14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

NON-LINEAR PRE-CODING FOR MIMO MULTI-USER DOWNLINK TRANSMISSIONS WITH DIFFERENT QoS REQUIREMENTS

Luca Sanguinetti and Michele Morelli

University of Pisa
Department of Information Engineering
Pisa, Italy
luca.sanguinetti@iet.unipi.it, michele.morelli@iet.unipi.it

ABSTRACT

An efficient non-linear pre-filtering technique based on Tomlinson-Harashima pre-coding (THP) has recently been proposed by Liu and Krzymien for multiple antenna multi-user systems. The algorithm is based on the Zero-Forcing (ZF) criterion and assumes a number of transmit antennas equals to the number of active users. In contrast to other methods, it ensures a fair treatment of the active users providing them the same signal-to-noise ratio. In multimedia applications, however, several types of information with different quality-of-service (QoS) must be supported. Motivated by the above problem, in the present work we design a ZF THP-based pre-filtering algorithm for multiple antenna multi-user networks in which the base station allocates the transmit power according to the QoS requirement of each active user. In doing so, we consider a system in which the number of active users may be less than the number of transmit antennas. As we will see, in such a case there exists an infinite number of solutions satisfying the ZF criterion. We address the problem of finding the *best* using as optimality criterion the maximization of the signal-to-noise ratios at all mobile terminals.

1. INTRODUCTION

Over the last years, multiple-input multiple-output (MIMO) techniques have received a lot of attention due to their potential benefits in terms of spectral efficiency and/or diversity gain [1]. Multiple antenna multi-user systems have recently been proposed as a mean to combine the high capacity provided by MIMO processing with the multiple-access capability of space-division-multiplexing (SDM) [2]. The main impairment of MIMO multi-user systems is represented by the multiple-access interference (MAI) arising from the simultaneous transmission of parallel data streams over the same frequency band. In uplink transmissions interference mitigation is typically accomplished at the base station (BS) using linear multi-user detectors or non-linear techniques based on layered architectures [3]. The above schemes require joint processing of the received signals and, accordingly, are not suited for downlink transmissions where coordination among spatially distributed users is not possible. In these circumstances, interference mitigation can only be accomplished at the transmit side using pre-filtering techniques. This requires some form of channel state information (CSI) at the transmitter, which can be achieved in time division duplex (TDD) systems by exploiting the channel reciprocity between alternate uplink and downlink transmissions.

Tomlinson-Harashima pre-coding (THP) schemes have recently been proposed as viable candidates to mitigate the detrimental effects of interference in downlink transmissions. References [4] -[8] provide a good sample of the results obtained in this area. In particular, in [4] the processing matrices of the THP algorithm are designed according to a minimum mean square error (MMSE) approach under a constraint on the overall transmit power. Unfortunately, the solution to this problem requires a large number of matrix inversions (equal to the number of active users) and may be infeasible when applied to heavy-loaded systems. An efficient implementation of the above algorithm is discussed in [5], where all matrix inversions are replaced by a single Cholesky factorization. The method reported in [6] is based on the QR-decomposition of the channel matrix and selects the pre-filtering coefficients according to the zero-forcing (ZF) criterion so as to completely remove the interference at the receivers. In [7], this approach is extended to a multi-user MC-CDMA network in which the mobile terminals (MTs) can employ conventional single-user detection (SUD) techniques. Albeit reasonable, the methods reported in [6]- [7] are not based over any optimality criterion and produce significant differences in the error rate performance of the receive terminals, which may be undesirable if fair treatment of the active users is required. A solution to this problem is provided in [8], where the THP matrices are computed so as to ensure the same signal-to-noise ratio (SNR) at each MT. The main drawback of this solution is that it requires a number N_T of transmit antennas equals to the number K of active users. This may prevent its applicability to practical multi-user systems, where the number of contemporarily active users varies dynamically according to the traffic evolution while N_T is fixed.

