Joint Optimization of Transceivers with Decision Feedback and Bit Loading

Ching-Chih Weng, Chun-Yang Chen and P. P. Vaidyanathan
Dept. of Electrical Engineering, MC 136-93
California Institute of Technology, Pasadena, CA 91125, USA
E-mail: cweng@caltech.edu, cyc@caltech.edu, and ppvnath@systems.caltech.edu

***Abstract*— The transceiver optimization problem for MIMO channels has been considered in the past with linear receivers as well as with decision feedback (DFE) receivers. Joint optimization of bit allocation, precoder, and equalizer has in the past been considered only for the linear transceiver (transceiver with linear precoder and linear equalizer). It has also been observed that the use of DFE even without bit allocation in general results in better performance that linear transceivers with bit allocation. This paper provides a general study of this for transceivers with the zero-forcing constraint. It is formally shown that when the bit allocation, precoder, and equalizer are jointly optimized, linear transceivers and transceivers with DFE have identical performance in the sense that transmitted power is identical for a given bit rate and error probability. The developments of this paper are based on the generalized triangular decomposition (GTD) recently introduced by Jiang, Li, and Hager. It will be shown that a broad class of GTD-based systems solve the optimal DFE problem with bit allocation. The special case of a linear transceiver with optimum bit allocation will emerge as one of the many solutions.** [1]

***Index Terms* — Decision Feed-Back, BER Optimization, Generalized Triangular Decomposition, Bit Allocation, MIMO Transceiver.**

I. Introduction

In this paper we consider the optimization of the multiple-input multiple-output (MIMO) communication systems with perfect channel state information (CSI) at both sides of the link. The focus of this paper will be on the system with decision feedback equalizer and linear precoding technique. The designing method of the system with DFE and linear precoding is considered by many authors when the bit constellations are fixed and identical for each sub-stream [17] [9] [10] [12] [18] [15] [11]. Similarly, when the channel and DFE are given, the bit loading scheme is a well treated problem [4]. However, to the best of the authors' knowledge, the joint optimization of transceivers with decision feedback and bit loading has not been reported before. The main goal of this paper is to provide the theoretical background for this problem.

For the linear transceiver case (which is a special case of the system with DFE and linear precoding), several researchers considered the joint optimization of bit loading and the precoder/equalizer design. In [13] the authors considered the joint design of the constellations and linear transceivers under the zero-forcing condition. The multiuser quality of service (QoS) problem, which is an extension of the problem in [13], is treated in [2]. In [3] the authors considered the bit loading scheme and the linear transceiver design with no zero-forcing constraint. It is shown in all of those papers that the diagonalized structure is optimal under the assumption that the bit allocation formula is realizable.

It has also been observed that the use of DFE without bit loading in general results in better performance than linear transceivers [8], [9]. In [8] the authors showed that the design based on the GMD (geometric mean decomposition) is asymptotically optimal for high SNR in terms of both channel throughput and bit error rate. In other words, GMD systems can achieve the optimal performance when the number bits assigned to each sub-stream are identical.

In this paper we consider the problem of minimizing the total power when the error probabilities of the substreams and the total bit rate are fixed. It is formally shown that when the bit allocation, precoder, and receiver matrices are jointly optimized, linear transceivers and transceivers with DFE have identical performance (i.e., minimum power), assuming of course that the bit loading formula is realizable. We then show that the optimal system can be designed by representing the channel in terms of the generalized triangular decomposition (GTD) proposed recently by Jiang et. al [10], and choosing the transceiver matrices appropriately in terms of the GTD. While the GTD representation of the channel is not unique, it is shown that the design based on any GTD is optimal, with bit allocation appropriately adjusted. The ZF-VBLAST system [1], the GMD system, and the SVD-based system [13] are some special cases. Some novel special cases such as the bi-diagonal representation and the Schur decomposition will also be mentioned. We will see that the flexibility offered by the GTD system is valuable. For example, it can often be exploited to make the bit allocation realizable (i.e., ensure that the optimal b_k are nonnegative integers).

This paper is structured as follows. In Section II, we will introduce the communication models and give explicit problem formulations. In Section III, we will prove that systems with DFE and linear precoding have the same performance as linear transceivers, if optimal bit loading formula is realizable. Section IV gives the transceiver structure based on the generalized triangular decomposition of the channel matrix, and proves that this kind of systems always achieves the optimal performance. Section V presents the numerical simulation results related to the topics discussed in the paper. The final conclusions of the paper are summarized in section VI.

