TOMLINSON HARASHIMA PRECODING FOR MIMO SYSTEMS WITH LOW RESOLUTION D/A-CONVERTERS

Amine Mezghani, Rafik Ghiat and Josef A. Nossek

Institute for Circuit Theory and Signal Processing
Technische Universität München, 80290 Munich, Germany
{mezghani,ghiat,nossek}@nws.ei.tum.de

ABSTRACT

We study the joint transmitter optimization for the flat multi-input multi-output (MIMO) channel under nonlinear distortion from the digital-to-analog converters (DACs). Our design is based on a minimum mean square error (MMSE) approach, taking into account the effects of the transmitter nonlinearities. The derivation does not make use of the assumption of uncorrelated white distortion (quantization) errors and considers the correlations of the quantization error with the other signals of the system. Through simulation, we compare the new optimized Tomlinson Harashima precoder to previously proposed transmitter designs when operating under DACs in terms of uncoded BER.

1. INTRODUCTION

The use of multiple antennas at both sides of the transmission link (MIMO systems) can improve the communication performance dramatically. If the channel state information is available at the transmitter and the receiver(s) have to be of low complexity, then transmit processing becomes advantageous. An other application of transmit processing is the broadcast scenario where no cooperation is possible between the receivers. The transmit filter will be then designed based on the knowledge of the a priori defined receiver (simple scalar receiver) and the Channel State Information (CSI). It is well known that nonlinear processing strategies can clearly improve the performance. In [1, 2] transmit filters combined with Tomlinson Harashima Precoding (THP) [3] were derived using a minimum mean square error (MMSE or Wiener) approach, where the receiver is restricted to be a simple common weight. However, most of these contributions on transmitter design for MIMO systems assume that the transmitted signals do not experience any nonlinear distortion. In practice, however, the signal undergoes many nonlinearities (e.g., Digital-to-Analog Conversion (DAC), Power Amplifier,...). For instance, the D/A-Conversion can be also regarded as a kind of quantization since it requires rounding the data to a suitable level of precision to be converted into the analog domain. The effect of the digital-to-analog conversion on the transmitter designs has been neglected by the research community up to now. This is due to the fact that the DAC is a non-linear operation which complicates theoretical analysis. In high speed applications, the DACs have to be of low resolution in order to save power and area. In this case the proposed design do not perform well when operating under coarse quantization. In fact, in order to reduce circuit complexity and save power and area, low resolution DACs have to be employed [4]. Therefore, the proposed designs do not necessarily have good performance when operating under DACs in a real system. Motivated by the same approach as in our recent works [5, 6], which concerns the linear and nonlinear MMSE receivers operating on quantized data under A/D conversion at the receiver side, we modify the Wiener (or MMSE) THP transmitter from [2] for the transmit-quantized flat MIMO channel (later denoted by WFQ-THP), taking into account the presence of the DACs. Under the choice of an optimally designed DAC we evaluate the resulting MSE between the estimated and the transmitted symbols and we minimize it subject to a Wiener THP transmitter and a common scalar receiver. In our model we assume perfect channel state information (CSI) at the transmitter.

2. SYSTEM MODEL

We consider a MIMO Gaussian channel where the transmitter employs $N$ antennas and the receiver has $B$ antennas (or users). Fig. 1 shows the general form of a THP quantized MIMO system, where $\mathbf{H} \in \mathbb{C}^{B \times N}$ is the channel matrix. The vector $\mathbf{s} \in \mathbb{C}^{B}$ comprises the $M$ transmitted symbols, which are uncorrelated and have zero-mean and covariance matrix $\mathbf{R}_{ss} = \mathbf{E}[\mathbf{s}\mathbf{s}^\dagger] = \sigma_s^2 \mathbf{I}$. The vector $\mathbf{\eta}$ refers to zero-mean complex circular Gaussian noise with covariance $\mathbf{R}_{\eta \eta} = \sigma_{\eta}^2 \mathbf{I}$. Furthermore, $\mathbf{y} \in \mathbb{C}^{N}$ is the unquantized transmit signal.

