Hybrid TH-VP Precoding for Multi-User MIMO

Rui Chen, Marco Moretti, *Member, IEEE,* and Xiaodong Wang, *Fellow, IEEE*

***Abstract*—Vector perturbation (VP) is a nonlinear precoding technique that achieves near-capacity performance in multi-user MIMO systems at the expense of large complexity due to the search for the optimum perturbation vector. In this paper we present the hybrid Tomlinson-Harashima VP (TH-VP) algorithm, a novel zero-forcing precoding scheme, which combines Tomlinson-Harashima (TH) precoding to remove inter-user interference, and VP precoding to equalize each user's multiple spatial streams. We show that the two nonlinear techniques can be integrated in a single optimization and that the proposed algorithm has lower computational requirements than any other. The performance of TH-VP is analyzed and simulation results show that TH-VP outperforms conventional zero-forcing VP and approaches the performance of dirty paper coding.**

***Index Terms*—Multi-user MIMO, vector perturbation, Tomlinson-Harashima precoding.**

I. Introduction

In recent years multi-user multiple-input multiple-output (MU-MIMO) systems have been the subject of intense research due to their potential for achieving high capacity and increased diversity in mobile communications. In particular, MU-MIMO techniques have found application in wireless systems seeking a spectrally efficient usage of the available radio resources. A key element of the MU-MIMO architecture is *transmit precoding*, which exploits the knowledge of channel state information (CSI) to separate the signals of the various users in the system and, in some cases, increase their received signal-to-noise ratio (SNR). It is known [1] that dirty paper coding (DPC) is the optimal capacity-achieving precoding technique for the MIMO broadcast channel. Unfortunately, DPC requires full non-causal knowledge of the transmitted signals, and hence it is difficult to implement in practice. Practical systems employ suboptimal *linear* techniques [2] such as zero forcing (ZF) and minimum-mean-squared-error (MMSE) precoding, both of which are based on inverting the MIMO channel matrix or a regularized version of it. Nevertheless, the performance gap between these linear precoding methods and DPC is large, mostly because of the transmit power enhancement due to channel inversion.

To limit the transmit power penalty of channel inversion, *nonlinear* precoding techniques, such as Tomlinson-Harashima (TH) [3] and vector perturbation (VP) [4] precoding, have been proposed. Recently, zero-forcing VP (ZF-VP) and MMSE-VP precoding have attracted the attention of many researchers [5]-[11] for their excellent performance. On the other hand, vector perturbation schemes have a large complexity: finding the perturbation vector requires the solution at the transmitter of an optimization problem over a large set of variables, whose size depends on the number of users and the number of antennas at the transmit and receive sides. To address this problem, [5], [8] and [12] address the case where mobile users have more than one antenna and VP precoding is performed on a per-user/group basis rather than over all the users globally. These approaches are based on block diagonalization (BD) of the MU-MIMO broadcast channel [13] and, although suboptimal, they have the advantage of a lower complexity and a larger flexibility compared to conventional VP. Unfortunately, use of BD sacrifices part of MIMO degrees of freedom to nullify interference and its sum rate is inferior to the capacity of DPC with a fixed gap at all SNR values.

In this paper, we present a hybrid TH-VP precoding algorithm for the MU-MIMO broadcast channel. Successive BD and TH precoding are designed to cancel the inter-user interference (IUI), while channel inversion of each user's effective channel and VP cancel the remaining inter-stream interference (ISI). The main advantage of this scheme with respect to other per-user VP algorithms proposed in the literature is that most users benefit from a large degree of spatial diversity and the achievable sum rate of TH-VP is shown to approach the sum-capacity of DPC asymptotically for high SNR. Regarding the implementation of the proposed precoder, we show that it is possible to integrate TH and VP precoding in just one single operation so that, employing the complex Lenstra-Lenstra-Lovász (CLLL) lattice-reduction-aided algorithm [14], the computational load of TH-VP is less than one-half of that of conventional ZF-VP [4] and even lower than that of low complexity BD-VP algorithm presented in [8], which is a benchmark for per-user VP schemes.

