix contains the modulated symbols of the K streams. The feedback filter $\boldsymbol{F} \in \mathbb{T}^{K\times K}$ vanishes in case of linear precoding whereas it is restricted to

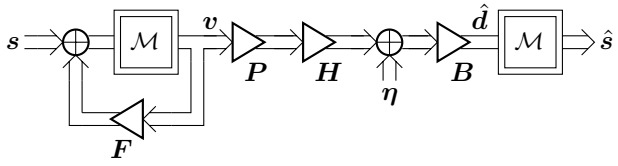

Fig. 1. Nonlinear point-to-point system model.

be *strictly* upper[1] triangular (with zeros on its main diagonal) when THP is applied. The modulo operator $\mathcal{M}(\cdot) : \mathbb{C}^K \rightarrow \mathbb{V}^K$ with $\mathbb{V} = \{z \in \mathbb{C} | -\tau \leq 2\Re\{z\} < \tau,\ -\tau \leq 2\Im\{z\} < \tau\}$ is an element-wise many-to-one mapping from the entire K-dimensional complex hyper-plane into the half-open complex hyper-cube by an element-wise addition of integer multiples of the modulo constant τ to both real and imaginary parts of its argument such that its image lies in $\mathbb{V}^K$. We make the common statistical assumptions on the output $\boldsymbol{v} \in \mathbb{V}^K$ of the modulo operator [6], i.e., we assume that the covariance matrix $\boldsymbol{C_v} = \mathrm{E}[\boldsymbol{v}\boldsymbol{v}^{\mathrm{H}}] \in \mathbb{S}_+^{K\times K}$ is diagonal which means that the individual entries are uncorrelated. Furthermore, the last stream is not affected by the modulo operator when the modulo constant τ is chosen sufficiently large, hence its variance remains one. In case of an M-ary QAM modulation with $M = 4^n, n \in \mathbb{N}$, the first $K-1$ diagonal entries of $\boldsymbol{C_v}$ are assumed to have variance $\sigma_v^2 = \frac{M}{M-1}$, see [6], *if* the respective feedback is *active*. This follows from the assumption that $[\boldsymbol{v}]_{1,1}, \ldots, [\boldsymbol{v}]_{K-1,1}$ are uniformly distributed over $\mathbb{V}$. The signal $\boldsymbol{v}$ is linearly precoded by $\boldsymbol{P} \in \mathbb{C}^{N_{\mathrm{Tx}}\times K}$ and propagates over the frequency flat channel $\boldsymbol{H}^{N_{\mathrm{Rx}}\times N_{\mathrm{Tx}}}$. At the receiver side, zero-mean additive white Gaussian noise $\boldsymbol{\eta} \in \mathbb{C}^{N_{\mathrm{Rx}}}$ is superimposed and the receive filter $\boldsymbol{B}^{K\times N_{\mathrm{Rx}}}$ generates $\hat{\boldsymbol{d}} \in \mathbb{C}^K$ representing an estimate for the virtual signal $\boldsymbol{d} \in \mathbb{A}^K + \tau\mathbb{Z}^K + \mathrm{j}\,\tau\mathbb{Z}^K \subset \mathbb{C}^K$. In the end, the modulo operator generates the estimated symbol vector $\hat{\boldsymbol{s}}$ by performing the remapping onto $\mathbb{V}^K$.

Fig. 2 shows a modulo operator free representation of the transmitter with the additive signal $\boldsymbol{a}(\boldsymbol{s}, \boldsymbol{F})$ from a K-dimensional infinite lattice $\tau\mathbb{Z}^K + \mathrm{j}\,\tau\mathbb{Z}^K$ such that $\boldsymbol{v} \in \mathbb{V}^K$ remains the same. As the modulo operators avoid simple relations between $\boldsymbol{s}$ and $\hat{\boldsymbol{s}}$, all our optimizations are based on the virtual signals $\boldsymbol{d}$ and $\hat{\boldsymbol{d}}$.

3. CAPACITY ACHIEVING *LINEAR* FILTERING REVISITED

If the precoder is restricted to act linearly, the feedback filter $\boldsymbol{F} = \boldsymbol{0}$ is inactive and all modulo operators are removed. In the following, we shortly review the capacity achieving linear filtering from an information theoretic point of view and from a signal processing point of view and highlight key properties of the resulting transmission chain.

[1] In general, any symmetrically permuted strictly upper triangular matrix may be chosen since causality and realizability are ensured by means of the resulting matrix structure. We choose the strictly upper triangularity due to the fact that the general triangular decomposition, which we will make use of later, was introduced for upper triangular matrices first.

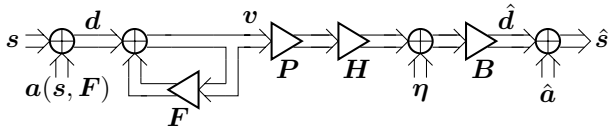

Fig. 2. Modulo-operator free precoder and receiver representation.

3.1. Information Theoretic Point of View

Capacity is defined as the maximum mutual information between the transmit signal $\boldsymbol{x} = \boldsymbol{P}\boldsymbol{s} \in \mathbb{C}^{N_{\mathrm{Tx}}}$ and the receive signal $\boldsymbol{y} = \boldsymbol{H}\boldsymbol{P}\boldsymbol{s} + \boldsymbol{\eta} \in \mathbb{C}^{N_{\mathrm{Rx}}}$, and maximization is done with respect to the input distribution of $\boldsymbol{x}$ and its covariance $\boldsymbol{C_x} = \mathrm{E}[\boldsymbol{x}\boldsymbol{x}^{\mathrm{H}}] = \boldsymbol{P}\boldsymbol{P}^{\mathrm{H}} \in \mathbb{S}_+^{N_{\mathrm{Tx}}\times N_{\mathrm{Tx}}}$ subject to an average power constraint $\mathrm{tr}(\boldsymbol{C_x}) \leq P_{\mathrm{Tx}}$:

$$\max_{\boldsymbol{C_x}} \log_2 \det\left(\boldsymbol{\mathrm{I}}_{N_{\mathrm{Tx}}} + \boldsymbol{C_x}\boldsymbol{H}^{\mathrm{H}}\boldsymbol{C_\eta}^{-1}\boldsymbol{H}\right) \quad \text{s.t.: } \mathrm{tr}(\boldsymbol{C_x}) \leq P_{\mathrm{Tx}} \text{ and } \boldsymbol{C_x} \succcurlyeq \boldsymbol{0} \tag{1}$$