In this work we return to the problem discussed in [8] and propose a THP-based scheme that can be used even when $K < N_T$. As we will see, in such a case there exists an infinite number of solutions satisfying the ZF criterion. We address the problem of finding the *best* using as optimality criterion the maximization of the SNRs at all mobile terminals. Compared to [8], our scheme has lower complexity and allows the BS to allocate the transmit power according to the Quality-of-Service (QoS) constraints of each user. This is a significant advantage for future communications systems where the traffic is a mixture of heterogenous multimedia applications, each of which with different QoS constraints.

The remainder of this work is organized as follows. Next section outlines the signal model and introduces basic notation. In Section 3 we derive the pre-filtering scheme using

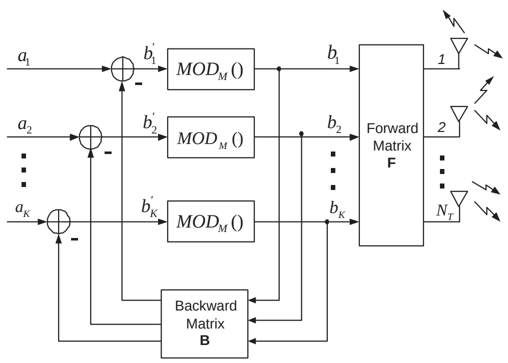

Figure 1: Block diagram of the transmitter.

the ZF criterion. Numerical results are provided in Section 4, while conclusions are drawn in Section 5.

2. SYSTEM MODEL

We consider the downlink of a network in which the BS employs N_T transmit antennas to communicate with K active users ($K \leq N_T$), each equipped with a single receive antenna. The user data are collected into the K-dimensional vector $\mathbf{a} = [a_1, a_2, ..., a_K]^T$ (the notation $[\cdot]^T$ denotes the transpose operation) and are taken from an M-QAM constellation with energy $\sigma_a^2 = 2(M-1)/3$. This amounts to saying that the real and imaginary parts of a_k belong to the set $A = \in \{\pm 1, \pm 3, ..., \pm\sqrt{M}-1\}$.

Figure 1 shows the THP-based scheme under investigation. It consists of a *backward* matrix $\mathbf{B}$, K non-linear operators $MOD_M(\cdot)$ and a *forward* matrix $\mathbf{F}$. As discussed in [9], $\mathbf{B}$ is *strictly* lower triangular to allow data pre-coding in a recursive fashion while the modulo operator acts independently over the real and imaginary parts of its input according to the following rule

$$MOD_M(x) = x - 2\sqrt{M} \cdot \left\lfloor \frac{x - \sqrt{M}}{2\sqrt{M}} \right\rfloor \qquad (1)$$

where the notation $\lfloor c \rfloor$ indicates the *smallest* integer larger than or equal to c. A close observation of (1) reveals that $MOD_M(x)$ performs a periodic mapping of x onto the interval $(-\sqrt{M}, \sqrt{M}]$. In practice, the pre-coded symbols b_k are constrained into the square region $\aleph = \{x^{(R)} + jx^{(I)} | x^{(R)}, x^{(I)} \in (-\sqrt{M}, \sqrt{M}]\}$ and the transmit power is consequently reduced with respect to linear pre-filtering.

Using (1) it is seen that the elements of $\mathbf{b}$ can be iteratively computed as

$$b_k = a_k - \sum_{l=1}^{k-1} [\mathbf{B}]_{k,l} b_l + d_k \quad k = 1, 2, \ldots, K \qquad (2)$$

where $[\cdot]_{k,\ell}$ is the (k,ℓ)th entry of the enclosed matrix, $d_k = 2\sqrt{M} p_k$ and p_k is a complex-valued quantity whose real and imaginary parts are suitable integers that reduce b_k to the square region $\aleph$. Clearly, a unique p_k exists with such a property.