In addition to the feedforward filter denoted by $\mathbf{P} \in \mathbb{C}^{M \times B}$ the transmitter is extended with a modulo device $M(\cdot)$ and a spatial feedback filter $\mathbf{F}$, where

$$\mathbf{P} = [p_1 \cdots p_B] \in \mathbb{C}^{N \times B}, \quad \text{and} \quad (1)$$

$$\mathbf{F} = \begin{bmatrix}
0 & 0 & \cdots & 0 \\
f_{2,1} & 0 & \cdots & 0 \\
\vdots & \ddots & \ddots & \vdots \\
f_{B,1} & \cdots & f_{B,B-1} & 0
\end{bmatrix} \in \mathbb{C}^{B \times B}. \quad (2)$$
In other words, \( p_k \) is the \( k \)-th column of \( P \) and \( f_{k,j} \) is the entry at the \( k \)-th row and \( j \)-th column \((k > j)\) of the feedback matrix \( F \).

The data streams (or users) are successively encoded with order \( \pi \), i.e. stream \( \pi_k \) (with order \( k \)) sees the interference caused by the data streams (or users) \( \pi_{k+1} \ldots \pi_B \). Then we construct the unquantized transmit signal
\[
y = Pu,
\]
where the outputs of the modulo devices are computed successively as
\[
u_k = (s_{\pi_k} - \sum_{j<k} f_{k,j}u_j) \mod(\tau).
\]

Through the DACs, the real parts \( y_{i,R} \) and the imaginary parts \( y_{i,I} \) of the transmit signals \( y_i, 1 \leq i \leq N \), are each quantized by a \( b \)-bit resolution scalar quantizer. Thus, the resulting analog signals read as
\[
r_{i,l} = Q(y_{i,l}) = y_{i,l} + q_{i,l}, \quad l \in \{R,I\}, \quad 1 \leq i \leq N,
\]
where \( Q(\cdot) \) denotes the quantization operation and \( q_{i,l} \) is the resulting quantization error. In the transmit processing architecture of Fig. 1, the receiver is restricted to a simple signal scaling with a common weight \( g \). The matrices \( P, F \) and the weight \( g \), together, delivers each estimate \( \hat{s}_{\pi_k} \) as
\[
\hat{s}_{\pi_k} = (gh_{\pi_k}^T r + \eta_{\pi_k}) \mod(\tau),
\]
with \( r = Q(y) = y + q \),

\[  \hat{s}_{\pi_k} = (gh_{\pi_k}^T r + \eta_{\pi_k}) \mod(\tau), \]

where \( h_{\pi_k} \) is the \( \pi_k \)-th row vector of the channel matrix \( H \). Our aim is to choose the DAC stepsizes, the matrix \( P \), and the scalar \( g \), minimizing the MSE \( E[\|s - \hat{s}\|_2^2] \), taking into account the quantization effect. Throughout this paper, \( r_{\alpha\beta} \) denotes \( E[\alpha\beta] \).

The operators \((\cdot)^T\), \((\cdot)^H\), \((\cdot)^*\), \text{Re}(\cdot), \text{Im}(\cdot)\) stand for transpose, Hermitian transpose, complex conjugate, real and imaginary parts of a complex number, respectively; \( e_k \) is the \( k \)-th column of the \( B \times B \) identity matrix \( 1_B \).

**Fig. 1. MIMO-Quantizer THP Transmit System**

### 3. REVIEW OF THE SPATIAL MMSE-THP

In this section, we review the MMSE-THP (or Wiener Filter-THP) algorithm [2], while ignoring the quantization (i.e. \( r \equiv y \)). Neglecting the modulo-loss, the error signal at the \( k \)-th encoding step can be approximated by
\[
e_{\pi_k} = -u_k + gh_{\pi_k}^T y - \sum_{j<k} f_{k,j}u_j + g\eta_{\pi_k}.
\]