First of all, Gaussian signaling has to be chosen such that the utility in (1) represents not only an upper bound, but also the truly achievable sum-rate. Telatar [7] came up with the well known result that the input covariance $\boldsymbol{C_x}$ needs to have the same eigenspace as $\boldsymbol{H}^{\mathrm{H}}\boldsymbol{C_\eta}^{-1}\boldsymbol{H}$. Moreover, the water-filling policy guarantees optimum power allocation. The optimum covariance $\check{\boldsymbol{C}}_{\boldsymbol{x}}$ directly follows from the KKT conditions assigned to (1) and can compactly be expressed as

$$\check{\boldsymbol{C}}_{\boldsymbol{x}} = \left[\mu^{-1}\boldsymbol{\mathrm{I}}_{N_{\mathrm{Tx}}} - (\boldsymbol{H}^{\mathrm{H}}\boldsymbol{C_\eta}^{-1}\boldsymbol{H})^{-1}\right]_{\perp}, \tag{2}$$

where $\mu \ln 2$ is the Lagrangian factor which is chosen such that the constraint $\mathrm{tr}(\check{\boldsymbol{C}}_{\boldsymbol{x}}) \leq P_{\mathrm{Tx}}$ is fulfilled with equality. The operator $[\cdot]_{\perp}$ performs the orthogonal projection of its Hermitian argument onto the proper cone of positive semidefinite matrices by setting all negative eigenvalues to zero [8]. It can be thought of as the multi-dimensional extension of the $\max(0, \cdot)$ operator known from scalar water-pouring. However, this compact notation already tells us which eigenspace needs to be chosen and the optimum power allocation follows as well, cf. [9].

3.2. Signal Processing Point of View

Solution (2) is important from an information theoretic point of view. However, from a signal processing perspective, it is more relevant to know how to realize the precoder $\boldsymbol{P}$ instead of its covariance $\boldsymbol{C_x} = \boldsymbol{P}\boldsymbol{P}^{\mathrm{H}}$ and how to design the receiver. In conjunction with an MMSE-type *receiver* $\boldsymbol{B}$ recovering $\boldsymbol{s}$, the utility in (1) can also be expressed as $-\log_2 \det \boldsymbol{C_e}$ (see [10]) with

$$\boldsymbol{C_e} = \left(\boldsymbol{\mathrm{I}}_K + \boldsymbol{P}^{\mathrm{H}}\boldsymbol{H}^{\mathrm{H}}\boldsymbol{C_\eta}^{-1}\boldsymbol{H}\boldsymbol{P}\right)^{-1}$$

denoting the covariance matrix of the error vector $\boldsymbol{e} = \boldsymbol{s} - \hat{\boldsymbol{s}}$.

In order to find the optimum transmit filter, we introduce the *sorted* reduced eigenvalue decomposition $\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\mathrm{H}}$ of $\boldsymbol{H}^{\mathrm{H}}\boldsymbol{C}_{\boldsymbol{\eta}}^{-1}\boldsymbol{H}$ with $\boldsymbol{U} \in \mathbb{C}^{N_{\mathrm{Tx}} \times r}$ containing the first r unit-norm eigenvectors and $\boldsymbol{\Lambda} \in \mathbb{D}^{r \times r}$ containing the r non-zero eigenvalues $\lambda_1, \ldots, \lambda_r$ in non-increasing order. Setting $\check{\boldsymbol{P}} = \boldsymbol{U}\boldsymbol{\Phi}\boldsymbol{S}$, where $\boldsymbol{\Phi} \in \mathbb{D}^{r \times K}$ is diagonal but not necessarily square with entries $\Phi_1, \ldots, \Phi_{\min(r,K)} \in \mathbb{R}$, the water-filling policy leads to a power allocation according to $\Phi_k^2 = \max(0, \frac{1}{\mu} - \frac{1}{\lambda_k})$, $k \in \{1, \ldots, \min(K, r)\}$. In order to let the covariance matrix $\boldsymbol{C}_{\boldsymbol{x}} = \boldsymbol{P}\boldsymbol{P}^{\mathrm{H}}$ achieve the rank of the optimum covariance matrix $\check{\boldsymbol{C}}_{\boldsymbol{x}}$ from (2), the number of streams K has to be chosen at least as large as the rank of the optimum covariance matrix $d = \mathrm{rank}(\check{\boldsymbol{C}}_{\boldsymbol{x}}) \leq N_{\mathrm{Tx}}$. Otherwise, capacity cannot be achieved. Finally, a unitary matrix $\boldsymbol{S} \in \mathbb{U}^{K \times K}$ can be chosen as a degree of freedom as it does not change the obtained sum-rate.