From (2) we see that the modulo operation in (1) is equivalent to adding a vector $\mathbf{d} = [d_1, d_2, \ldots, d_K]^T$ to the input data symbols $\mathbf{a}$. This results in the equivalent block diagram illustrated in Figure 2, from which it follows that

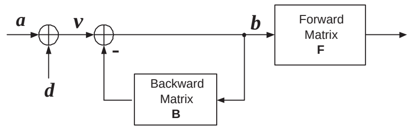

Figure 2: Equivalent block diagram of the transmitter.

$$\mathbf{b} = \mathbf{C}^{-1}\mathbf{v} \qquad (3)$$

where we have defined $\mathbf{v} = \mathbf{a} + \mathbf{d}$ and $\mathbf{C} = \mathbf{B} + \mathbf{I}$ ($\mathbf{I}$ denotes the identity matrix of order K). Note that $\mathbf{C}$ is a unit-diagonal and lower triangular matrix, i.e., $[\mathbf{C}]_{k,k} = 1$ and $[\mathbf{C}]_{k,\ell} = 0$ for $k < \ell$.

The pre-coded symbols b_k are then passed to the forward matrix $\mathbf{F}$. The resulting N_T- dimensional vector $\mathbf{s} = \mathbf{F}\mathbf{b}$ is finally transmitted over the channel using the N_T antennas of the BS array. The channel is flat-fading and it is mathematically described by a matrix $\mathbf{H}$ of dimensions $K \times N_T$. In particular, $[\mathbf{H}]_{k,i}$ represents the channel gain from the ith transmit antenna to the kth MT. Then, the received signal at the kth MT can be written as

$$y_k = \mathbf{H}_k \mathbf{F}\mathbf{b} + w_k \qquad (4)$$

where $\mathbf{H}_k$ indicates the kth row of $\mathbf{H}$ while w_k accounts for the thermal noise and it is modelled as a Gaussian random variable with zero-mean and variance σ_w^2. At the receiver side, each sample y_k is scaled by the automatic gain control (AGC) unit and fed to the same modulo operator employed at the transmitter so as to remove the effect of d_k. The output is finally passed to a threshold device which delivers an estimate of a_k.

Stacking the received signals of all users into a single vector $\mathbf{y} = [y_1, y_2, ..., y_K]^T$ and bearing in mind (3), we may write

$$\mathbf{y} = \mathbf{H}\mathbf{F}\mathbf{C}^{-1}\mathbf{v} + \mathbf{w} \qquad (5)$$

where $\mathbf{w} = [w_1, w_2, ..., w_K]^T$ is a Gaussian vector with zero-mean and covariance matrix $\sigma_w^2\mathbf{I}$. In the next Section we show how the backward and forward matrices can be designed according to the ZF criterion.

3. DESIGN OF THE BACKWARD AND FORWARD MATRICES

3.1 Optimality Criterion

From (5) we see that the received signals depends on the THP matrices $\mathbf{F}$ and $\mathbf{C}$. The latter must be designed so as to mitigate the MAI while enhancing the desired signal component. To this purpose, we adopt a procedure in which a ZF approach is first employed to completely eliminate the MAI and the result is then exploited to maximize the SNR at all MTs. The resulting scheme is suitable for multimedia applications with different QoS requirements since it can provide any set of relative SNRs at the MTs.

To maintain the same power as in the case where no pre-filtering is performed, we impose a constraint on the overall transmit power. Assuming that the pre-coded symbols b_k are statistically independent with zero mean and power σ_a^2 [6], the power constraint can be mathematically expressed as

$$\mathrm{tr}\left\{\mathbf{F}^H\mathbf{F}\right\} = K \tag{6}$$

where $\mathrm{tr}\{\cdot\}$ denotes the trace of the matrix.

3.2 ZF Design

The complete elimination of MAI at all MTs implies that

$$\mathbf{y} = \sqrt{e}\boldsymbol{\Lambda}\mathbf{v} + \mathbf{w} \tag{7}$$

where e is a non-negative real-valued parameter that must be chosen so as to meet the power constraint (6) while $\boldsymbol{\Lambda} = \mathrm{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_K\}$ is a diagonal matrix with real-valued components satisfying the identity $\mathrm{tr}\{\boldsymbol{\Lambda}\} = K$. The sample y_k at the kth MT is fed to the AGC unit and ideally divided by $\lambda_k\sqrt{e}$. Next, it is passed to a non-linear device that operates according to (1). Neglecting for simplicity the modulo-folding effect on the thermal noise, the output from the non-linear device takes the form

$$z_k = a_k + \frac{w_k}{\lambda_k\sqrt{e}} \tag{8}$$

and the corresponding SNR is given by

$$SNR_k = e \cdot \frac{\lambda_k^2\sigma_a^2}{\sigma_w^2}. \tag{9}$$

The above equation indicates that BS can ensure any set of relative SNRs at the MTs by properly selecting the elements of $\boldsymbol{\Lambda}$. Note that this is in contrast with the method discussed in [8], where the same SNR is guaranteed at each remote unit.