II. Problem Formulations

The transceiver considered in this paper is shown in Fig. 1, with the sizes of matrices indicated (e.g., $\mathbf{F}$ is $P \times M$, etc.). The additive channel noise is assumed to have covariance $\sigma_n^2\mathbf{I}$. Here $\mathbf{F}$ is the linear precoder, $\mathbf{H}$ is the channel, $\mathbf{G}$ is the feedforward part of the equalizer, and $\mathbf{B}$ is the feedback part. The decision device processes the vector $\widehat{\mathbf{s}}$ bottom-up sequentially, and the past decisions within a block are fedback via $\mathbf{B}$ to correct future decisions in the block. This causality

[1] This work is supported in parts by the ONR grant N00014-06-1-0011, and the TMS scholarship 94-2-A-018 of the National Science Council of Republic of China, Taiwan.

of decision feedback is ensured by restricting $\mathbf{B}$ to be strictly upper triangular.

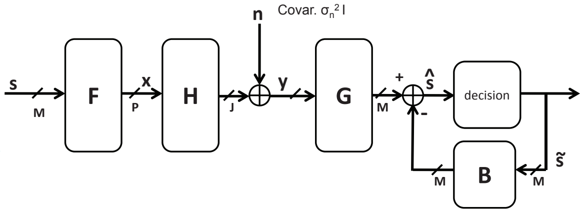

Fig. 1. The MIMO transceiver with linear precoder and DFE.

In the following sections, we will focus on the problem of minimizing the transmitted power subject to the total bit rate and error probabilities in each sub-stream [4]. To understand how the problems of bit allocation and power minimization arise, we first examine the relationships between the error probabilities, bit rates and user powers. Assume the input signals are zero-mean uncorrelated processes representing independent users with power P_k so that the input covariance is

$$\mathbf{\Lambda}_s = \text{diag}(P_1, P_2, \cdots, P_M). \tag{1}$$

Consider the situation where each user is represented with a different constellation size. Let us say the kth user transmits b_k-bit QAM symbols with average power P_k. If the error at the kth sub-stream has variance $\sigma^2_{e_k}$, based on the low error and high bit rate assumption, it can be shown [13] that when the probability of error is given, we have

$$\frac{P_k}{\sigma^2_{e_k}} \approx \frac{2^{b_k}}{3}\left(Q^{-1}\left(\frac{P_e(k)}{4}\right)\right)^2. \tag{2}$$

The total power transmitted can be written as

$$P_{trans} = \text{Tr}(\mathbf{F}\mathbf{\Lambda}_s\mathbf{F}^\dagger) = \text{Tr}(\mathbf{F}^\dagger\mathbf{\Lambda}_s\mathbf{F}) = \sum_{k=1}^{M} P_k[\mathbf{F}^\dagger\mathbf{F}]_{kk}.$$

Substituting from (2) we can rewrite this as

$$P_{trans} = \sum_{k=1}^{M} d_k 2^{b_k}\sigma^2_{e_k}[\mathbf{F}^\dagger\mathbf{F}]_{kk}, \tag{3}$$

where $d_k = \frac{1}{3}(Q^{-1}(\frac{P_e(k)}{4}))^2,$, which is determined by the specified probability of error.

It is usually assumed that the previous detected symbols $\widetilde{\mathrm{s}}$ in Fig. 1 are always correct. When we assume there is no error propagation, the zero forcing constraint can be written as

$$\mathbf{GHF} - \mathbf{B} = \mathbf{I} \tag{4}$$

This means that the interference from other transmitted symbols is canceled out completely. Under the zero-forcing constraint, the error before the decision device for each sub-stream entirely comes from the channel noise. Since the channel noise has covariance $\sigma^2_n\mathbf{I}$, the error variance before the kth input of the decision device is given by

$$\sigma^2_{e_k} = \sigma^2_n[\mathbf{G}\mathbf{G}^\dagger]_{kk} \tag{5}$$

From (3) the transmitted power can then be written as

$$P_{trans} = \sum_{k=1}^{M} c_k 2^{b_k}[\mathbf{F}^\dagger\mathbf{F}]_{kk}[\mathbf{G}\mathbf{G}^\dagger]_{kk}, \tag{6}$$

where $c_k = \sigma^2_n d_k = \frac{\sigma^2_n}{3}(Q^{-1}(\frac{P_e(k)}{4}))^2$. Therefore our first problem is, given the specified QoS(probability of error), i.e., c_k, how should we design the precoder, the equalizer, and the bit loading scheme to minimize the transmitted power. This problem can be written as follows:

$$\begin{aligned} \min_{\mathbf{F},\mathbf{G},\mathbf{B},\{b_k\}} \quad & P_{trans} = \sum_{k=1}^{M} c_k 2^{b_k}[\mathbf{F}^\dagger\mathbf{F}]_{kk}[\mathbf{G}\mathbf{G}^\dagger]_{kk} \\ \text{s.t.} \quad (a) \quad & \frac{1}{M}\sum_{k=1}^{M} b_k = b \\ (b) \quad & \mathbf{GHF} - \mathbf{B} = \mathbf{I} \end{aligned} \tag{7}$$