3.3. Key Properties of the Transmission Chain

Choosing $\boldsymbol{U}$ and $\boldsymbol{\Phi}$ as in the previous section, capacity is achieved. The resulting impacts on the transmission chain are now summarized:

1.) For $\boldsymbol{S} = \mathbf{I}_K$, the number of streams K has to be chosen equal to the dimension d of the subspace where transmission takes place. If $K < d$, capacity cannot be achieved, if $K > d$, $K - d$ streams get zero power allocated. So for $\boldsymbol{S} = \mathbf{I}_K$ and $K = d$, the error covariance matrix $\boldsymbol{C}_{\boldsymbol{e}} = (\mathbf{I}_K + \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{\Phi})^{-1}$ gets diagonalized and all K streams can be decoded *separately*, which has an enormous practical relevancy and drastically reduces complexity. However, the K streams have different individual MSEs and hence different signal-to-noise ratios in general. The MSE of stream k reads as $\varepsilon_k = \frac{\mu}{\lambda_k}$ and obviously depends on the associated eigenvalue λ_k. As a consequence, both transmitter and receiver must be capable of handling different coding schemes and different coding rates in order to achieve capacity. Furthermore, different modulation schemes must be available if the theoretical limit which is based on Gaussian signaling shall be approached by a practical QAM modulation scheme. Whereas one stream may require QPSK, another stream might demand for 64QAM in order to come close to the Gaussian limit.

2.) Identical stream MSEs and SINRs, i.e., equal diagonal elements of the error covariance matrix $\boldsymbol{C}_{\boldsymbol{e}}$, can be achieved by choosing $\boldsymbol{S}$ as a DFT or Hadamard matrix [11]. For this case, $K > d$ would also make sense, since power for the additional $K - d$ streams is allocated by the unitary matrix $\boldsymbol{S}$. Despite the fact that capacity is still obtained, the price one has to pay is that the error covariance matrix is no longer diagonal. Since all streams are coupled then, they have to be decoded jointly, leading to an intractable complexity.

3.) The number of transmitted streams K is therefore limited by the dimension $d = \mathrm{rank}(\check{\boldsymbol{C}}_{\boldsymbol{x}})$ and reaches the rank of the channel $r = \mathrm{rank}(\boldsymbol{H}) \leq \min(N_{\mathrm{Tx}}, N_{\mathrm{Rx}})$ in the high SNR region. No more streams can be spanned for the sake of a reduced cardinality of the modulation alphabet if separate stream decoding is preferred.

4. OPTIMUM THP TRANSCEIVER DESIGN

4.1. Precoder Structure for Near-Capacity Transmission

For THP, the modulo operator limits the real and imaginary part of each entry in $\boldsymbol{v} \in \mathbb{V}^K$ and hence prevents Gaussian signaling. The induced *shaping loss* [3] leads to the fact that the channel capacity can only be obtained up to this 1.53 dB shaping loss [3]. Nonetheless, we focus on the maximization of the upper bound. Inserting the MMSE receiver, depending on $\boldsymbol{F}$ and $\boldsymbol{P}$, into the covariance matrix $\boldsymbol{C}_{\boldsymbol{e}}$ of the error signal $\boldsymbol{e} = \boldsymbol{d} - \hat{\boldsymbol{d}}$ yields

$$\boldsymbol{C}_{\boldsymbol{e}} = (\mathbf{I} - \boldsymbol{F})(\boldsymbol{P}^{\mathrm{H}}\boldsymbol{H}^{\mathrm{H}}\boldsymbol{C}_{\eta}^{-1}\boldsymbol{H}\boldsymbol{P} + \boldsymbol{C}_{\boldsymbol{v}}^{-1})^{-1}(\mathbf{I} - \boldsymbol{F})^{\mathrm{H}}. \quad (3)$$

Again, the obtained rate can be expressed as a function of the determinant of $\boldsymbol{C}_{\boldsymbol{e}}$, see [10]. However, as mentioned in Section 2, the covariance matrix $\boldsymbol{C}_{\boldsymbol{v}}$ of $\boldsymbol{v}$ is no longer an identity matrix and the obtained rate, neglecting the shaping loss, now reads as $\log_2 \det(\boldsymbol{C}_{\boldsymbol{v}}) - \log_2 \det(\boldsymbol{C}_{\boldsymbol{e}})$. Due to the fact that $\boldsymbol{F}$ is strictly upper triangular, the determinant of $\boldsymbol{C}_{\boldsymbol{e}}$ does not depend on $\boldsymbol{F}$, i.e., $\det(\mathbf{I}_K - \boldsymbol{F}) = 1$. This means that, no matter how $\boldsymbol{F}$ is chosen, capacity up to the shaping loss can always be reached by properly choosing $\boldsymbol{P}$. The optimum precoder $\check{\boldsymbol{P}}$ follows from a determinant maximization problem [12] with covariance matrix $\boldsymbol{C}_{\boldsymbol{v}}$ and reads as

$$\check{\boldsymbol{P}} = \boldsymbol{U}\boldsymbol{\Phi}\boldsymbol{S}\boldsymbol{C}_{\boldsymbol{v}}^{-\frac{1}{2}}, \quad (4)$$

where $\boldsymbol{U}$ and $\boldsymbol{\Phi}$ are defined as in Section 3.2 and $\boldsymbol{S} \in \mathbb{U}^{K \times K}$ is again an arbitrary unitary matrix. Summing up, capacity-achieving precoding, except from the non-Gaussianity, does not uniquely determine the feedforward filter $\boldsymbol{P}$, a unitary matrix $\boldsymbol{S}$ remains as degree of freedom. Moreover, the feedback filter $\boldsymbol{F}$ does not have any influence on the throughput.