Comparing (5) with (7) leads to the following *modified* ZF condition

$$\mathbf{HF} = \sqrt{e}\boldsymbol{\Lambda}\mathbf{C}. \tag{10}$$

Letting $\tilde{\mathbf{H}} = \boldsymbol{\Lambda}^{-1}\mathbf{H}$ and $\mathbf{U} = (1/\sqrt{e})\cdot\mathbf{F}$, we may rewrite the above equation in the equivalent form

$$\mathbf{C}^{-1}\tilde{\mathbf{H}}\mathbf{U} = \mathbf{I} \tag{11}$$

while the power constraint (6) becomes

$$\mathrm{tr}\left\{\mathbf{U}^H\mathbf{U}\right\} = \frac{K}{e}. \tag{12}$$

The ZF condition in (11) is now exploited to compute $\mathbf{U}$. For this purpose, we observe that (11) is a system of K^2 linear equations whose unknowns are the $N_T \times K$ entries of $\mathbf{U}$. Since $K \leq N_T$, we observe that (11) may have more equations than unknowns. In such a case there exists an infinite number of matrix $\mathbf{U}$ satisfying (11) and the problem is to find the *best*. As mentioned earlier, we provide a solution to this problem looking for the maximum of the SNR at each MT.

Solving (12) with respect to e and substituting into (9) yields

$$SNR_k = \frac{\lambda_k^2\sigma_a^2}{\sigma_w^2} \cdot \frac{K}{\mathrm{tr}\{\mathbf{U}^H\mathbf{U}\}} \tag{13}$$

from which it follows that the matrix $\mathbf{U}$ maximizing the right-hand side (RHS) of (13) is the minimum-norm solution of (11). The latter is found as the pseudo-inverse of $\mathbf{C}^{-1}\tilde{\mathbf{H}}$ [10] and reads $\mathbf{U} = \tilde{\mathbf{H}}^H(\tilde{\mathbf{H}}\tilde{\mathbf{H}}^H)^{-1}\mathbf{C}$ or, equivalently,

$$\mathbf{U} = \mathbf{H}^H\left(\mathbf{H}\mathbf{H}^H\right)^{-1}\boldsymbol{\Lambda}\mathbf{C} \tag{14}$$

where we have borne in mind that $\tilde{\mathbf{H}} = \boldsymbol{\Lambda}^{-1}\mathbf{H}$. Substituting (14) into (13), we obtain

$$SNR_k = \frac{\lambda_k^2\sigma_a^2}{\sigma_w^2} \cdot \frac{K}{\mathrm{tr}\{\mathbf{C}^H\boldsymbol{\Lambda}^H(\mathbf{H}\mathbf{H}^H)^{-1}\boldsymbol{\Lambda}\mathbf{C}\}}. \tag{15}$$

In Appendix it is shown that the unit-diagonal and lower triangular matrix $\mathbf{C}$ maximizing the RHS of (15) is given by

$$\mathbf{C} = \boldsymbol{\Lambda}^{-1}\mathbf{L}\mathbf{D} \tag{16}$$

where $\mathbf{L}$ is taken from the Cholesky factorization of the matrix $\mathbf{H}\mathbf{H}^H$ and $\mathbf{D}$ is a $K \times K$ diagonal matrix which scales the elements on the main diagonal of $\mathbf{C}$ to unity, i.e., $\mathbf{D} = \mathrm{diag}\left\{\lambda_k/[\mathbf{L}]_{k,k}; 1 \leq k \leq K\right\}$.