II. Background

We consider the downlink of a MU-MIMO system where K independent users share the same channel. The BS is equipped with N_t transmit antennas and each user k has $n_k \geq 1$ receive antennas so that the total number of receive antennas is $N_r = \sum_{k=1}^{K} n_k$ ($N_r \leq N_t$) and we assume that the BS transmits $L_k = n_k$ independent streams to user k. At the BS, nonlinear precoding is employed to separate the signals of the various users and optimize their performance. Precoding consists of two operations: first the n_k-dimensional signal $\mathbf{s}_k$ of user k is nonlinearly precoded into the signal $\mathbf{x}_k$ and then $\mathbf{x}_k$ is multiplied by the matrix $\mathbf{F}_k$, so that the kth user's transmitted signal is

$$\mathbf{y}_k = \mathbf{F}_k \mathbf{G}_k^{-1} \mathbf{x}_k \tag{1}$$

where $\mathbf{G}_k = g_k \mathbf{I}_{n_k}$ and g_k is an automatic gain control (AGC) scalar. In matrix notation, let $\mathbf{x} = [\mathbf{x}_1^T, \mathbf{x}_2^T, \ldots, \mathbf{x}_K^T]^T$ and $\mathbf{F} = [\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_K]$, the transmitted vector $\mathbf{y} = \sum_{k=1}^{K} \mathbf{y}_k$ is

$$\mathbf{y} = \mathbf{F}\mathbf{G}^{-1}\mathbf{x} \tag{2}$$

where $\mathbf{G} = \operatorname{diag}(\mathbf{G}_1, \mathbf{G}_2, \ldots, \mathbf{G}_K)$.

The channel from the base station to the kth user is flat fading and is modeled by the $n_k \times N_t$ channel matrix $\mathbf{H}_k$,

whose elements are zero-mean complex Gaussian variables with variance σ_k^2. We assume that the *combined channel matrix* $\mathbf{H} = [\mathbf{H}_1^T, \mathbf{H}_2^T, \cdots, \mathbf{H}_K^T]^T$, composed by the channel matrices of all users, is known at the BS by means of uplink channel estimation or limited feedback [2]-[5]. The received signal at the kth terminal is multiplied by the AGC g_k to yield

$$\hat{\mathbf{x}}_k = g_k \mathbf{H}_k \mathbf{y}_k + \underbrace{g_k \mathbf{H}_k \sum_{i=1, i\neq k}^{K} \mathbf{y}_i}_{\text{IUI}} + g_k \mathbf{z}_k, \tag{3}$$

where $\mathbf{z}_k$ is the additive complex Gaussian noise vector with zero mean and covariance matrix $\sigma_z^2 \mathbf{I}_{n_k}$.

In MU-MIMO systems IUI and ISI are among the major problems the system designer has to deal with. ZF precoding, which employs the pseudo-inverse of the channel matrix as linear filter, i.e., $\mathbf{F} = \mathbf{H}^\dagger$, is particularly suited to address this kind of problems, with the disadvantage that it tends to cause an increment of the transmit power.

Vector perturbation precoding is a nonlinear technique designed to moderate the ZF transmit power enhancement. In detail, the perturbed signal vector is $\mathbf{x} = \mathbf{s} + \tau \boldsymbol{l}^{\text{VP}}$, where the *perturbation vector* $\boldsymbol{l}^{\text{VP}}$ is added to the modulation signal $\mathbf{s}$ with the objective of minimizing $\mathcal{E} = \|\mathbf{F}\mathbf{x}\|^2$, i.e. $\boldsymbol{l}^{\text{VP}}$ is obtained by solving

$$\boldsymbol{l}^{\text{VP}} = \underset{\boldsymbol{l} \in \mathbb{C}\mathbb{Z}^{N_r}}{\arg\min} \left\| \mathbf{F}(\mathbf{s} + \tau \boldsymbol{l}) \right\|^2, \tag{4}$$

where $\boldsymbol{l}$ belongs to an N_r-dimensional complex integer lattice $\mathbb{C}\mathbb{Z}^{N_r}$ and τ is a positive design parameter chosen to provide a symmetric decoding region around every signal constellation point [4]. For the zero-forcing vector perturbation (ZF-VP) precoder the AGC value is $g_k = \sqrt{\mathcal{E}/P}$ for all users k, where P is the transmit signal power. Since the precoding filter $\mathbf{F}$ is designed to completely eliminate both IUI and ISI, the received signal at the kth user in (3) becomes $\hat{\mathbf{x}}_k = \mathbf{x}_k + g_k \mathbf{z}_k$ and the modulation signal vector can be recovered by applying the modulo operator on the received signal vector $\hat{\mathbf{x}}_k$

$$\hat{\mathbf{s}}_k = \text{mod}_\tau \left(\mathbf{x}_k + g_k \mathbf{z}_k \right) = \mathbf{s}_k + g_k \mathbf{z}_k, \tag{5}$$

where the last equality holds for a sufficiently high SNR. The combination of channel inversion and vector perturbation delivers excellent performance since it is able to relax input alphabet and capture the full diversity of the channel. Unfortunately, the optimization in (4) is extremely complex and becomes computationally prohibitive for even small values of N_r.