4.2. MSE Minimization and Uniform Stream Properties

Defining the *arithmetic* MSE ε_A as the sum MSE via

$$\varepsilon_A = \mathrm{E}[\|\boldsymbol{d} - \hat{\boldsymbol{d}}\|_2^2] = \mathrm{tr}\left(\mathrm{E}[\boldsymbol{e}\boldsymbol{e}^{\mathrm{H}}]\right) = \mathrm{tr}(\boldsymbol{C}_{\boldsymbol{e}}),$$

it becomes evident that ε_A corresponds to the trace of the error covariance matrix $\boldsymbol{C}_{\boldsymbol{e}}$. A lower bound on the arithmetic MSE can be derived from the trace-determinant inequality

$$\varepsilon_A = \mathrm{tr}(\boldsymbol{C}_{\boldsymbol{e}}) \geq K \sqrt[K]{\det(\boldsymbol{C}_{\boldsymbol{e}})} = K \sqrt[K]{\det(\varepsilon_G)} \quad (5)$$

relating the arithmetic *mean* to the geometric *mean*. Here, $\varepsilon_G = \det(\boldsymbol{C}_{\boldsymbol{e}})$ denotes the *geometric* MSE. The lower bound is obtained if, and only if, $\boldsymbol{C}_{\boldsymbol{e}}$ is a scaled identity matrix. The

authors in [1,2] were the first to introduce this framework. Instead of directly minimizing the sum-MSE in a decision feedback system at the receiver, they minimized the lower bound, i.e., the geometric MSE, and showed that the minimum lower bound can also be achieved by means of the *geometric mean decomposition* (GMD) [5]. The GMD is a matrix factorization where an arbitrary matrix is decomposed into the product of a unitary matrix, an upper triangular matrix with *identical* main diagonal entries, and a second unitary matrix. But since we employ THP at the transmitter and not decision feedback at the receiver, the signal covariance matrix $\boldsymbol{C_v}$ is not a scaled identity matrix. This follows from the fact that the power loss does not affect that last stream which is not influenced by the modulo operator. As we will see later, we have to decompose a matrix in a similar way except that the main diagonal entries of the upper triangular matrix have (different) prescribed values now. As a consequence, the GMD cannot be applied and the more powerful *generalized triangular decomposition* (GTD) [4] must be utilized. Its existence and properties as well as a very fast implementation are discussed in Section 5.

As $\breve{\boldsymbol{P}}$ from (4) achieves capacity up to the shaping loss, it maximizes $-\log_2 \det \boldsymbol{C_e}$ and thus minimizes the geometric mean $\varepsilon_G = \det(\boldsymbol{C_e})$. In order to let the arithmetic mean $\varepsilon_A = \mathrm{tr}(\boldsymbol{C_e})$ merge with this minimum lower bound, the error covariance matrix $\boldsymbol{C_e}$ has to boil down to a scaled identity matrix. Plugging the optimum precoder $\breve{\boldsymbol{P}}$ into the covariance expression (3), we get

$$\boldsymbol{C_e} = (\mathbf{I}_K - \boldsymbol{F})\boldsymbol{C_v}^{\frac{1}{2}}\boldsymbol{S}^{\mathrm{H}}\boldsymbol{D}^{-2}\boldsymbol{S}\boldsymbol{C_v}^{\frac{1}{2}}(\mathbf{I}_K - \boldsymbol{F})^{\mathrm{H}}, \quad (6)$$

where $\boldsymbol{D} = (\mathbf{I}_K + \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{\Phi})^{\frac{1}{2}}$ is diagonal and positive definite. If a unitary $\boldsymbol{S}$ and and a strictly upper triangular $\boldsymbol{F} \in \mathbb{T}^{K\times K}$ exist such that $\boldsymbol{C_e} = \sigma_\varepsilon^2\mathbf{I}_K = \sigma_\varepsilon^2\boldsymbol{Q}^{\mathrm{H}}\boldsymbol{Q}$ with unitary $\boldsymbol{Q} \in \mathbb{U}^{K\times K}$, i.e., $\boldsymbol{C_e}$ becomes a scaled identity, the arithmetic mean achieves its global minimum. To this end, a unitary decomposition [2]

$$\boldsymbol{D} = \boldsymbol{Q}(\mathbf{I}_K - \boldsymbol{F})\sigma_\varepsilon^{-1}\boldsymbol{C_v}^{\frac{1}{2}}\boldsymbol{S}^{\mathrm{H}} = \boldsymbol{Q}\boldsymbol{R}\boldsymbol{S}^{\mathrm{H}} \quad (7)$$

should exist with unitary $\boldsymbol{S}$ and $\boldsymbol{Q}$ and an upper triangular matrix

$$\boldsymbol{R} = \sigma_\varepsilon^{-1}(\mathbf{I}_K - \boldsymbol{F})\boldsymbol{C_v}^{\frac{1}{2}}, \quad (8)$$

the diagonal of which corresponds to the diagonal of $\boldsymbol{C_v}^{\frac{1}{2}}$ divided by σ_ε and therefore has different entries on its main diagonal. That's why the GMD cannot be used and the GTD has to be employed. In Section 5, exact conditions for its existence are presented. From simulations, we find that these requirements are only violated for very low SNR values below -10 dB. More precisely, the common MSE level σ_ε^2 cannot grow above one, see Section 5. As $\boldsymbol{R}$ follows from two unitary transformations applied to $\boldsymbol{D}$ (cf. Eq. 7), and since both $\boldsymbol{R}$ and $\boldsymbol{D}$ have positive real-valued main diagonal entries, $\boldsymbol{Q}$ and $\boldsymbol{S}$ can be chosen real-valued as well. Additionally, $\boldsymbol{R}$ and $\boldsymbol{D}$ have the same determinant from which we can compute the common MSE level

$$\sigma_\varepsilon^2 = \sqrt[K]{\prod_{k=1}^{K}[\boldsymbol{C_v}]_{k,k}\prod_{k=1}^{K}[\boldsymbol{D}]_{k,k}^{-2}} = \sqrt[K]{\sigma_v^{2(K-1)}\prod_{k=1}^{K}[\boldsymbol{D}]_{k,k}^{-2}}, \quad (9)$$