Substituting (16) into (14) and recalling that $\mathbf{F} = \sqrt{e}\cdot\mathbf{U}$, we obtain

$$\mathbf{F} = \sqrt{e}\cdot\mathbf{H}^H\left(\mathbf{L}^{-1}\right)^H\mathbf{D} \tag{17}$$

where e is found from (12) using the identity $\mathrm{tr}\left\{\mathbf{U}^H\mathbf{U}\right\} = \mathrm{tr}\left\{\mathbf{D}^H\mathbf{D}\right\}$ and reads

$$e = \frac{K}{\sum\limits_{n=1}^{K}(\lambda_n/[\mathbf{L}]_{n,n})^2}. \tag{18}$$

Finally, substituting (18) into (9) yields the SNR at the kth MT

$$SNR_k = \frac{\lambda_k^2\sigma_a^2}{\sigma_w^2} \cdot \frac{K}{\sum\limits_{n=1}^{K}(\lambda_n/[\mathbf{L}]_{n,n})^2}. \tag{19}$$

The above equation indicates that BS can ensure any set of relative SNRs at the MTs by properly selecting the elements of $\boldsymbol{\Lambda}$. In the sequel, the scheme that employs the matrices $\mathbf{F}$ and $\mathbf{C}$ given in (17) and (16) is called the Zero-Forcing THP (ZF-THP) algorithm.

3.3 Remarks

1) Letting $\mathbf{B} = \mathbf{0}$ (corresponding to $\mathbf{C} = \mathbf{I}$) leads to the ZF linear (ZF-L) pre-coding scheme. In these circumstances the forward matrix $\mathbf{F}$ takes the form

$$\mathbf{F} = \sqrt{e'}\cdot\mathbf{H}^H\left(\mathbf{H}\mathbf{H}^H\right)^{-1}\boldsymbol{\Lambda} \tag{20}$$

with $e' = K/\mathrm{tr}\{\boldsymbol{\Lambda}^H(\mathbf{H}\mathbf{H}^H)^{-1}\boldsymbol{\Lambda}\}$, while the SNR at the kth MT is given by

$$SNR'_k = \frac{\lambda_k^2\sigma_a^2}{\sigma_w^2} \cdot \frac{K}{\sum\limits_{n=1}^{K}(\lambda_n/[\mathbf{L}]_{n,n})^2 + \sum\limits_{n=2}^{K}\sum\limits_{\ell=1}^{n-1}\left|[\mathbf{L}^{-1}\boldsymbol{\Lambda}]_{n,\ell}\right|^2}. \tag{21}$$

Comparing (21) with (19) indicates that $SNR'_k \leq SNR_k$, meaning that ZF-L cannot perform better than ZF-THP.

2) Setting $\boldsymbol{\Lambda} = \mathbf{I}$ into (16) produces $\mathbf{C} = \mathbf{L}\mathbf{D}$ while $\mathbf{F}$ is still given in (17) with $\mathbf{D} = \mathrm{diag}\left\{1/[\mathbf{L}]_{k,k}; 1 \leq k \leq K\right\}$ and

$e = K / \sum_{k=1}^{K} (1/[\mathbf{L}]_{k,k})^2$. It is interesting to compare this result with the THP-based scheme discussed in [6] , where $\mathbf{C} = \mathbf{DL}$ and $\mathbf{F} = \mathbf{H}^H \left(\mathbf{L}^{-1}\right)^H$. As it is seen, our solution only differs for the presence of the scalar factor $\sqrt{e}$ and for the different position of $\mathbf{D}$ into the expression of the backward and forward matrices. However, selecting $\mathbf{C}$ and $\mathbf{F}$ as indicated in [6] leads to

$$SNR_k = \frac{\sigma_a^2 \, [\mathbf{L}]_{k,k}^2}{\sigma_w^2}. \tag{22}$$

Since $\mathbf{L}$ is taken from the Cholesky factorization of $\mathbf{HH}^H$, the above equation indicates that the set of relative SNRs depends on the actual channel matrix $\mathbf{H}$. This means that they cannot be designed by the BS according to the QoS requirements of the active users.