III. Optimal Bit-Loaded DFE Transceivers

To minimize the transmitted power, we first observe that

$$\begin{aligned} P_{trans} &= \sum_{k=1}^{M} c_k 2^{b_k}[\mathbf{F}^\dagger\mathbf{F}]_{kk}[\mathbf{G}\mathbf{G}^\dagger]_{kk} \\ &\geq c2^b(\prod_{k=1}^{M}[\mathbf{F}^\dagger\mathbf{F}]_{kk})^{\frac{1}{M}}(\prod_{k=1}^{M}[\mathbf{G}\mathbf{G}^\dagger]_{kk})^{\frac{1}{M}} \end{aligned}$$

where we have used the AM-GM inequality. Here $c = M(\prod_{k=1}^{M} c_k)^{\frac{1}{M}}$, and we have used the fact that

$$b = \frac{1}{M}\sum_{k=1}^{M} b_k. \tag{8}$$

Equality can be achieved in the AM-GM inequality if and only if the terms are identical for all k, that is, $c_k 2^{b_k}[\mathbf{F}^\dagger\mathbf{F}]_{kk}[\mathbf{G}\mathbf{G}^\dagger]_{kk} = A$ for some constant A. Taking logarithms on both sides we get

$$b_k = D - \log_2 c_k - \log_2[\mathbf{F}^\dagger\mathbf{F}]_{kk} - \log_2[\mathbf{G}\mathbf{G}^\dagger]_{kk} \tag{9}$$

where D is a constant, which is chosen such that (8) is satisfied. Eq. (9) is called the optimum bit loading formula.

For any fixed precoder $\mathbf{F}$ and receiver $\{\mathbf{G}, \mathbf{B}\}$, and specified probabilities of error $P_e(k)$, the bit allocation that minimizes the transmitted power is given by (9). With the bit allocation so chosen the quantities P_k are computed from (2) where $\sigma^2_{e_k}$ is as in (5). With P_k so chosen, the specified probabilities of error are met, and the total power P_{trans} is minimized. This minimized power is

$$P_{trans} = c2^b(\prod_{k=1}^{M}[\mathbf{F}^\dagger\mathbf{F}]_{kk})^{\frac{1}{M}}(\prod_{k=1}^{M}[\mathbf{G}\mathbf{G}^\dagger]_{kk})^{\frac{1}{M}}), \tag{10}$$

which depends only on $\mathbf{F}$ and $\mathbf{G}$.

In the following we show how to minimize (10) further subject to the zero-forcing constraint. First we derive the optimal feedforward filter $\mathbf{G}$ when the precoder $\mathbf{F}$ and the feedback filter $\mathbf{B}$ are given. The result is stated in the following lemma.

Lemma 1: When the precoder $\mathbf{F}$ and the feedback filter $\mathbf{B}$ are given, the optimal feed-forward filter $\mathbf{G}$ for minimizing the transmitted power subject to the zero forcing constraint will be

$$\mathbf{G}_{opt} = (\mathbf{I}+\mathbf{B})(\mathbf{HF})^{\sharp}, \tag{11}$$

where $(\mathbf{HF})^{\sharp} = (\mathbf{F}^{\dagger}\mathbf{H}^{\dagger}\mathbf{HF})^{-1}\mathbf{F}^{\dagger}\mathbf{H}^{\dagger}$, which is the pseudo inverse of $(\mathbf{HF})$.

Proof: First note that the zero-forcing constraint is satisfied by (11):

$$\mathbf{G}_{opt}\mathbf{HF} - \mathbf{B} = (\mathbf{I}+\mathbf{B})(\mathbf{HF})^{\sharp}\mathbf{HF} - \mathbf{B} = \mathbf{I}.$$