III. Hybrid Tomlinson-Harashima Vector Perturbation Precoding

In this section, we introduce TH-VP precoding, which combines successive block diagonalization with a modified vector perturbation technique. The idea is to pursue a ZF approach by cancelling IUI and ISI in two different steps: IUI is removed by successive block diagonalization and Tomlinson-Harashima (TH) precoding and ISI is canceled by channel inversion and vector perturbation on a per-user basis, so that the search in (4) is performed on a much smaller set of values.

Block diagonalization (BD) precoding [13] achieves full cancellation of the IUI by forcing each user's signal to lie in a subspace orthogonal to the channels of all other users, effectively transforming the MU-MIMO channel into a set of parallel single-user MIMO channels. Nevertheless, BD sacrifices part of the available spatial diversity so that the diversity for each user k is reduced to $N_t - \sum_{i=1, i\neq k}^{K} n_i$ from the potential value of N_t, fully achieved by ZF-VP precoding. Accordingly, BD-VP, the precoding algorithm that combines BD precoding and per-user channel inversion and vector perturbation, shows worse performance of ZF-VP.

This transmit diversity issue can be partially addressed by combining together *successive* BD with nonlinear TH precoding as in [15] and [16]. Successive BD ranks the users according a certain sequential order $u_1, u_2, \ldots, u_K$ [17] and designs the precoding matrix so that user u_k is interfered only by users $u_1, u_2, \cdots, u_{k-1}$. To simplify the notation, in the following derivations the indexes u_k will be relabelled according to the map $u_k \rightarrow k$. Let the *partial interference matrix* for user k be the $\sum_{i=1}^{k-1} n_i \times N_t$-dimensional matrix $\tilde{\mathbf{H}}_k = [\mathbf{H}_1^T, \mathbf{H}_2^T, \cdots, \mathbf{H}_{k-1}^T]^T$ that collects the MIMO channels of the first ordered $k-1$ users and whose SVD is

$$\tilde{\mathbf{H}}_k = \tilde{\mathbf{U}}_k \tilde{\mathbf{\Lambda}}_k [\tilde{\mathbf{V}}_k^{(1)} \tilde{\mathbf{V}}_k^{(0)}]^H. \tag{6}$$

By employing the precoding matrix $\tilde{\mathbf{V}}_k^{(0)}$, which contains the $(N_t - \sum_{i=1}^{k-1} n_i)$ right singular vectors associated to the null eigenvalues of $\tilde{\mathbf{H}}_k$, the precoded signal of user k is projected on a subspace orthogonal to the channels of the users with index $i < k$ and does not interfere with them. By allowing a certain level of interference, all users but the last one are projected on a subspace with a number of spatial dimensions larger than in the case of conventional BD. The effective channel of the kth user is described by the $n_k \times (N_t - \sum_{i=1}^{k-1} n_i)$ matrix $\mathbf{H}_{\text{eff},k} = \mathbf{H}_k \tilde{\mathbf{V}}_k^{(0)}$.

Following a ZF strategy, the ISI for user k is canceled by employing the pseudo-inverse $\mathbf{H}_{\text{eff},k}^\dagger$ of the effective channel matrix, so that the linear precoder for user k is

$$\mathbf{F}_k = \tilde{\mathbf{V}}_k^{(0)} \mathbf{H}_{\text{eff},k}^\dagger \tag{7}$$

and the kth received signal in (3) becomes

$$\hat{\mathbf{x}}_k = \mathbf{x}_k + g_k \mathbf{H}_k \sum_{i=1}^{k-1} \mathbf{y}_i + g_k \mathbf{z}_k. \tag{8}$$

A. Nonlinear Precoding

The remaining IUI for user k, represented by the signals of users 1 to $k-1$ in (8), is canceled by employing nonlinear TH precoding, implemented as follows

$$\mathbf{x}_k^{\text{TH}} = \text{mod}_{\tau_k} \left(\mathbf{s}_k - g_k \mathbf{H}_k \sum_{i=1}^{k-1} \mathbf{y}_i \right), \tag{9}$$

where τ_k is a positive real number related to the modulation order and constellation size of the signal of user k. In particular, we choose the value of τ_k equal to the one required by VP precoding as in [4].