Then, under the assumption that the outputs of the modulo devices \( u_k \) are i.i.d. with variance \( \sigma_u^2 \), it is easy to show by the KKT conditions that, for given ordering \( \pi \), the optimal precoding and feedback matrices minimizing the sum-MSE \( \sum_k E[|e_k|^2] \) read as
\[
f_{k,j} = gh_{\pi_k}^T p_j, \quad p_k = \frac{1}{g} \left( \sum_{j \leq k} h_{\pi_j}^* h_{\pi_j} + \xi 1_N \right)^{-1} h_{\pi_k}^*,
\]
where \( \xi = \frac{\tau E[|\eta|^2]}{E[u]} \). With these optimum matrices, the sum mean square error reads as
\[
\text{MSE} = \sigma_u^2 \left[ B - \sum_k h_{\pi_k}^T \left( \sum_{j \leq k} h_{\pi_j}^* h_{\pi_j}^T + \xi 1_N \right)^{-1} h_{\pi_k}^* \right].
\]

Now, the problem of finding the optimal ordering \( \pi \) minimizing the sum-MSE is NP hard, since we must check all \( B! \) possible permutations. To reduce the complexity, we minimize each summand in (10) successively, i.e., \( \pi_k \) is chosen under the assumption that \( \pi_{k+1}, \ldots, \pi_B \) are fixed
\[
\pi_k = \arg\max_{i \in \{\pi_{k+1}, \ldots, \pi_B\}} h_i^T \left( \sum_{j \notin \{\pi_{k+1}, \ldots, \pi_B\}} h_j^* h_j^T + \xi 1_N \right)^{-1} h_i^*.
\]

### 4. DAC OPTIMIZATION AND CHARACTERIZATION

D/A-conversion can be viewed as a quantization process since the resolution of the values of \( y \) has to be reduced to a smaller one, let say \( b \) bits, to be transformed into the analog domain, which is inevitable for high speed transmission. This operation can be characterized by a distortion factor \( \rho_q^{(i,l)} \) to indicate the relative amount of quantization error generated, which is defined as follows
\[
\rho_q^{(i,l)} = \frac{E[|y_{i,l}|^2]}{E[|y_{i,l}|^2]},
\]
where \( r_{y_{i,l}y_{i,l}} = E[|y_{i,l}|^2] \) is the variance of \( y_{i,l} \) and the distortion factor \( \rho_q^{(i,l)} \) depends on the number of quantization bits \( b \), the quantizer type (uniform or non-uniform) and the probability density function of \( y_{i,l} \). Note that the signal-to-quantization noise ratio (SQNR) has an inverse relationship with regard to the distortion factor
\[
\text{SQNR}^{(i,l)} = \frac{1}{\rho_q^{(i,l)}}.
\]
Similar to our work [5], the quantizer design is based on minimizing the mean square error (distortion) between the input \( y_{i,t} \) and the output \( r_{i,t} \) of each quantizer. In other words, the SQNR values are maximized. Under this optimal design of the scalar finite resolution quantizer, whether uniform or not, the following equations hold for all \( 0 \leq i \leq N \), \( l \in \{ R, I \} \) [7, 8]

\[
E[y_{i,t}] = 0 \quad (14)
\]

\[
E[r_{i,t} | y_{i,t}] = 0 \quad (15)
\]

\[
E[y_{i,t} | r_{i,t}] = -\rho_q^{(i,l)} r_{y_{i,t},y_{i,t}} \quad (16)
\]

Obviously, (16) follows from (12) and (15). For the uniform quantizer case, (14) holds only if the probability density function of \( y_{i,t} \) is even.

Assuming a dense matrix \( P \) and for large number of antennas, the DAC input signals \( y_{i,t} \) are weighted sums of weakly dependent random symbols and thus, due to central limit theorem, they are approximately Gaussian distributed. Besides, they undergo nearly the same distortion factor \( \rho_q \), i.e., \( \rho_q^{(i,l)} = \rho_q \) \( \forall i \forall l \). Furthermore, the optimal parameters of the uniform as well as the non-uniform quantizer and the resulting distortion factor \( \rho_q \) for Gaussian distributed signal are tabulated in [7] for different bit resolutions \( b \). Recent research work on optimally quantizing the Gaussian source can be found in [9, 10].