3) Setting $\mathbf{\Lambda} = \mathbf{I}$ and $K = N_T$ into (16)-(18) leads to the THP matrices employed by Liu and Krzymien (L&K) in [8]. Compared to L&K, however, our scheme is simpler to implement as it only involves the Cholesky factorization of $\mathbf{HH}^H$ whereas $K-1$ matrix inversions (of decreasing order from K to 2) are required in [8]. Moreover, the proposed algorithm can adjust the relative SNRs at the remote units by properly selecting the elements of $\mathbf{\Lambda}$ and has a wider application range than L&R as it can be used even when $K < N_T$.

4) It is well known that the ordering strategy adopted during the pre-coding process has a significant impact on the performance of THP. The optimal order is the one that maximizes the SNR at all MTs and it can only be found through an exhaustive search over all possible permutations of the parallel data streams. Unfortunately, this may require excessive computations in the presence of many active users (heavy-loaded system). An interesting alternative to the exhaustive search is the *best-first* ordering strategy, which was originally proposed in [3] for the vertical Bell Labs layered space-time (V-BLAST) architecture and recently extended to THP in [5] and [8]. These methods operate on the rows of the channel matrix and in most cases achieve the optimal order with a significant reduction of complexity with respect to the exhaustive search. In the sequel, we adopt the optimal method proposed by [8] after replacing $\mathbf{H}$ with the modified channel matrix $\tilde{\mathbf{H}}$.

4. SIMULATION RESULTS

Computer simulations have been run to assess the performance of ZF-THP in terms of uncoded bit-error-rate (BER) vs. E_t/N_0, where E_t is the average transmitted energy per bit and N_0 the one-sided noise power spectral density. The information bits are mapped onto 16-QAM symbols and the number of transmit antennas is fixed to $N_T = 4$. A new channel matrix $\mathbf{H}$ is generated at each simulation run. Its entries are modeled as independent Gaussian random variables with zero-mean and unit variance. Perfect channel knowledge is assumed at the transmitter end. As mentioned earlier, the optimal order is found by using the *best-first* ordering strategy discussed in [8].

Figure 3 illustrates the BER of ZF-THP for $\mathbf{\Lambda} = \mathbf{I}$ and three active users ($K = 3$). Comparisons are made with ZF-L and L&K. Since the latter is specifically designed for $N_T = K$, only the first K transmit antennas out of a total of N_T are used for transmission. As is intuitively clear, in this case L&K does not exploit the spatial diversity offered by the MIMO channel and its performance is even worse than that achievable with ZF-L. To overcome this problem, we propose an *improved* version of the L&K algorithm in which the *best K* antennas (i.e., those maximizing the SNR of the active users for the actual channel realization) are selected for transmissions. This scheme is referred to as L&K-AS (L&K with Antenna Selection) in the sequel. In this way some form of spatial diversity is achieved and significant gains are observed with respect to both L&K and ZF-L. The best results are obtained with ZF-THP, which exhibits a gain of approximately 1 dB as compared to L&K-AS. Note that ZF-THP is much simpler to implement than L&K-AS since the latter requires $K-1$ matrix inversions for each possible antenna selection.

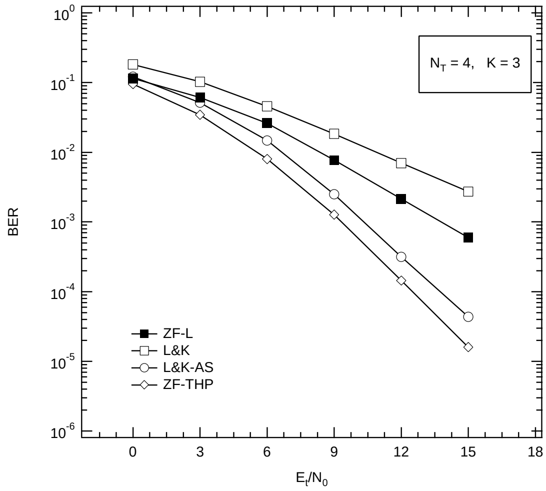

Figure 3: BER performance vs. E_t/N_0 with $K = 3$.