We can now observe that the modulo τ operation of the n_k-dimensional vector $\mathbf{q}$ can be expressed as $\text{mod}_\tau(\mathbf{q}) = \mathbf{q} + \tau \boldsymbol{l}$,

where $\boldsymbol{l}$ belongs to the n_k-dimensional complex integer lattice $\mathbb{C}\mathbb{Z}^{n_k}$ and it is such that each element of $\mathbf{q}+\tau\boldsymbol{l}$ lies in the interval $[0,\tau)$. Therefore, the TH precoded signal in (9) can be equivalently rewritten as

$$\mathbf{x}_k^{\text{TH}} = \mathbf{s}_k - g_k\mathbf{H}_k\sum_{i=1}^{k-1}\mathbf{y}_i + \tau_k\boldsymbol{l}_k^{\text{TH}} \tag{10}$$

where the term $\tau_k\boldsymbol{l}_k^{\text{TH}}$ accounts for the modulo τ_k operation in (9). Moreover, to reduce the power increment due to the inversion of the effective channel $\mathbf{H}_{\text{eff},k}$, the signal $\mathbf{x}_k^{\text{TH}}$ is further nonlinearly precoded as in (4) by means of the VP algorithm, i.e. $\mathbf{x}_k = \mathbf{x}_k^{\text{TH}} + \tau_k\boldsymbol{l}_k^{\text{VP}}$ where

$$\boldsymbol{l}_k^{\text{VP}} = \underset{\boldsymbol{l}\in\mathbb{C}\mathbb{Z}^{n_k}}{\arg\min}\left\|\mathbf{H}_{\text{eff},k}^{\dagger}\left(\mathbf{x}_k^{\text{TH}}+\tau_k\boldsymbol{l}\right)\right\|^2. \tag{11}$$

By replacing (10) in (11), the two operations can be combined in a single nonlinear precoding operation that leads to the perturbed vector for user k

$$\mathbf{x}_k = \mathbf{s}_k - g_k\mathbf{H}_k\sum_{i=1}^{k-1}\mathbf{y}_i + \tau_k\boldsymbol{l}_k^{\text{TH-VP}}, \tag{12}$$

where the perturbation vector $\boldsymbol{l}_k^{\text{TH-VP}}$ is calculated as

$$\boldsymbol{l}_k^{\text{TH-VP}} = \underset{\boldsymbol{l}\in\mathbb{C}\mathbb{Z}^{n_k}}{\arg\min}\left\|\mathbf{H}_{\text{eff},k}^{\dagger}\left(\mathbf{s}_k - g_k\mathbf{H}_k\sum_{i=1}^{k-1}\mathbf{y}_i+\tau_k\boldsymbol{l}\right)\right\|^2. \tag{13}$$

Unlike standard TH precoding schemes, (13) shows that there is no need of the modulo operation at the transmitter: the perturbation vector $\boldsymbol{l}_k^{\text{TH-VP}}$ that minimizes (13) encompasses also the term $\boldsymbol{l}_k^{\text{TH}}$ in (10) and deals with the power amplification due to both *inter-stream* and *inter-user* interference precancellation. Since the computation of $\boldsymbol{l}_k^{\text{TH-VP}}$ requires the full knowledge of the signals $\mathbf{y}_i$ with index $i<k$, it is performed sequentially, one user after the other.

At the kth user's receiver, assuming that SNR is sufficiently high, the informative message $\mathbf{s}_k$ is recovered by computing the modulo of the received signal vector $\hat{\mathbf{x}}_k$

$$\hat{\mathbf{s}}_k = \text{mod}_{\tau_k}\left(\mathbf{x}_k + g_k\mathbf{H}_k\sum_{i=1}^{k-1}\mathbf{y}_i + g_k\mathbf{z}_k\right) = \mathbf{s}_k + g_k\mathbf{z}_k. \tag{14}$$

B. Efficient Implementation

To implement efficiently the TH-VP architecture, it is possible to compute the filters $\mathbf{F}_k$ $(k=1,2,\ldots,K)$ with just a single QR decomposition of the combined channel matrix $\mathbf{H}$.

Let the QR decomposition of $\mathbf{H}$ be

$$\mathbf{H} = \mathbf{T}\mathbf{Q}^H, \tag{15}$$

where $\mathbf{Q}=[\mathbf{Q}_1,\mathbf{Q}_2,\ldots,\mathbf{Q}_K]$ is an $N_t\times N_r$ (semi-)unitary matrix with $\mathbf{Q}_k$ composed by the n_k columns of $\mathbf{Q}$ corresponding to the channel of user k, and $\mathbf{T}$ is a lower triangular matrix,

$$\mathbf{T} = \mathbf{H}\mathbf{Q} = \begin{bmatrix} \mathbf{T}_{1,1} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{T}_{2,1} & \mathbf{T}_{2,2} & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{T}_{K,1} & \mathbf{T}_{K,2} & \ldots & \mathbf{T}_{K,K} \end{bmatrix}, \tag{16}$$

where the block $\mathbf{T}_{j,k} = \mathbf{H}_j\mathbf{Q}_k$ is an $n_j\times n_k$ matrix. Accordingly, the channel matrix $\mathbf{H}_k$ can be factorized as

$$\mathbf{H}_k = \sum_{i=1}^{k}\mathbf{T}_{k,i}\mathbf{Q}_i^H. \tag{17}$$

We can now state the following theorem, whose proof is omitted for space limitations.