Suppose there is another $\mathbf{G}'$ satisfying the zero forcing constraint with the given $\mathbf{F}$ and $\mathbf{B}$, i.e., $\mathbf{G}'\mathbf{HF} = \mathbf{I}+\mathbf{B}$. Define $\mathbf{\Delta} = \mathbf{G}_{opt} - \mathbf{G}'$. Since both $\mathbf{G}_{opt}$ and $\mathbf{G}'$ satisfy the zero-forcing constraint, it follows that

$$\begin{aligned}
\mathbf{\Delta}\mathbf{G}^{\dagger}_{opt} &= \mathbf{\Delta}\mathbf{HF}(\mathbf{F}^{\dagger}\mathbf{H}^{\dagger}\mathbf{HF})^{-\dagger}(\mathbf{I}+\mathbf{B})^{\dagger} \\
&= (\mathbf{G}_{opt}\mathbf{HF} - \mathbf{G}'\mathbf{HF})(\mathbf{F}^{\dagger}\mathbf{H}^{\dagger}\mathbf{HF})^{-\dagger}(\mathbf{I}+\mathbf{B})^{\dagger} \\
&= \mathbf{0}.
\end{aligned}$$

Therefore

$$\begin{aligned}
[\mathbf{G}'\mathbf{G}'^{\dagger}]_{kk} &= [(\mathbf{G}_{opt} - \mathbf{\Delta})(\mathbf{G}_{opt} - \mathbf{\Delta})^{\dagger}]_{kk} \\
&= [(\mathbf{G}_{opt}\mathbf{G}^{\dagger}_{opt} + \mathbf{\Delta}\mathbf{\Delta}^{\dagger}]_{kk} \\
&\geq [\mathbf{G}_{opt}\mathbf{G}^{\dagger}_{opt}]_{kk},
\end{aligned}$$

where we have used $\mathbf{\Delta}\mathbf{G}^{\dagger}_{opt} = \mathbf{0}$ in these inequalities. Therefore we have smaller sub-channel noise variances if we replace $\mathbf{G}'$ with $\mathbf{G}_{opt}$, hence with given bit rate and probabilities of error, the lower transmitted power can be achieved. ■

Therefore, when $\mathbf{F}$ and $\mathbf{B}$ are given, the best $\mathbf{G}$ is given as (11). Also, we can easily calculate that

$$\begin{aligned}
\mathbf{G}_{opt}\mathbf{G}^{\dagger}_{opt} &= (\mathbf{I}+\mathbf{B})(\mathbf{HF})^{\sharp}(\mathbf{HF})^{\sharp\dagger}(\mathbf{I}+\mathbf{B})^{\dagger} \\
&= (\mathbf{I}+\mathbf{B})(\mathbf{F}^{\dagger}\mathbf{H}^{\dagger}\mathbf{HF})^{-1}(\mathbf{I}+\mathbf{B})^{\dagger}.
\end{aligned}$$

Based on lemma 1, we can substitute $\mathbf{G}_{opt}$ to minimize the transmitted power in (10). We can rewrite the transmitted power as

$$\begin{aligned}
P_{trans} &= c2^{b}(\prod_{k=1}^{M}[\mathbf{F}^{\dagger}\mathbf{F}]_{kk})^{\frac{1}{M}} \\
&\quad \times(\prod_{k=1}^{M}[(\mathbf{I}+\mathbf{B})(\mathbf{F}^{\dagger}\mathbf{H}^{\dagger}\mathbf{HF})^{-1}(\mathbf{I}+\mathbf{B})^{\dagger}]_{kk})^{\frac{1}{M}}.
\end{aligned}$$

Using the Hadamard's inequality for positive definite matrices, we have $\prod_{k=1}^{M}[\mathbf{F}^{\dagger}\mathbf{F}]_{kk} \geq \det(\mathbf{F}^{\dagger}\mathbf{F})$ and

$$\begin{aligned}
&\prod_{k=1}^{M}[(\mathbf{I}+\mathbf{B})(\mathbf{F}^{\dagger}\mathbf{H}^{\dagger}\mathbf{HF})^{-1}(\mathbf{I}+\mathbf{B})^{\dagger}]_{kk} \\
&\geq \det((\mathbf{I}+\mathbf{B})(\mathbf{F}^{\dagger}\mathbf{H}^{\dagger}\mathbf{HF})^{-1}(\mathbf{I}+\mathbf{B})^{\dagger}) \\
&= \det((\mathbf{F}^{\dagger}\mathbf{H}^{\dagger}\mathbf{HF})^{-1}),
\end{aligned}$$

where we use the fact that the matrix $\mathbf{I}+\mathbf{B}$ is upper triangular and with diagonal terms all equal to unity. Thus $\det(\mathbf{I}+\mathbf{B}) = 1$. Therefore, we have

$$P_{trans} \geq c2^{b}\left(\frac{\det(\mathbf{F}^{\dagger}\mathbf{F})}{\det(\mathbf{F}^{\dagger}\mathbf{H}^{\dagger}\mathbf{HF})}\right)^{\frac{1}{M}}.$$

This is exactly the form of equation (6) in [13]. We also prove that P_{trans} is no less than some constant times the Mth-root of the product of the first dominant M channel singular values, i.e.,

$$P_{trans} \geq P_{min} = c2^{b}(\frac{1}{\prod_{k=1}^{M}\sigma_{h,k}^{2}})^{\frac{1}{M}}. \tag{12}$$

The proof is given in [19].