where we exploited the fact that the first $K-1$ diagonal entries of $\boldsymbol{C_v}$ are σ_v^2 and the last entry equals one. Given the unfavorable conditions at very low SNR that the GTD does not exist, the minimum MSE cannot be achieved in combination with a balancing of all individual MSEs. As soon as σ_ε^2 is one, i.e., the sum-MSE reaches K, a further reduction of the transmit power prevents the GTD to exist, and consequently, all streams cannot be balanced any more. Interestingly, it is possible to switch off the last stream such that its MSE is one, and balance the remaining streams $1,\ldots,K-1$ at an MSE level below one. Doing so, the $K-1$st stream is precoded linearly, and therefore, the $K-1$st diagonal entry in $\boldsymbol{C_v}$ reduces from σ_v^2 to one, as the feedback for the linearly precoded stream $K-1$ is disabled. This change in the variance leads to the fact that the MSE level of the remaining streams changes, cf. (9). If this level has dropped below $[\boldsymbol{D}]_{K,K}^{-2}$, the remaining $K-1$ streams can be balanced at this level, otherwise, the $K-1$st stream will also be switched off, changing the variance of the $K-2$nd stream to one, and so on. If only a single stream is active in the end, the covariance matrix $\boldsymbol{C_v}$ has been transformed to the identity matrix, and the modulo operators become obsolete. From this procedure it becomes obvious that the sum-MSE can never grow above the number of streams K no matter how small the transmit power P_{Tx} gets.

4.3. Key Properties of the Nonlinear Transmission

The following statements hold under the assumption that the GTD exists.

1.) Except for the shaping loss due to the non-Gaussianity of $\boldsymbol{v}$, the precoder achieves capacity. At high SNRs, the modulo operator at the receiver side does not significantly change the Gaussianity of the receive noise. These propositions remain valid even if the GTD does not exist.

2.) The error covariance matrix $\boldsymbol{C_e}$ reduces to a scaled identity matrix, meaning that all streams have the same properties, especially the same MSEs and the same rates. Hence, no bit-loading is necessary which is the major advantage compared to capacity achieving linear filtering. Moreover, each stream can be decoded separately which has a drastically lower complexity than a joint decoding of all streams simultaneously.

3.) The number K of active streams must be larger than or equal to the dimension $d = \mathrm{rank}(\breve{\boldsymbol{C}}_{\boldsymbol{x}})$ of the optimum covariance matrix from (2). Capacity is achieved for an arbitrary high number of streams. Larger stream numbers facili-

[2] Remember that $\boldsymbol{D} = (\mathbf{I}_K + \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{\Phi})^{\frac{1}{2}}$ is diagonal.

tate the use of small cardinality modulation alphabets making this approach very important with respect to the practical implementation. Clearly, the MSE per stream is increased when increasing the number of streams.

4) In contrast to decision-feedback equalization (DFE) based nonlinear filtering [1,2], THP does not suffer from error propagation. Thus, considerable gains can be achieved in the mid-SNR region despite the modulo loss implicating the generation of new nearest neighbors and despite the power loss ($\boldsymbol{C_v} \neq \mathbf{I}_K$) since the crucial assumption of perfect decision in DFE is not fulfilled in this SNR region, see the simulation results in Section 6.

5. THE GENERALIZED TRIANGULAR DECOMPOSITION

5.1. Existence and Uniqueness

Taking a closer look at (7), we find from $\boldsymbol{R} = \boldsymbol{Q}^{\mathrm{H}}\boldsymbol{D}\boldsymbol{S}$ that the diagonal matrix $\boldsymbol{D}$ comprises the singular values of the matrix $\boldsymbol{R}$, which itself has the eigenvalues on its diagonal, since it is upper triangular. Weyl [13] proved that the singular values multiplicatively majorize[3] the eigenvalues of a matrix, and Horn [14] stated that an upper triangular matrix $\boldsymbol{R}$ with prescribed diagonal and a specific singular value set exists, if the diagonal is majorized by the singular values, see [5]. In our context, this means that

$$\begin{aligned} \prod_{k=1}^{\ell}[\boldsymbol{R}]_{k,k}^2 &\leq \prod_{k=1}^{\ell}[\boldsymbol{D}]_{k,k}^2 \forall \ell : 1 \leq \ell < K \\ \prod_{k=1}^{K}[\boldsymbol{R}]_{k,k}^2 &= \prod_{k=1}^{K}[\boldsymbol{D}]_{k,k}^2, \end{aligned} \tag{10}$$

because the diagonal entries of both $\boldsymbol{R}$ and $\boldsymbol{D}$ are already sorted in an non-increasing fashion. For $\boldsymbol{D}$, this follows from the water-pouring policy and the fact that we made use of the *sorted* eigenvalue decomposition. The second row in (10) is always fulfilled due to the choice of σ_ε^2 in (9). Since the entries in $\boldsymbol{D}$ are non-increasing and the first $K-1$ entries in $\boldsymbol{C_v}$ are identical, we may replace (10) in conjunction with (8) by the necessary and sufficient condition

$$\frac{1}{\sigma_\varepsilon^2} \geq [\boldsymbol{D}]_{K,K}^2 \tag{11}$$

for the existence of the GTD. Note that $[\boldsymbol{D}]_{K,K}^{-2} \leq 1$ and hence $\sigma_\varepsilon^2 \leq 1$ must hold. In the following, we show that the transmit power P_{Tx} can always be chosen so small that the GTD does not exist. To this end, we assume that the transmit power is already so small that $d = 1$ holds, i.e., $\Phi_1^2 = P_{\mathrm{Tx}}$ and $\Phi_k = 0\ \forall k \neq 1$. Moreover, $K = 2$ streams shall be transmitted. Then $[\boldsymbol{D}]_{1,1}^2 = 1 + \lambda_1 P_{\mathrm{Tx}}$, $[\boldsymbol{D}]_{2,2}^2 = 1$, and $\sigma_\varepsilon^2 = \sqrt{\frac{\sigma_v^2}{1+\lambda_1 P_{\mathrm{Tx}}}}$ hold. Eq. 10 now requires that $\frac{\sigma_v^2}{\sigma_\varepsilon^2} \leq [\boldsymbol{D}]_{1,1}^2$, which is violated for $P_{\mathrm{Tx}} < \lambda_1^{-1}(\sigma_v^2 - 1)$.