Theorem 1. *The $N_t\times n_k$ precoding filter $\mathbf{F}_k$ for the TH based schemes can be computed as*

$$\mathbf{F}_k = \mathbf{Q}_k\mathbf{T}_{k,k}^{-1}. \tag{18}$$

Employing the results of Theorem 1 and (17), one can rewrite the IUI part in (8) as

$$g_k\mathbf{H}_k\sum_{i=1}^{k-1}\mathbf{y}_i = \sum_{i=1}^{k-1}\sum_{\ell=1}^{k-1}\tilde{\mathbf{T}}_{k,\ell}\mathbf{Q}_\ell^H\mathbf{Q}_i\tilde{\mathbf{T}}_{i,i}^{-1}\mathbf{x}_i = \sum_{i=1}^{k-1}\tilde{\mathbf{T}}_{k,i}\tilde{\mathbf{T}}_{i,i}^{-1}\mathbf{x}_i, \tag{19}$$

where $\tilde{\mathbf{T}}_{k,i} = g_k\mathbf{T}_{k,i}$. Since the matrix $\mathbf{Q}_k$ is quasi-unitary and has no effect on the norm in (13), the calculation of user k's TH-VP perturbation vector can be reformulated as

$$\boldsymbol{l}_k^{\text{TH-VP}} = \underset{\boldsymbol{l}\in\mathbb{C}\mathbb{Z}^{n_k}}{\arg\min}\left\|\mathbf{T}_{k,k}^{-1}\left(\mathbf{s}_k - \sum_{i=1}^{k-1}\tilde{\mathbf{T}}_{k,i}\tilde{\mathbf{T}}_{i,i}^{-1}\mathbf{x}_i+\tau_k\boldsymbol{l}\right)\right\|^2, \tag{20}$$

where all filters are computed with the QR factorization (15).

Employing the results of Theorem 1, we investigate what is the relationship between conventional ZF precoding and the linear precoding adopted for the TH-VP scheme. Let $\mathbf{x}' = \mathbf{s}+\boldsymbol{\Gamma}^{\text{TH-VP}}$ be the TH-VP perturbed vector, where $\boldsymbol{\Gamma}^{\text{TH-VP}}$ is the vector obtained stacking the perturbation Vector of all K users, so that the precoded vector in (12) can be written as

$$\mathbf{x} = \mathbf{x}' + (\mathbf{I}_{N_r} - \mathbf{R})\,\mathbf{x}, \tag{21}$$

where $\mathbf{R}$ is a feedback lower triangular block matrix whose (i,j)-th block is the $n_i\times n_j$-dimensional matrix

$$\mathbf{R}_{i,j} = \begin{cases} \mathbf{G}_i\mathbf{T}_{i,j}\mathbf{T}_{j,j}^{-1}\mathbf{G}_j^{-1} & \text{if } i\geq j \\ \mathbf{0} & \text{otherwise} \end{cases}. \tag{22}$$

Using the results of Theorem 1, the feedforward filter $\mathbf{F}$ is

$$\mathbf{F} = \mathbf{Q}\boldsymbol{\Psi}, \tag{23}$$

where $\boldsymbol{\Psi} = \text{diag}\left(\mathbf{T}_{1,1}^{-1},\mathbf{T}_{2,2}^{-1},\cdots,\mathbf{T}_{K,K}^{-1}\right)$. Since from (21) is $\mathbf{x}=\mathbf{R}^{-1}\mathbf{x}'$ and from (22) is $\mathbf{R}=\mathbf{G}\mathbf{T}\boldsymbol{\Psi}\mathbf{G}^{-1}$, the transmitted signal vector $\mathbf{y}$ can be written as

$$\mathbf{y} = \mathbf{F}\mathbf{G}^{-1}\mathbf{x} = \mathbf{Q}\mathbf{T}^{-1}\mathbf{G}^{-1}\mathbf{x}'. \tag{24}$$

Inspection of (24) shows that, combining BD, TH and channel inversion, the linear filter of the TH-VP precoder is equivalent to the pseudo-inverse of the channel matrix $\mathbf{H}$ just as in the ZF-VP precoder. In practice, what differs between the two schemes is that the ZF-VP perturbation vector (4) is computed jointly for all users while the TH-VP precoding (20) is carried out independently per each user optimizing a metric substantially different from the one of ZF-VP. This adds more flexibility and lower complexity to the latter scheme as we will see in next sections.