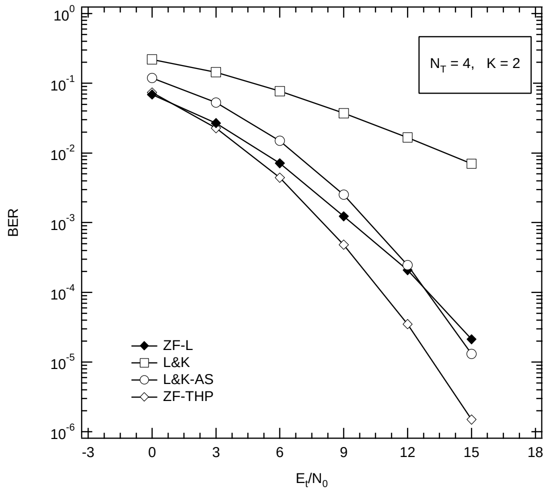

Figure 4: BER performance vs. E_t/N_0 with $K = 2$.

Figure 4 illustrates the performance of ZF-THP, ZF-L, L&K and L&K-AS in the same operating conditions of Figure 3 except that now two users are simultaneously active ($K = 2$). As expected, the system performance of ZF-THP, ZF-L and L&K-AS improve as the spatial diversity increases. In contrast to the previous results, we see that now ZF-L performs even better than L&K-AS for $E_t/N_0 < 12$ dB. Again,

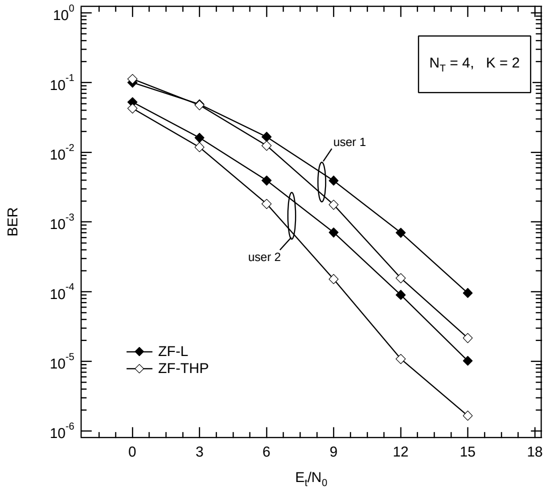

Figure 5: BER performance vs. E_t/N_0 in presence of two active users with different QoS requirements.

ZF-THP gives the best results. In particular, at an error rate of 10^{-3} the gain of ZF-THP with respect to ZF-L is approximately 1.5 dB while it becomes 2.5 dB with L&K-AS.

Figure 5 shows the BER of ZF-THP in case of two users with different QoS requirements. In particular, we set $\lambda_2 = \sqrt{2}\lambda_1$ so that SNR_2 is 3 dB higher than SNR_1. Comparisons are only made with ZF-L since L&K and L&K-AS cannot ensure different SNRs at the MTs. As expected, ZF-THP performs better than ZF-L.

5. CONCLUSIONS

We have derived a non-linear pre-filtering scheme for MIMO multi-user downlink transmissions based on Tomlinson-Harashima pre-coding. The proposed solution aims at maximizing the SNR at all mobile terminals under a ZF constraint. At the same time, it allows the BS to allocate the available power according to the QoS requirements of each user. In contrast to other existing methods, this scheme is simpler to implement and can effectively exploit the spatial diversity provided by the MIMO channel when the number of simultaneously active users is less than the number of transmit antennas.

6. APPENDIX

In this Appendix we highlight the major steps leading to (16). Our goal is to find the matrix $\mathbf{C}$ that maximizes the RHS of (15). This is equivalent to minimizing

$$J = \mathrm{tr}\{\mathbf{C}^H \mathbf{\Lambda}^H (\mathbf{H}\mathbf{H}^H)^{-1} \mathbf{\Lambda}\mathbf{C}\}. \tag{23}$$