Now, let \( q_i = q_{i,R} + j q_{i,I} \) be the complex quantization error. Under the assumption of uncorrelated real and imaginary part of \( y_{i,t} \), we easily obtain

\[
r_{q_i,q_i} = E[q_i q_i^*] = \rho_q r_{y_{i,t},y_{i,t}} \quad \text{and} \quad r_{y_{i,t},q_i} = E[y_{i,t} q_i^*] = -\rho_q r_{y_{i,t},y_{i,t}}.
\]

(17)

For the uniform quantizer case, it was shown in [10], that the optimal quantization step \( \Delta \) for a Gaussian source decreases as \( \sqrt{2}^{-b} \) and that \( \rho_q \) is asymptotically well approximated by \( \frac{\pi}{12} \) and decreases as \( b 2^{-2b} \). On the other hand, the optimal non-uniform quantizer achieves, under high-resolution assumption, approximately the following distortion [11]

\[
\rho_q \approx \frac{\pi \sqrt{3}}{2} 2^{-2b}.
\]

(18)

This particular choice of the (non)-uniform scalar quantizer minimizing the distortion between \( r \) and \( y \) for the DACs, combined with the transmitter of the next section, is also optimal with respect to the total MSE between the transmitted symbol vector \( s \) and the estimated symbol vector \( \hat{s} \), as we will see later.

## 5. NEARLY OPTIMAL QUANTIZED TBP-TRANSMITTER

In this section, we optimize the transmit matrix \( P \) and the scalar receiver \( g \) based on the MMSE criterion, taking into account the quantization process. To this end, we minimizes the sum-MSE under the power constraint \( E \left[ \| r \|^2 \right] \leq E_{tr} \)

\[
\min_{\{ P, F, g \}} \text{MSE}(P, F, g) \quad \text{s.t.:} \quad E \left[ \| r \|^2 \right] \leq E_{tr},
\]

with

\[
E \left[ \| r \|^2 \right] = tr(\mathbf{R}_{rr}).
\]

For this optimization we establish the Lagrangian function to find the global minimum

\[
L(P, g, \lambda) = \text{MSE}(P, g) + \lambda \left( tr(\mathbf{R}_{rr}) - E_{tr} \right) \quad \text{with} \quad \lambda \in \mathbb{R}^+. \quad (20)
\]

As the \( u_k \)s are results of modulo-operations, they are nearly independent and have approximately a uniform distribution in the complex square with side \( \tau \). Thus we assume that the power of \( u_k \) is \( \sigma^2 = \frac{\tau^2}{2} \). Let us first consider the received symbol \( \hat{s}_{\pi_k} \) of the \( \pi_k \)-th user

\[
\hat{s}_{\pi_k} = (g h_{\pi_k}^T r + g n_{\pi_k}) \text{mod}(\tau).
\]

After adding and subtracting the term \( \sum_{j<k} f_{k,j} u_j \) we obtain\(^1\)

\[
\hat{s}_{\pi_k} = (\sum_{j>k} f_{k,j} u_j + g h_{\pi_k}^T r - \sum_{j<k} f_{k,j} u_j + g n_{\pi_k}) \text{mod}(\tau)
\]

\[
= (s_{\pi_k} - u_k + g h_{\pi_k}^T r - \sum_{j<k} f_{k,j} u_j + g n_{\pi_k}) \text{mod}(\tau).
\]

When we neglect the modulo-loss, the error \( \varepsilon_{\pi_k} = \hat{s}_{\pi_k} - s_{\pi_k} \) is nearly given by

\[
\varepsilon_{\pi_k} = -u_k + g h_{\pi_k}^T r - \sum_{j<k} f_{k,j} u_j + g n_{\pi_k}.
\]

(21)