Note that P_{min} in (12) is exactly equal to the form derived for a linear transceiver with optimal bit loading [13]. This means, the extra freedom provided by the decision feedback receiver structure does not reduce the power needed to achieved the specified bit rate and probability of error. In other words, when bit loading is allowed, DFE with linear precoding systems has the same performance as linear transceivers! To the best of authors' knowledge, this fact has not been formally proved in earlier work. However, the DFE with linear precoding system actually provides more choices of possible configurations that achieve the P_{min} in (12). This interesting observation will be discussed extensively in the following sections.

IV. GTD-Based Systems

The GTD (generalized triangular decomposition) was introduced in [10] and used in [11] for the optimization of transceivers under QoS constraints. We show how the GTD of the channel can be used as a starting point to design transceivers which jointly optimize bit allocation and the matrices $\mathbf{F}$, $\mathbf{G}$, and $\mathbf{B}$.

Theorem 1: *The generalized triangular decomposition (GTD)*: Let $\mathbf{H} \in \mathcal{C}^{m\times n}$ be a given rank K matrix with singular values $\sigma_{h,1}, \sigma_{h,2}, \cdots, \sigma_{h,K}$ in descending order. Let $\mathbf{r} = [r_1, r_2, \cdots, r_K]$ be a given vector which satisfies

$$\mathbf{a} \prec_{\times} \mathbf{h}, \tag{13}$$

where $\mathbf{a} = [|r_1|, |r_2|, \cdots, |r_K|]$ and $\mathbf{h} = [\sigma_{h,1}, \sigma_{h,2}, \cdots, \sigma_{h,K}]$, and ”$\prec_{\times}$” denotes the multiplicative majorization relationship [14], [6]. Then there exist matrices $\mathbf{R}$, $\mathbf{Q}$, and $\mathbf{P}$ such that

$$\mathbf{H} = \mathbf{QRP}^{\dagger}, \tag{14}$$

where $\mathbf{R}$ is a $K\times K$ upper triangular matrix with diagonal terms equal to r_k, and $\mathbf{Q} \in \mathcal{C}^{m\times K}$ and $\mathbf{P} \in \mathcal{C}^{n\times K}$ both have orthonormal columns.

Proof: See [10]. ■

This decomposition is the extended version of the results by Weyl in 1949 [16] and Horn in 1954 [6], which give the complete relationship between the matrix singular values and eigenvalues. Special instances of the GTD include:

(a) The singular value decomposition (SVD) $\mathbf{H} = \mathbf{U\Sigma V}^{\dagger}$ where $\mathbf{\Sigma}$ is a diagonal matrix containing the singular values on the diagonal [7].

(b) The Schur decomposition $\mathbf{H} = \mathbf{Q\Delta Q}^{\dagger}$ where $\mathbf{\Delta}$ is an upper triangular matrix with eigenvalues of a square matrix $\mathbf{H}$ on the diagonal [7].

(c) The QR decomposition $\mathbf{H} = \mathbf{QR}$ where $\mathbf{R}$ is an upper triangular matrix (here $\mathbf{P} = \mathbf{I}$) [7].

(d) The complete orthogonal decomposition $\mathbf{H} = \mathbf{Q}_2\mathbf{R}_2\mathbf{Q}_1^{\dagger}$ [5], where $\mathbf{H}^{\dagger} = \mathbf{Q}_1\mathbf{R}_1$ is the QR factorization of $\mathbf{H}^{\dagger}$ and $\mathbf{R}_1^{\dagger} = \mathbf{Q}_2\mathbf{R}_2$ is the QR factorization of $\mathbf{R}_1^{\dagger}$.

(e) The geometric mean decomposition (GMD) $\mathbf{H} = \mathbf{QRP}^{\dagger}$ where $\mathbf{R}$ is an upper triangular matrix with the diagonal elements equal to the geometric means of the positive singular values [8].

(f) The bi-diagonal decomposition (BID) $\mathbf{H} = \mathbf{QRP}^\dagger$, where $\mathbf{R}$ is a bi-diagonal and upper triangular matrix [5].