[3] In this context, majorization is related to an non-increasing ordering.

Similar to the GMD [5], the GTD is not unique if it exists. In particular, infinitely many decompositions exist but the Frobenius norm of all those matrices $\boldsymbol{R}$ is identical to $\|\boldsymbol{D}\|_{\mathrm{F}}$. One type of invariance are unitary diagonal matrices: Let $\boldsymbol{D} = \boldsymbol{Q}\boldsymbol{R}\boldsymbol{S}^{\mathrm{H}}$ be a GTD decomposition of the diagonal matrix $\boldsymbol{D}$, and let $\boldsymbol{V}$ and $\boldsymbol{W}$ be unitary diagonal matrices. With $\bar{\boldsymbol{R}} = \boldsymbol{V}\boldsymbol{R}\boldsymbol{V}^{\mathrm{H}}$, $\bar{\boldsymbol{Q}} = \boldsymbol{W}\boldsymbol{Q}\boldsymbol{V}^{\mathrm{H}}$, and $\bar{\boldsymbol{S}} = \boldsymbol{W}\boldsymbol{S}\boldsymbol{V}^{\mathrm{H}}$, we find that $\boldsymbol{D} = \boldsymbol{W}\boldsymbol{D}\boldsymbol{W}^{\mathrm{H}} = \boldsymbol{W}\boldsymbol{Q}\boldsymbol{R}\boldsymbol{S}^{\mathrm{H}}\boldsymbol{W}^{\mathrm{H}} = \bar{\boldsymbol{Q}}\bar{\boldsymbol{R}}\bar{\boldsymbol{S}}^{\mathrm{H}}$ is also a valid GTD decomposition of $\boldsymbol{D}$. Furthermore, a completely different off-diagonal structure in $\boldsymbol{R}$ can be obtained for $K > 2$ by changing the order selection of the Givens rotations, see the following section.

5.2. Efficient and Stable Computation of the GTD

Independently from but similar to [2,4], we found a low complexity and extremely stable way to compute the GTD when trying to recude the complexity of the GMD from quartic order in [15] to quadratic order. This goal was achieved by the use of Givens rotations as in [2]. We interpret the unitary matrices $\boldsymbol{S}$ and $\boldsymbol{Q}$ as the product of $K-1$ real-valued Givens rotations $\boldsymbol{Q}_i$ and $\boldsymbol{S}_i$, respectively:

$$\boldsymbol{S} = \prod_{i=1}^{K-1}\boldsymbol{S}_i, \text{ and } \boldsymbol{Q} = \prod_{i=1}^{K-1}\boldsymbol{Q}_i.$$

A (real-valued) Givens rotation $\boldsymbol{G}_{j,k}$ is a rank-two perturbation of the K-dimensional identity matrix with $[\boldsymbol{G}_{j,k}]_{j,j} = [\boldsymbol{G}_{j,k}]_{k,k} = a$, $[\boldsymbol{G}_{j,k}]_{j,k} = b$, and $[\boldsymbol{G}_{j,k}]_{k,j} = -b$ holds. Imposing the constraints $a, b \in \mathbb{R}$, $-1 \leq a \leq 1$, $-1 \leq b \leq 1$, and $a^2 + b^2 = 1$ leads to the fact that $\boldsymbol{G}_{j,k}$ is orthogonal. We assume that $j < k$. Note that $\boldsymbol{G}_{j,k}$ performs a counterclockwise rotation in the j,k plane with a and b corresponding to the cosine and the sine of the rotation angle [16].

Multiplying $\boldsymbol{G}_{j,k}$ from the right-hand side (RHS) onto a matrix has influence only on the columns j and k whereas a left-hand side (LHS) multiplication has impact on rows j and k only. Hence, we can focus on rows j and k and columns j and k separately. Extracting the four entries of the LHS and RHS multiplication of two different Givens rotations applied to the matrix $\boldsymbol{D}$ yields

$$\begin{aligned} \begin{bmatrix} a & b \\ -b & a \end{bmatrix} \begin{bmatrix} z_j & 0 \\ 0 & z_k \end{bmatrix} \begin{bmatrix} c & d \\ -d & c \end{bmatrix} = \\ \begin{bmatrix} acz_j - bdz_k & adz_j + bcz_k \\ -(bcz_j + adz_k) & -bdz_j + acz_k \end{bmatrix}. \end{aligned} \tag{12}$$

z_j and z_k are the jth and the kth diagonal entries of the matrix $\boldsymbol{Z}$ which is initialized with the matrix $\boldsymbol{D}$ and to which the Givens rotations are applied such that after $N-1$ rotations, $\boldsymbol{Z} = \boldsymbol{R}$ holds. Since the upper triangular structure of $\boldsymbol{Z}$ must be preserved, the lower left entry in (12) has to be zero. W.l.o.g., we set the upper left[4] element to the desired

[4] It is also possible to set the lower right element to the desired value. However, the order selection changes then.

value $\beta_j = \sigma_\varepsilon^{-1}[\boldsymbol{C}_{\boldsymbol{v}}^{\frac{1}{2}}]_{j,j}$, see (8). The set of equations

$$acz_j - bdz_k = \beta_j, \quad bcz_j + adz_k = 0$$
$$a^2 + b^2 = 1, \qquad c^2 + d^2 = 1$$

leads to the solutions

$$c = \sqrt{\frac{\beta_j^2 - z_k^2}{z_j^2 - z_k^2}}, \quad d = -\sqrt{\frac{z_j^2 - \beta_j^2}{z_j^2 - z_k^2}}, \quad a = \frac{z_j c}{\beta_j}, \quad b = -\frac{z_k d}{\beta_j}.$$

The verification of the constraints $0 \leq 1 - b^2 \leq 1$ and $0 \leq d^2 \leq 1$ tells us how to choose the indices j and k:

$$z_j \leq \beta_j \leq z_k \text{ or } z_j \geq \beta_j \geq z_k. \tag{13}$$

Apparently, we need to find k such that the desired value β_j is enclosed by the two diagonal elements z_k and z_j. Starting with $j = 1$, we search for an index $k > j$ for which (13) is fulfilled. If the majorization criterion (10) is fulfilled, such a k always exists. After the application of the Givens rotations to $\boldsymbol{Z}$, j is increased by one and a $k > j$ satisfying (13) has to be found again. This procedure terminates after $K - 1$ steps and during the last step, $j = K - 1$ and $k = K$ hold since the determinants of $\boldsymbol{D}$ and $\boldsymbol{R}$ are the same. The overall complexity to compute both $\boldsymbol{Q}$, $\boldsymbol{S}$, and $\boldsymbol{R}$ is only $\mathcal{O}(K^2)$.