In order to compute each individual mean square error \( \varepsilon_{\pi_k} = E[\varepsilon_{\pi_k}^2] \pi_k \), we assume that the \( u_k \) have the variance \( \sigma^2 \) and are mutually uncorrelated. Furthermore, we consider the correlation of the quantization error with the other signal of the system by means of (21)

\[
\varepsilon_{\pi_k} = |g|^2 h_{\pi_k}^T R_{ee} h_{\pi_k} - 2 \text{Re}(g h_{\pi_k}^T R_{eu} e_k) \quad -2 \text{Re}(g h_{\pi_k}^T R_{eu} \sum_{j<k} e_j f_{k,j}^* + \sigma_n^2 (1 + \sum_{j<k} |f_{k,j}|^2) |g|^2 \sigma_n^2).
\]

(22)

with the correlation matrix

\[
R_{ee} = E[ru^H] = E[(y + q)u^H] = R_{yu} + R_{qy}, \quad (23)
\]

and \( R_{rr} \) the covariance matrix of the quantized signal given by

\[
R_{rr} = E[(y + q)(y + q)^H] = R_{yy} + R_{qy} + R_{qy}^H + R_{qq}. \quad (24)
\]

## 5.1. Derivation of the Covariance Matrices involving the Quantization Error

Before investigating the MMSE optimization, we first derive all needed covariance matrices by using the fact that the quantization error \( q_i \), conditioned on \( y_i \), is statistically independent from\(^1\)

\[
\text{Note that} \quad (s_{\pi_k} - u_k) \text{mod}(\tau) = (\sum_{j<k} f_{k,j} u_j) \text{mod}(\tau) \text{ and } (a + b) \text{mod} \tau = (a \text{mod} \tau + b \text{mod} \tau) \text{mod} \tau.
\]
all other random variables of the system.

First we calculate 
\[ E[y_i q_j^*] \] for \( i \neq j \)

\[
E[y_i q_j^*] = E[y_i q_j^* | y_j] = E[y_i q_j^* | y_j] \approx \frac{1}{p_q} \sum_{j \leq k} h^*_\pi_j h^T \pi_j + \rho_q \text{diag}(H\text{HH}) + \xi_{1N}^\prime
\]

(25)

(26)

In (25), we approximate the Bayesian estimator \( E[y_i | y_j] \) with the linear estimator \( r_{y_i y_j} r_{y_j y_j}^{-1} y_j \), which holds with equality if the vector \( y \) is jointly Gaussian distributed. Eq. (26) follows from (17). Summarizing the results of (17) and (26), we obtain

\[
R_{yq} \approx -\rho_q R_{yy}.
\]

Similarly, we evaluate \( r_{q_i q_j} \) for \( i \neq j \) to end up in

\[
E[q_i q_j^*] = E[q_j q_i^* | y_j] \approx \rho_q r_{y_i y_j} = \rho_q r_{y_i y_j},
\]

(28)

where we used (27) and (17). From (28) and (17) we deduce the covariance matrix of the quantization error

\[
R_{qq} \approx \rho_q \text{diag}(R_{yy}) + \rho_q^2 \text{nondiag}(R_{yy}) = \rho_q R_{yy} - (1 - \rho_q) \rho_q \text{nondiag}(R_{yy}),
\]

(29)

with \( \text{diag}(A) \) denotes a diagonal matrix containing all the diagonal elements of \( A \) and \( \text{nondiag}(A) = A - \text{diag}(A) \). Inserting the expressions (27) and (29) into (24), we obtain

\[
R_{rr} \approx (1 - \rho_q)(R_{yy} - \rho_q \text{nondiag}(R_{yy})).
\]

(30)

In a very similar way, we get the covariance matrix \( R_{yu} = E[qu^H] \) as

\[
R_{yu} = E[qu^H] \approx -\rho_q R_{yu}.
\]

(31)

Thus, Equation (23) becomes

\[
R_{ru} \approx (1 - \rho_q) R_{yu}.
\]

(32)

Finally, \( R_{yy} \) and \( R_{yu} \) can be easily obtained from our system model

\[
R_{yy} = \sigma_u^2 P P^H\text{,}
\]