In all these cases, the majorization property (13) can be verified to be true. In the following sections we will discuss those systems induced from the GTD concept, and show that each one of those systems can achieve the optimal minimized power in the problem considered in Section III. We observe that many existing systems are actually special cases of the GTD-based system, such as SVD systems [13], ZF-VBLAST systems [1], and GMD systems [8]. Furthermore, many novel systems, such as Schur transceiver, and bi-diagonal (BID) transceiver can also be conceived.

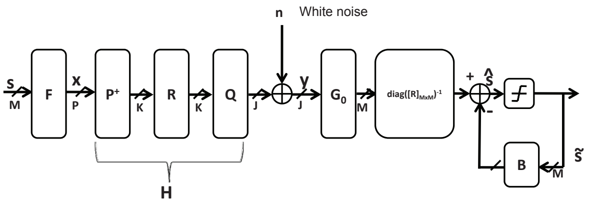

Fig. 2. The system with orthonormal linear precoding and DFE.

With the channel decomposed using the GTD as in (14) we now describe a method to construct the matrices $\mathbf{F}, \mathbf{G}, \mathbf{B}$. This design, with appropriate bit allocation, will be optimal in the sense described in Theorem 2 below. In (14) the matrix $\mathbf{R}$ is a $K \times K$ upper triangular matrix with the vector $\{[\mathbf{R}]_{M+1,M+1}, \cdots, [\mathbf{R}]_{K,K}\}$ equal to some permutation of the vector $\{\sigma_{h,M+1}, \cdots, \sigma_{h,K}\}$, which contains the smallest $K - M$ singular values of $\mathbf{H}$. The first M diagonal elements of $\mathbf{R}$, $\mathbf{r} = \{[\mathbf{R}]_{1,1}, \cdots, [\mathbf{R}]_{M,M}\} \in \mathcal{R}^+$, is multiplicatively majorized by the vector $\sigma = \{\sigma_{h,1}, \cdots, \sigma_{h,M}\}$, which contains the first M dominant singular values of $\mathbf{H}$. Here we assume the rank of the channel matrix $\mathbf{H}$ is K, and $K \geq M$. Note that this decomposition is possible because of the GTD theory [10]. Also we want to point out that with this decomposition we have

$$\prod_{k=1}^{M} [\mathbf{R}]_{kk}^2 = \prod_{k=1}^{M} \sigma_{h,k}^2, \tag{15}$$

which is a direct consequence of the multiplicative majorization relationship. This fact will be useful in later discussions.

Now consider Fig. 2. Suppose we choose the precoder $\mathbf{F}$ to be such that

$$\mathbf{P}^\dagger \mathbf{F} = \begin{pmatrix} \mathbf{I_M} \\ \mathbf{0} \end{pmatrix}.$$

i.e.,

$$\mathbf{F} = [\mathbf{P}]_{P \times M}. \tag{16}$$

Since $\mathbf{P}$ has orthonormal columns, $\mathbf{F}$ has orthonormal columns as well. The transmitted power will be

$$P_{trans} = \sum_{k=1}^{M} P_k [\mathbf{F}^\dagger \mathbf{F}]_{kk} = \sum_{k=1}^{M} P_k.$$

The matrix $\mathbf{G}_0$ will be chosen so that

$$\mathbf{G}_0 \mathbf{Q} = \begin{pmatrix} \mathbf{I_M} & \mathbf{0} \end{pmatrix},$$

i.e.,

$$\mathbf{G}_0 = [\mathbf{Q}^\dagger]_{M \times J}. \tag{17}$$

Since $\mathbf{Q}$ has orthonormal columns, $\mathbf{G}_0$ has orthonormal rows, therefore the noise covariance after the filter $\mathbf{G}_0$ will be

$$E[\mathbf{G}_0 \mathbf{n}\mathbf{n}^\dagger \mathbf{G}_0^\dagger] = \mathbf{G}_0 E[\mathbf{n}\mathbf{n}^\dagger] \mathbf{G}_0^\dagger = \sigma_n^2 \mathbf{I}.$$

Thus the noise remains white after passing through the filter $\mathbf{G}_0$. The signal sub-streams then will pass through some multipliers $\{[\mathbf{R}]_{ii}^{-1}\}$ before the decision devices. Those multipliers can be equivalently viewed as a diagonal matrix multiplied with the signal vector. Thus the feedforward filter can be written as

$$\mathbf{G} = (\text{diag}([\mathbf{R}]_{M \times M}))^{-1} \mathbf{G}_0. \tag{18}$$