6. SIMULATION RESULTS

In our simulation results, we compare the THP GTD-based joint transmitter and receiver design (cross marker) presented in this paper, with the joint transmitter and receiver GMD-based decision feedback equalization (circle marker) from [1,2]. Both versions guarantee uniform stream properties (the GTD requires the common MSE level to be below one for stream balancing). For purposes of comparison, we also simulated the nonlinear MMSE THP (square marker) from [17] and the linear MMSE filter (dashed curve) from [18] where the receive filters are forced to be scaled identity matrices. Clearly, these filters only exist for $K = N_{\text{Rx}}$ but they do not balance the individual streams' MSEs. Nonetheless, we plot stream-averaged uncoded bit-error ratios (BERs).

Fig. 3 shows a QPSK transmission of $K = 4$ streams in a system with $N_{\text{Tx}} = 4$ transmit and $N_{\text{Rx}} = 4$ receive antennas averaged over uncorrelated channels. Linear MMSE precoding (dashed curve) [18] with a non-cooperative receiver is clearly inferior to nonlinear THP-based precoding or nonlinear DFE-based equalization. Both the GMD-based and the GTD-based schemes clearly outperform the MMSE THP filter [17] with a non-cooperative receiver being a scaled identity matrix. There are two reasons why the DFE-based system performs better than the GTD-based one. First, the generation of next neighbors due to the modulo operator is quite crucial for QPSK, since every symbol now has four nearest neighbors instead of two. Second, the variance σ_v^2 of the first three

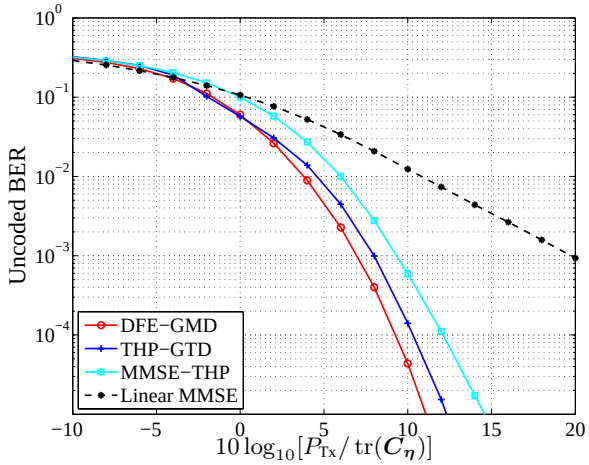

Fig. 3. QPSK transmission of $K = 4$ streams over a system with $N_{\text{Tx}} = 4$ transmit and $N_{\text{Rx}} = 4$ receive antennas (uncorrelated channel).

streams raises from one to $\sigma_v^2 = \frac{M}{M-1} = \frac{4}{3}$ for QPSK bringing an increased MSE level σ_ε^2 with it compared to the case when $\sigma_v^2 = 1$ would hold, see the definition of σ_ε^2 in Section 4.2.

This changes when we switch the modulation alphabet to 16-QAM, see Fig. 4. All other parameters are left unchanged. First, we notice that all filters perform slightly worse. In the previous figure, the transmit signal-to-noise ratio $P_{\text{Tx}}/\operatorname{tr}(\boldsymbol{C}_\eta)$ was to small to see that the MMSE THP with non-cooperative receiver (square marker) from [17] flattens due to the fact that it does not achieve the full diversity order. In Fig. 4, this can already be anticipated, and in fact, the square marker curve achieves the diversity order $1 + \min(N_{\text{Tx}}, N_{\text{Rx}}) - K = 1$, since one stream is precoded linearly. By contrast, both the GMD (circle marker) and therefore also the GTD based (cross marker) version have the full diversity order, see [2]. Another interesting observation is the intersection between the DFE-GMD curve (circle marker) and the MMSE THP curve (square marker). It results from the error propagation of the DFE in the small and moderate SNR range. The assumption of perfect interference subtraction is clearly violated in this region. Furthermore, the generation of next neighbors due to the modulo operator is not so severe for the THP-GTD (cross marker) in case of 16-QAM, since only the outer symbol get new nearest neighbors. In addition, the power loss induced by $\sigma_v^2 > 1$ is very small for 16-QAM, since $\sigma_v^2 = \frac{M}{M-1} = \frac{16}{15}$. Both effects are not as detrimental as the error propagation in case of decision feedback equalization. For an uncoded BER of 10^{-1}, a gain of 2 dB can be achieved if THP is applied instead of the DFE.