(33)

\[
R_{yu} = \sigma_u^2 P.
\]

(34)

5.2. Derivation of the TxFWQ transmitter

Now, we return to our MMSE problem and insert these results into (22) to obtain

\[
\varepsilon_{\pi_k} = \frac{\alpha q^2}{\alpha^2} \left[ \alpha g^2 h^T_{\pi_k} (\alpha P P^H + \rho_q \text{diag}(P P^H))(h^T_{\pi_k}) - 2 \alpha \text{Re}(g h^T_{\pi_k} p_k) - 2 \alpha \text{Re}(g h^T_{\pi_k} \sum_{j < k} p_j f_{jk}^*) + 1 + \sum_{j < k} |f_{jk}|^2 \right] + |g|^2 \sigma_u^2
\]

(35)

with \( \alpha = 1 - \rho_q \). Afterwards, we establish the Lagrangian function for the optimization (19) with the Lagrangian multiplier \( \lambda \in \mathbb{R}^+ \)

\[
L(p_{\pi_k}, g, f_{jk}) = \sum_{j=1}^{B} \varepsilon_{\pi_k} + \lambda \left( \frac{\sigma_u^2 \text{tr}(PP^H)}{\varepsilon_{\pi_k}} - \frac{E_{tr}}{(1 - \rho_q)} \right).
\]

(36)

When differentiating this lagrangian with respect to \( f_{jk} \), we obtain

\[
\frac{\partial}{\partial f_{jk}} L(p_{\pi_k}, g, f_{jk}) = \alpha q^2 (\alpha \sum_{j \geq k} p_j p_j^H + \rho_q \text{diag}(PP^H)) h^T_{\pi_k} - 2 \alpha \text{Re}(g h^T_{\pi_k} p_k) + 1 + |g|^2 \sigma_u^2
\]

(37)

which must be equal to zero in order to optimize the MSE. Thus we obtain the optimal feedback matrix

\[
f_{jk} = (1 - \rho_q) g h^T_{\pi_k} p_j.
\]

(38)

This delivers the following MSE

\[
\varepsilon_{\pi_k} = \frac{\alpha q^2}{\alpha^2} \left[ \alpha g^2 h^T_{\pi_k} (\alpha \sum_{j \geq k} p_j p_j^H + \rho_q \text{diag}(PP^H)) h^T_{\pi_k} - 2 \alpha \text{Re}(g h^T_{\pi_k} p_k) + 1 + |g|^2 \sigma_u^2
\]

(39)

Then, we formulate the KKT equations with respect to \( g \) and \( P \)

\[
\frac{\partial L(P, g, \lambda)}{\partial g} = \frac{\alpha^2}{\alpha^2} \sum_{k} \alpha \left[ g h^T_{\pi_k} (\alpha \sum_{j \geq k} p_j p_j^H) + \rho_q \text{diag}(PP^H) h^T_{\pi_k} - h^T_{\pi_k} p_k \right] + |g|^2 \sigma_u^2 B \mathbf{1} = 0,
\]

(40)

\[
\frac{\partial L(P, g, \lambda)}{\partial p_k} = \frac{\alpha^2}{\alpha^2} \sum_{k} \left[ |g|^2 (\alpha \sum_{j \leq k} h_{\pi_j} h^H_{\pi_j} + \rho_q \text{diag}(H^T H^*)) p_k \right] - g h_{\pi_k} + \lambda \sigma_u^2 p_k^T = 0,
\]

(41)

where we used the following identity for two matrices \( A, B \)

\[
\text{tr}(A \text{ diag}(B)) = \text{tr}(B \text{ diag}(A)).
\]

(42)

Summing the complex conjugate of the derivative with respect to \( p_k \) (41), then multiplying it by \( P^H \) from the right followed by the trace operation, and finally comparing it with (40) multiplied by \( g \), we can show the following relation

\[
\lambda = |g|^2 \frac{B \sigma_u^2}{\alpha^2 \text{tr}(PP^H)} = |g|^2 \frac{\text{tr}(R_{yy})}{E_{tr}} = |g|^2 \xi.
\]