We begin by considering the Cholesky factorization of the matrix $\mathbf{H}\mathbf{H}^H$ in (23), i.e.,

$$\mathbf{H}\mathbf{H}^H = \mathbf{L}\mathbf{L}^H \tag{24}$$

where $\mathbf{L}$ is a $K \times K$ lower triangular matrix with real positive elements on the main diagonal. Substituting (24) into (23) produces

$$J = \mathrm{tr}\left\{\mathbf{C}^H \mathbf{\Lambda}^H \left(\mathbf{L}^H\right)^{-1} \mathbf{L}^{-1} \mathbf{\Lambda}\mathbf{C}\right\} \tag{25}$$

or, equivalently,

$$J = \mathrm{tr}\left\{\left(\mathbf{L}^{-1}\mathbf{\Lambda}\mathbf{C}\right)^H \mathbf{L}^{-1}\mathbf{\Lambda}\mathbf{C}\right\}. \tag{26}$$

Next, we observe that $\mathbf{L}^{-1}\mathbf{\Lambda}\mathbf{C}$ is still a lower triangular matrix and, in consequence, equation (26) can be rewritten as

$$J = \sum_{k=1}^{K} \frac{\lambda_k^2}{[\mathbf{L}]_{k,k}^2} + \sum_{k=2}^{K}\sum_{\ell=1}^{k-1} \left| \left[\mathbf{L}^{-1}\mathbf{\Lambda}\mathbf{C}\right]_{k,\ell} \right|^2 \tag{27}$$

where we have born in mind that $\mathbf{C}$ is a unit-diagonal and lower triangular matrix, i.e., $[\mathbf{C}]_{k,k}$=1 for $k = 1, 2, \ldots, K$.

From (27) it follows that the minimum of J is achieved when $\mathbf{L}^{-1}\mathbf{\Lambda}\mathbf{C}$ is diagonal, i.e.,

$$\mathbf{L}^{-1}\mathbf{\Lambda}\mathbf{C} = \mathbf{D}. \tag{28}$$

Finally, premultiplying both sides of (28) by $\mathbf{\Lambda}^{-1}\mathbf{L}$ produces the result (16) in the text.

REFERENCES

[1] G. J. Foschini and M. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless Personal Commun.*, vol. 6, no. 3, pp. 311–335, Mar. 1998.

[2] Q. Spencer, C. Peel, A. Swindlehurst, and M. Haardt, "An introduction to the multi-user MIMO downlink," *IEEE Communications Magazine*, vol. 42, no. 10, pp. 60–67, Oct. 2004.

[3] P. Wolniansky, G. Foschini, G. Golden, and R. Valenzuela, "V-BLAST: an architecture for realizing very high data rates over the rich-scattering wireless channel," in *Proc. of Int. Symposium on Signals, Systems, and Electronics (ISSSE'98)*, 29 Sept.-2 Oct 1998, pp. 295 – 300.

[4] R. Hunger, M. Joham, and W. Utschick, "Extension of linear and non-linear transmit filters for decentralized receivers," in *Proc. of European Wireless Conference*, vol. 1, April 2005, pp. 40–46.

[5] K. Kusume, M. Joham, W. Utschick, and G. Bauch, "Efficient Tomlinson-Harashima precoding for spatial multiplexing on flat MIMO channel," in *Proc. of IEEE Int. Conf. on Communications (ICC'05)*, Seoul, Korea, 16-20 May 2005.

[6] C. Windpassinger, R. F. H. Fisher, T. Vencel, and J. B. Huber, "Precoding in multiantenna and multiuser communications," *IEEE Trans. on Wireless Communications*, vol. 3, no. 4, pp. 1305–1316, July 2004.

[7] I. Cosovic, S. Sand, and R. Raulefs, "A non-linear precoding technique for downlink MC-CDMA," in *Proc. of IEEE Vehicular Technology Conference (VTC'05, Spring)*, Stockholm, Sweden, 28 May- 1 June 2005.

[8] J. Liu and W. A. Krzymien, "A novel non-linear precoding algorithm for the downlink of multiple antenna multi-user systems," in *Proc. of IEEE Vehicular Technology Conference (VTC'05, Spring)*, Stockholm, Sweden, 28 May- 1 June 2005.

[9] R. F. H. Fisher, *Precoding and Signal Shaping for Digital Transmission*. Wiley-IEEE Press, July 2002.

[10] S. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice-Hall, 1993.