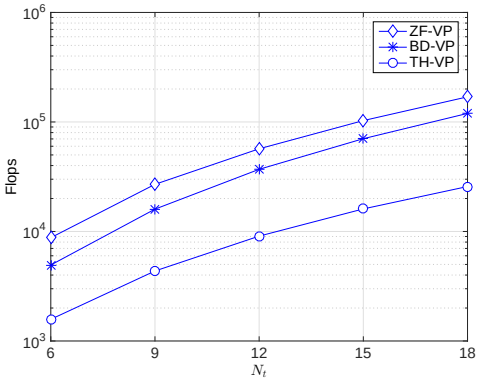

Fig. 1: The computational complexity of ZF-VP, BD-VP, TH-VP in the case of $N_r = N_t, K = 3, n_k = N_t/3$.

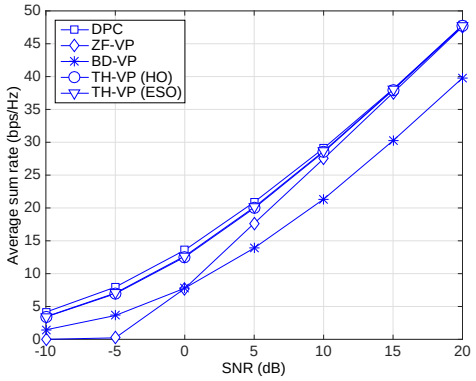

Fig. 2: Sum rate of DPC, ZF-VP, BD-VP and TH-VP (ESO and HO) vs. SNR.

C. Complexity Analysis

In this section, we evaluate the computational complexity of TH-VP and compare it with the complexity of ZF-VP and BD-VP presented in [4] and [5], respectively. The complexity of TH-VP is computed calculating the QR decomposition in (15), the filter implementation in (18) and the precoding in (20) to find the perturbation vector $\boldsymbol{l}_k^{\text{TH-VP}}$. In detail, the required floating point operations (flops) of each matrix operation are: inversion of an $m \times m$ matrix using Gauss-Jordan elimination, $4m^3/3$; multiplication of an $m \times n$ matrix and an $n \times p$ matrix, $2mnp$; QR decomposition of an $m \times n$ matrix ($m \geq n$), $2n^2(m - n/3)$. The minimum distance search of perturbation vector $\boldsymbol{l}^{\text{opt}}$ is performed with the complex LLL (CLLL) lattice-reduction-aided algorithm [14] to reduce the complexity of sphere encoding. Although one common approximation for its complexity is $\mathcal{O}(n^4 \log \mathcal{K})$ flops, with $\mathcal{K}$ being the norm of the longest basis, the exact computational load of the CLLL algorithm is difficult to analyze and we evaluate its complexity through simulation in the results plotted in Fig. 1.

Table I presents the computational complexity of TH-VP, ZF-VP and BD-VP assuming $N_r = N_t$. Moreover, if we assume that the population of users is static and the channel remains constant over several signalling intervals, the computation of the linear precoder coefficients can be neglected and the algorithm complexity evaluated focusing on the nonlinear precoding that needs to be updated every signalling interval.

Fig. 1 plots the number of flops for ZF-VP, BD-VP and TH-VP in this static scenario. The complexity is computed as function of the number of transmit antennas N_t when $N_r = N_t$, $K = 3$ and each user has the same number of receiver antennas, i.e., $n_k = N_t/K$. The TH-VP scheme is the one that requires the least number of flops and its complexity is in certain cases one tenth of that of ZF-VP.

IV. Numerical Results

In this section we present the performance of the proposed algorithm in terms of achievable sum rate and BER. The sum rate results of TH-VP, computed following the derivation presented in [5] for BD-VP, are compared with the results of BD-VP and the upper bound for ZF-VP presented in [7]. Unlike the other schemes, the sum rate results of TH-VP have been obtained by employing a power optimization algorithm, whose derivation we omit for lack of space, formulated to maximize the sum rate with a constraint on the total transmit power P that can be solved in a waterfilling fashion. The BER results are computed assuming uniform power distribution among users for all the schemes.

Unless differently noted, we assume that the BS, equipped with $N_t = 6$ transmit antennas, transmits simultaneously to $K = 3$ users each equipped with $n_k = 2$ receive antennas. The elements of each user's channel matrix are modeled as independent complex Gaussian random variables with zero mean and unity variance. The SNR is defined as transmit power per user versus noise power.