(43)
Note that the inverse always exists independently of $H$. Due to the transmit energy constraint we finally get

$$g = \sqrt{\frac{\sigma_p^2 (1-\rho_q)}{E_x}},$$

$$\sqrt{\sum_k h_k^T (1-\rho_q) \sum_{j\leq k} h_j^* h_j^T + \rho_q \text{diag}(H^H H) + \xi 1_N}^{-1} h_k^*, \quad (44)$$

Notice that, when we set $\rho_q = 0$ (infinite precision DACs) in the expressions of $P$ and $g$, we obtain exactly the same expressions as derived in (9) for the unquantized system. Using (39), (43) and (44), the sum-MSE resulting from the optimal design of the receiver filter and feedback matrix becomes after some computations

$$\text{MSE} = \sum_k \varepsilon_k = \sigma_n^2 \left[ B - (1-\rho_q) \sum_k h_k^T \left( (1-\rho_q) \sum_{j\leq k} h_j^* h_j^T + \rho_q \text{diag}(H^H H) + \xi 1_N \right) \right]^{-1} h_k^T.$$  \quad (45)

Now, the problem of finding the optimal ordering $\pi$ minimizing the sum-MSE (45) is NP hard, since we must check all $B!$ possible permutations. To reduce the complexity, we minimize each summand of the sum-MSE successively, i.e., $\pi_k$ is chosen under the assumption that $\pi_B, \ldots, \pi_{k+1}$ are fixed

$$\pi_k = \arg \max_{i \neq \pi_{k+1}, \ldots, \pi_B} h_i^T (1-\rho_q) \sum_{j \neq \pi_{k+1}, \ldots, \pi_B} h_j^* h_j^T + \rho_q \text{diag}(H^H H) + \xi 1_N \right]^{-1} h_i^T. \quad (46)$$

### 5.3. Effects of Quantization on the MSE

In order to verify whether the chosen quantizer minimizes the MSE of our system, we examine the first derivative of the MSE in (45) with respect to $\rho_q$

$$\frac{\partial \text{MSE}}{\partial \rho_q} = g^2 \sigma_n^2 \sum_k p_k^U (\text{diag}(H^H H) + \frac{\text{tr}(R_{yy})}{E_x} 1_N) p_k,$$

where we used (45). Obviously, the derivative of the MSE with respect to the distortion is positive. Therefore, the $\text{MSE}_{\text{WFQ-THP}}$ achieved by THP transmit processing is monotonically increasing with respect to $\rho_q$. Since we chose the quantizer to minimize the distortion factor $\rho_q$, our receiver quantizer designs are jointly optimum with respect to the total MSE.

Now, we expand the MSE expression (45) into a Taylor series around $\rho_q = 0$ up to the order one, to get an approximation of the MSE

$$\text{MSE}_{\text{WFQ-THP}} \approx \text{MSE}_{\text{WF-THP}} + \rho_q g^2 \sigma_n^2 E_x (\text{tr}(P^H \text{diag}(H^H H) + \frac{\text{tr}(R_{yy})}{E_x} 1_N) P), \quad (47)$$

where $\text{MSE}_{\text{WF-THP}} = \text{MSE}_{\text{WFQ-THP}}|_{\rho_q=0}$ is the achievable MSE without quantization given in (10). The second term gives the increase in the MSE due to the quantization as a function of $\rho_q$ and the channel parameters. It reveals also the residual error at infinite SNR.