Therefore the signal transfer function without the decision feedback will be

$$\begin{aligned} \mathbf{GHF} &= (\text{diag}([\mathbf{R}]_{M \times M}))^{-1} \mathbf{G}_0 \mathbf{QRP}^\dagger \mathbf{F} \\ &= (\text{diag}([\mathbf{R}]_{M \times M}))^{-1} \begin{pmatrix} \mathbf{I_M} & \mathbf{0} \end{pmatrix} \mathbf{R} \begin{pmatrix} \mathbf{I_M} \\ \mathbf{0} \end{pmatrix} \\ &= (\text{diag}([\mathbf{R}]_{M \times M}))^{-1} [\mathbf{R}]_{M \times M} \end{aligned}$$

The feedback filter $\mathbf{B}$ is the one that makes the zero-forcing constraint satisfied, i.e.,

$$\mathbf{B} = \mathbf{GHF} - \mathbf{I} = (\text{diag}([\mathbf{R}]_{M \times M}))^{-1} [\mathbf{R}]_{M \times M} - \mathbf{I}. \tag{19}$$

Since $\mathbf{R}$ is an upper triangular matrix, it can be seen that $\mathbf{B}$ in (19) will be strictly upper triangular. In this scenario, the noise variance in the k-th substream will be

$$\sigma_{e_k}^2 = \frac{\sigma_n^2}{[\mathbf{R}]_{kk}^2} \tag{20}$$

Substituting this into equation (3), the transmitted power needed to satisfy the specified QoS and bit rate constraints can be expressed as

$$P_{trans} = \sum_{k=1}^{M} d_k 2^{b_k} [\mathbf{F}^\dagger \mathbf{F}]_{kk} \sigma_{e_k}^2 = \sum_{k=1}^{M} \frac{d_k 2^{b_k}}{[\mathbf{R}]_{kk}^2} \sigma_n^2 = \sum_{k=1}^{M} \frac{c_k 2^{b_k}}{[\mathbf{R}]_{kk}^2}$$

In the following, we will prove that the system in Fig. 2 with $\mathbf{F}$ as in (16) and $\mathbf{G}$ as in (18) achieves the optimality.

Theorem 2: With b_k chosen as the following for $k = 1, 2, \cdots, M$:

$$\begin{aligned} b_k = \quad & \log_2 \left(\frac{c}{M} 2^b \left(\frac{1}{\prod_{k=1}^{M} \sigma_{h,k}^2} \right)^{\frac{1}{M}} \right) \\ & - \log_2(c_k) + \log_2([\mathbf{R}]_{kk}^2), \end{aligned} \tag{21}$$

the system in Fig. 2 with $\mathbf{F}$ as in (16) and $\mathbf{G}$ as in (18), achieves the minimized power with the given QoS and bit rate constraint. This means, the system is the optimal solution to the problem discussed in Section III. ♠

The proof can be found in [19]. It may be a little bit counter-intuitive to know that systems with DFE and linear precoding perform no better than linear transceivers when bit loading is allowed. However, as we shall point out, systems with DFE and linear precoding actually offer lots of flexibility in designing the transceivers. Firstly, the bit loading formula for the linear transceiver to achieve the minimum transmitted power is implicitly given in [13], and we reproduce it here:

$$b_k = D - \log_2 c_k + \log_2(\sigma^2_{h,k}), \qquad (22)$$

where D is some constant that satisfies the average bit rate constraint. The meaning of this formula is: allocate more bits to the better channel eigen-modes. The main drawback of this formula is that, it ignores the fact that the bits b_k are required be nonnegative and discrete. For the GTD-based system, the bit loading scheme is as in (21). The formula (21) tells us we should allocate more bits for the sub-stream with higher $[\mathbf{R}]_{kk}$. The freedom of the GTD-based system is that, we can re-shape the value of $[\mathbf{R}]_{kk}$ as long as the multiplicative majorization property is satisfied. The minimum transmitted power is still achievable by that specific GTD system. Those extra freedoms may be used to ensure that the bit loading scheme in (22) is realizable.

V. Numerical Results

In this section we consider a wireless communication system with multiple antenna at both sides of the link with perfect channel state information. We use 100 randomly generated MIMO channels for the simulation. The matrix channel is of size 5×4, and normalized so that $\mathbf{E}[|[\mathbf{H}]_{i,j}|^2] = 1$.