Fig. 2 compares the average sum rate for the various algorithms discussed in this paper. The results for DPC, the theoretical optimum, are added as a reference. In particular, the results for TH-VP depend on the particular ordering of the users. We implement two different user ordering strategies: the *exhaustive search ordering* (ESO) strategy, that exhaustively computes the achievable rate for all the possible user orderings and selects the best one and the *heuristic ordering* (HO) strategy, that sets the users in descending order according the Frobenius norm of their channel matrix, i.e., the user with the best channel is labelled with index $k = 1$ and gets the largest order of diversity while the user with the worst channel is labelled with the index $k = K$ and gets the least order of diversity. Exhaustive search requires the evaluation of $K!$ precoding vectors and accordingly is valid only as a benchmark. Results show that TH-VP (HO) has performance very close to what is obtained with optimal exhaustive search and outperforms all other practical schemes, including the much more complex ZF-VP, being very close to the theoretical optimum represented by DPC.

Fig. 3 shows the sum rate of the various scheme as function of the number of users K for SNR $= 5$ dB. Each mobile user is equipped with $n_k = 2$ antennas and the number of antennas at the BS is $N_t = 2 \times K$. As in the previous cases, TH-VP is very close to DPC and largely outperforms all other schemes. In particular, as K grows the gap between BD-VP precoding and the other algorithms grows because BD-VP is not able to exploit the full diversity of the system.

Fig. 4 shows the BER results of each for the various precod-

TABLE I: The number of floating point operations of ZF-VP, BD-VP and TH-VP

	ZF-VP	BD-VP	TH-VP
Channel inversion/QR decomposition	$4N_t^3/3$	$4N_t^3/3$	$4N_t^3/3$
CLLL algorithm	$\mathcal{O}(N_t^4 \log \mathcal{K})$	$\sum_{k=1}^{K} \mathcal{O}(n_k^4 \log \mathcal{K}')$	$\sum_{k=1}^{K} \mathcal{O}(n_k^4 \log \mathcal{K}'')$
Vector perturbation	$4N_t^3/3 + 4N_t^2 + 2N_t$	$\sum_{k=1}^{K}(4n_k^3/3 + 4n_k^2 + 2n_k)$	$\sum_{k=1}^{K}[4n_k^3/3 + 4kn_k^2 + 6n_k]$
Linear precoding	$2N_t^2 + 8N_t$	$4KN_t + \sum_{k=1}^{K}(2N_t n_k + 4n_k)$	$4KN_t + \sum_{k=1}^{K}(2kN_t n_k + 2n_k^2 + 4n_k)$

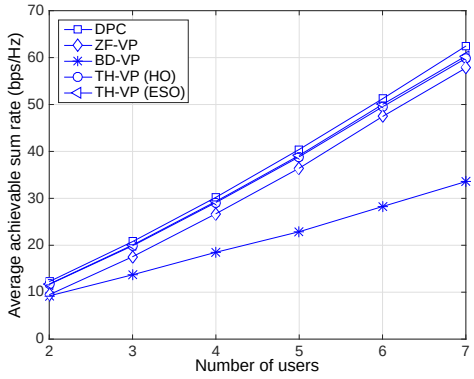

Fig. 3: Sum rate of DPC, ZF-VP, BD-VP and TH-VP (ESO and HO) vs. the number of users K for SNR $= 5$ dB, $K = N_t/2$.

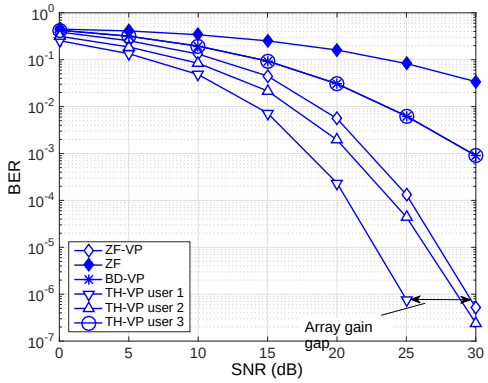

Fig. 4: BER of ZF, ZF-VP, BD-VP and TH-VP vs. SNR.

ing schemes, including conventional ZF, when the transmitted symbols belong to the 64-QAM constellation. Unlike all the other schemes considered, the performance of TH-VP (with HO) are different for the various users: Users $k = 1$ and $k = 2$ outperform both ZF-VP and BD-VP, user $k = 3$ performs worse than ZF-VP and has the same performance of BD-VP. User $k = 3$ and BD-VP have the same results because in the simulated scenario both schemes apply VP precoding in combination with the inversion of a 2×2 channel matrix. The behavior of TH-VP with respect to ZF-VP can be interpreted by an analysis of the diversity order and the array gain of the two schemes. The BER curves' slope, i.e., their diversity order, for the two schemes is $d_1^{\text{TH-VP}} = 6, d_2^{\text{TH-VP}} = 4, d_3^{\text{TH-VP}} = 2$ and $d^{\text{ZF-VP}} = 6$, respectively. The array gain values for TH-VP are $a_1^{\text{TH-VP}} = 3$, $a_2^{\text{TH-VP}} = 2$ and $a_3^{\text{TH-VP}} = 1$, while the array gain for ZF-VP is $a^{\text{ZF-VP}} = 1$. Accordingly, user $k = 1$ of TH-VP has an approximate 4.8dB SNR gain over all the users of ZF-VP, as shown by the simulation results.