### 6. SIMULATION RESULTS

The performance of the modified THP transmitter for a 4-bit quantized output MIMO system (WFQ-THP), in terms of BER averaged over $10^6$ channel realizations, is shown in Fig. 2 for a $10 \times 10$ MIMO system (QPSK), compared with the conventional Tomlinson Harashima Precoder (WF-THP) from [2]. The symbols and the noise are assumed to be uncorrelated, that is $R_{xx} = \sigma_n^2 I$ and $R_{yy} = \sigma_s^2 I$. Besides, the BER curve of the modified linear MMSE transmitter from [12] is shown. Thereby, the entries of $H$ are complex-valued realization of independent zero-mean Gaussian random variables with unit variance. Clearly, the modified transmitter outperforms the conventional ones at high SNR. This is because the effect of quantization error is more pronounced at higher SNR values when compared to the additive Gaussian noise variance. Since the conventional THP-precoder looses its regularized structure at high SNR values, its performance degrades asymptotically, when operating under DACs. For comparison, we also plotted the BER curve for the THP-precoder transmitter, if no quantization is applied (infinite precision DAC).

![Fig. 2. WFQ-THP vs. WFQ. WF-THP with quantization and WF-THP without quantization; 4QAM modulation with $B = 10$, $N = 10$, 4-bit ($\rho_q = 0.01154$) uniform quantizer.](image-url)
it shows the simulated BER with a THP transmitter considering
the quantization error $q$ as an additive white noise, which is un-
correlated with the other signals of the system. For the additive
quantization noise model, we take as effective noise covari-
ance matrix (see first equality of (17))

$$\bar{R}_{\eta\eta} = R_{\eta\eta} + H\bar{R}_{qq}H^H = R_{\eta\eta} + \rho_q^2\sigma_n^2\text{diag}(PP^H)H^H.$$  

(48)

Thus the well-known solution (9) can be adjusted to the follow-
ing effective noise power

$$\text{tr}(\bar{R}_{\eta\eta}) \approx B\sigma_n^2 + \frac{\rho_q}{1 - \rho_q} E_{\tr}(HH^H).$$  

(49)

Obviously, this model, which has been commonly used in the
literature is also outperformed by the presented design at any
SNR level and independently of the resolution.

![THP proposed vs. the conventional THP](image)

Fig. 3. THP proposed vs. the conventional THP [2] adjusted
to an additive quantization noise term; 4QAM modulation with
$B = 10, N = 10, 3$-bit ($\rho_q = 0.03744$) uniform quantizer.

7. CONCLUSION

We addressed the problem of designing a THP transmitter for
MIMO channels under DAC nonlinearity. We provided an ap-
proximation for the mean squared error for each data stream,
where the quantizer is optimized for a Gaussian input. Then,
we proposed an optimized THP transmitter in terms of the MSE
criteria that takes into account the quantization effects. The pro-
posed precoder operating for quantized transmit signals shows
better performance in terms of BER than simply applying the
WF-THP (as well-known in literature [1, 2]) designed for an
unquantized MIMO system. It also outperforms the most obvi-
ous approach of simply adjusting this WF-THP to an effective
noise power of $B\sigma_n^2 + \frac{\rho_q}{1 - \rho_q} E_{\tr}(HH^H)$. An essential aspect
of our derivation is that we do not make the assumption of un-
correlated white quantization error. Moreover, our transmitter
does not present any extra complexity from the implementa-
tion point of view.

8. REFERENCES

[1] R. Fischer, C. Windpassinger, A. Lampe, and J. Huber,
“Space-time transmission using Tomlinson-Harashima
precoding,” in 4th ITG Conference on Source and Channel
Coding, Jan. 2002.

“Finite-length MMSE Tomlinson-Harashima precoding
for frequency selective vector channels,” IEEE Trans. on

ulo arithmetic,” Electronic Letters, vol. 7, no. 5/6, pp. 138–
139, March 1971.


MIMO Decision Feedback equalizer Operating on Quan-
tized Data,” Submitted to ICASSP, 2008.


Bit Precision and Clip Level Computation for a Gaussian
Signal,” IEEE Trans. on Sig. Proc., vol. 44, no. 2, February
1996.

mal Fixed-Rate Uniform Scalar Quantization,” IEEE

Compression, Kluwer Academic Publishers, Dordrecht,

[12] R. Ghait, A. Mezghani, and J. A. Nossek, “Transmit Pro-
cessing with Low Resolution D/A-Converters,” in prepara-
tion, 2009.