Even larger gains can be obtained when the channel becomes correlated as in Fig. 5. Here, $K = 8$ streams are transmitted from a $N_{\text{Tx}} = 8$ antenna sender to a $N_{\text{Rx}} = 8$ antenna receiver. Due to the correlated scenario, error propa-

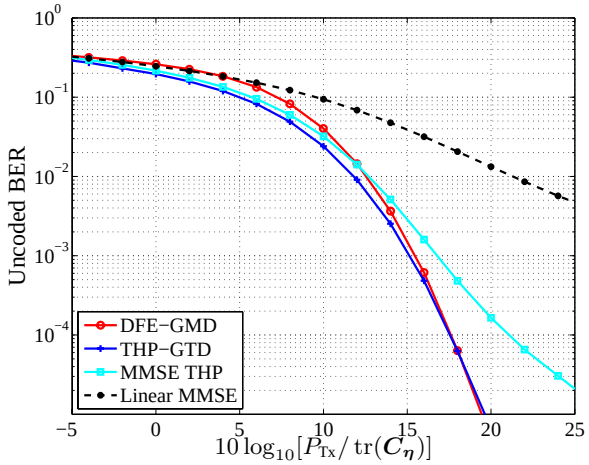

Fig. 4. 16-QAM transmission of $K = 4$ streams over a system with $N_{\mathrm{Tx}} = 4$ transmit and $N_{\mathrm{Rx}} = 4$ receive antennas (uncorrelated channel).

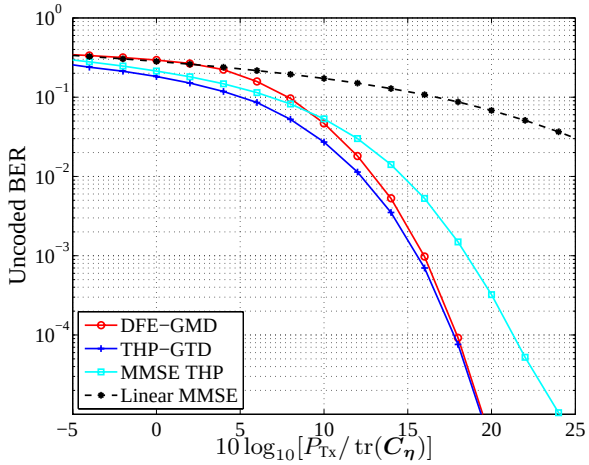

Fig. 5. 16-QAM transmission of $K = 8$ streams over a system with $N_{\mathrm{Tx}} = 8$ transmit and $N_{\mathrm{Rx}} = 8$ receive antennas (correlated channel).

gation turns out to be even more severe. For an uncoded BER of 10^{-1}, 3 dB can be gained by the THP-GTD compared to the DFE-GMD, and if 10^{-2} is the desired operation point, we can still save 1 dB.

7. REFERENCES

[1] F. Xu, T. N. Davidson, J.-K. Zhang, and K. M. Wong, "Design of Block Transceivers with Decision Feedback Detection," *IEEE Transactions on Signal Processing*, vol. 54, no. 3, pp. 965–978, March 2006.

[2] Y. Jiang, J. Li, and W. W. Hager, "Uniform Channel Decomposition for MIMO Communications," *IEEE Transactions on Signal Processing*, vol. 53, no. 11, November 2005.

[3] Wei Yu, David P. Varodayan, and John M. Cioffi, "Trellis and convolutional precoding for transmitter-based interference presubtraction.," *IEEE Transactions on Communications*, vol. 53, no. 7, pp. 1220–1230, 2005.

[4] Y. Jiang, W. W. Hager, and J. Li, "The Generalized Triangular Decomposition," Accepted for publication in Mathematics of Computation, 2005.

[5] Y. Jiang, W. W. Hager, and J. Li, "The Geometric Mean Decomposition," *Linear Algebra and its Application*, vol. 396, pp. 373–384, 2005.

[6] Robert F. Fischer, *Precoding and Signal Shaping for Digital Transmission*, John Wiley & Sons, Inc., New York, NY, USA, 2002.

[7] Emre Telatar, "Capacity of multi-antenna gaussian channels," *European Transactions on Telecommunications*, vol. 10, no. 6, pp. 585–596, November/December 1999.

[8] N. J. Higham, "Matrix Nearness Problems and Applications," in *Applications of Matrix Theory*, M. J. C. Gover and S. Barnett, Eds. 1989, pp. 1–27, Oxford University Press.

[9] R. Hunger, D. A. Schmidt, and W. Utschick, "Sum Capacity and MMSE for the MIMO Broadcast Channel without Eigenvalue Decompositions," *Submitted to ISIT 2007*.

[10] P. Stoica, Y. Jiang, and J. Li, "On MIMO Channel Capacity: An Intuitive Discussion," *IEEE Signal Processing Mag.*, pp. 83–84, May 2005.

[11] D. Palomar, J. Cioffi, and M. Lagunas, "Joint Tx-Rx beamforming Design for Multicarrier MIMO Channels: A Unified Framework for Convex Optimization," *IEEE Transactions on Signal Processing*, vol. 51, no. 9, pp. 2381–2401, 2003.

[12] H. S. Witsenhausen, "A Determinant Maximization Problem Occurring in The Theory of Data Communication," *Siam Journal on Applied Mathematics*, vol. 29, no. 3, pp. 515–522, November 1975.

[13] H. Weyl, "Inequalities between Two Kinds of Eigenvalues of a Linear Transformation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 35, pp. 408–411, 1949.

[14] A. Horn, "On the Eigenvalues of a Matrix with Prescribed Singular Values," in *Proceedings of the American Mathematical Society*, February 1954, vol. 5, pp. 4–7.

[15] Jian Kang Zhang, Aleksandar Kavčić, and Kon Max Wong, "Equal-Diagonal QR Decomposition and its Application to Precoder Ddesign for Successive-Cancellation Detection," January 2005.

[16] Gene H. Golub and Charles F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, 1991.

[17] M. Joham, J. Brehmer, and W. Utschick, "MMSE approaches to multiuser spatio-temporal Tomlinson-Harashima precoding," in *Proc. ITG SCC'04*, January 2004, pp. 387–394.

[18] M. Joham, W. Utschick, and J. A. Nossek, "Linear Transmit Processing in MIMO Communications Systems," *IEEE Transactions on Signal Processing*, vol. 53, no. 8, pp. 2700–2712, August 2005.