V. Conclusions

In this paper we present the TH-VP precoding algorithm, a novel precoding scheme for MU-MIMO broadcast channels, which combines TH and VP precoding to nullify both inter-user and inter-stream interference. We show that TH-VP has a computational load smaller than similar nonlinear techniques while the sum rate results approach asymptotically the sum capacity of DPC and outperform other existing state-of-the-art nonlinear precoding techniques.

References

[1] M. Costa, "Writing on dirty paper," *IEEE Trans. Inf. Theory*, vol. 29, no. 3, pp. 439–441, 1983.

[2] C. B. Peel, B. M. Hochwald, and A. L. Swindlehurst, "A vector-perturbation technique for near-capacity multiantenna multiuser communication-part I: channel inversion and regularization," *IEEE Trans. Commun.*, vol. 53, no. 1, pp. 195–202, 2005.

[3] R. F. Fischer, *Precoding and signal shaping for digital transmission*. John Wiley & Sons, 2005.

[4] B. Hochwald, C. Peel, and A. Swindlehurst, "A vector-perturbation technique for near-capacity multiantenna multiuser communication-part II: perturbation," *IEEE Trans. Commun.*, vol. 53, no. 3, 2005.

[5] C.-B. Chae, S. Shim, and R. Heath, "Block diagonalized vector perturbation for multiuser MIMO systems," *IEEE Trans. on Wireless Commun.*, vol. 7, no. 11, pp. 4051–4057, Nov. 2008.

[6] D. A. Schmidt, M. Joham, and W. Utschick, "Minimum mean square error vector precoding," *Europ. Trans. on Telecommun.*, vol. 19, no. 3, pp. 219–231, 2008.

[7] A. Razi, D. Ryan, I. Collings, and J. Yuan, "Sum rates, rate allocation, and user scheduling for multi-user MIMO vector perturbation precoding," *IEEE Trans. Wireless Commun.*, vol. 9, no. 1, pp. 356–365, 2010.

[8] R. Chen, C. Li, J. Li, and Y. Zhang, "Low complexity user grouping vector perturbation," *IEEE Wireless Commun. Letters*, vol. 1, no. 3, 2012.

[9] S. P. Herath, D. H. N. Nguyen, and T. Le-Ngoc, "Vector perturbation precoding for multi-user CoMP downlink transmission," *IEEE Access*, vol. 3, no. 1, 2015.

[10] S. Herath, D. Nguyen, and T. Le-Ngoc, "Vector-perturbation precoding under quantized CSI," *IEEE Trans. Veh. Technol.*, vol. 65, no. 1, pp. 420–427, 2016.

[11] J. Yang, X. Wang, S. Chen, and Z. Zhong, "Design of low-complexity vector perturbation precoding technique for multiuser MIMO systems," in *Proc. of IEEE WCSP*, 2016.

[12] J. Park, B. Lee, and B. Shim, "A MMSE vector precoding with block diagonalization for multiuser MIMO downlink," *IEEE Trans. Commun.*, vol. 60, no. 2, pp. 569–577, 2012.

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

[14] C. Windpassinger, R. Fischer, and J. Huber, "Lattice-reduction-aided broadcast precoding," *IEEE Trans. Commun.*, vol. 52, no. 12, 2004.

[15] V. Stankovic and M. Haardt, "Successive optimization Tomlinson-Harashima precoding (SO THP) for multi-user MIMO systems," in *Proc. of IEEE ICASSP '05*, March 2005.

[16] M. Moretti, L. Sanguinetti, and X. Wang, "Resource allocation for power minimization in the downlink of THP-based spatial multiplexing MIMO-OFDMA systems," *IEEE Trans. on Vehic. Techn.*, vol. 64, no. 1, pp. 405–411, Jan. 2015.

[17] M. Moretti and A. Perez-Neira, "Efficient margin adaptive scheduling for MIMO-OFDMA systems," *IEEE Trans. Wireless Commun.*, vol. 12, no. 1, pp. 278 –287, Jan. 2013.