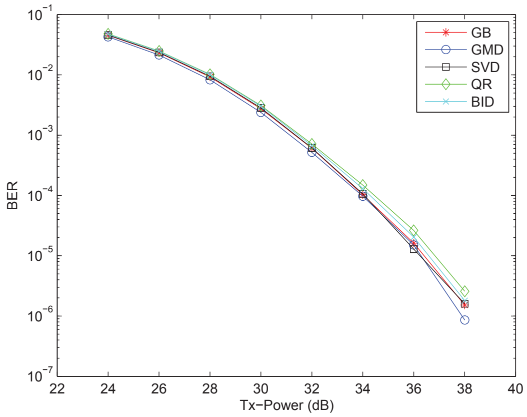

Fig. 3. BER versus Tx-Power when 32 bits are transmitted.

We implement five methods in the numerical results. “SVD”, “GMD”, “QR”, and “BID” stand for the each one of the special cases of GTD-based transceiver structures discussed in Section IV, respectively; “GB” stands for the following method: first we compute the ideal bit loading formula (22). If b_k is less than zero, we will drop this sub-stream without allocating any bits for it. Then we allocate the bits *uniformly* to the remaining sub-streams. With the bit allocation scheme fixed, we then compute the GTD-based precoder/equalizer structure for “GB” (using the equality (21) to find the required $[\mathbf{R}]_{kk}$ with same c_k for each sub-stream). Note that with this chosen GTD-based precoder/equalizer structure, the system is optimal for the integer bits being allocated in the remaining sub-streams according to the theorems developed in the paper. Therefore, the “GB” method actually suffers no problem due to integer constraints on bits. This is the freedom we have from the GTD-based system. Since the bit loading formula (9) is not realizable due to finite constellation granularity, we adopt the optimal bit loading algorithm in [4] under the given precoder/equalizer realization for the other four methods.

The additive noise is complex circulant Gaussian with average power normalized to 0 dB. The results are given in terms of bit error rate versus transmitted power. In Fig. 3 we consider the high bit rate case. There are 40 bits to be allocated into the four signal sub-streams. It can be observed that each systems performs about the same. This is consistent with theorem 1 and theorem 2. Notice in particular that the SVD system without DFE is almost as good as the systems with DFE.

VI. Concluding Remarks

We have presented a method for the joint optimization of the matrices $\{\mathbf{F}, \mathbf{G}, \mathbf{B}\}$ and the bits $\{b_k\}$ in a transceiver with DFE. It is formally shown that when the bit allocation, precoder, and equalizer are jointly optimized, linear transceivers and transceivers with DFE have identical performance in the sense that transmitted power is identical for a given bit rate and error probability. We also proved that any GTD-based system achieves the optimal performance. Many existing systems are identified to be special cases of the GTD-based system, while many novel GTD-based transceivers are also proposed in the paper. Both the theoretical analysis and numerical simulations have been provided to validate the effectiveness of our result.

References

[1] P. W. Wolniansky, G. J. Foschini, G. D. Golden, R. A. Valenzuela, “V-BLAST: an architecture for realizing very high data rates overthe rich-scattering wireless channel,” *IEEE Symp. on Signals, Systems, and Electronics,* pp. 295 - 300, Pisa, Italy, Oct. 1998.

[2] S. Dasgupta, and A. Pandharipande, “Optimum Biorthogonal DMT Systems for Multi-service Communication,” *IEEE Proc. ICASSP.* vol. IV. Hong Kong, pp. 552 - 555, Apr. 2003.

[3] D. P. Palomar, S. Barbarossa “Designing MIMO Communication Systems: Constellation Choice and Linear Transceiver Design,” *IEEE Trans. Sig. Proc.* pp. 3804 - 3818, Oct. 2005.

[4] Jorge Campello, “Optimal Discrete Bit Loading for Multicarrier Modulation Sysrtems” *IEEE ISIT.* pp. 193, Aug. 1998.

[5] G. H. Golub, and C. F. Van Loan, “Matrix computations,” The Johns Hopkins Univ. Press 1996.

[6] R. A. Horn, “On the eigenvalues of a matrix with prescribed singular values,” *Proceedings of the American Mathematical Society,* Vol. 5, No. 1. pp. 4-7. Feb. 1954.

[7] R. A. Horn, C. R. Johnson, *Matrix analysis,* Cambridge Univ. Press 1985.

[8] Y. Jiang, J. Li, and W. W. Hager “Joint transceiver design for MIMO communications using geometric mean decomposition,” *IEEE Trans. Sig. Proc.,* pp.3791-3803, Oct. 2005.

[9] Y. Jiang, J. Li, and W. W. Hager “Uniform channel decomposition for MIMO communications,” *IEEE Trans. Sig. Proc.,* pp.4283-4294, Nov. 2005.

[10] Y. Jiang, W. W. Hager, and J. Li “Generalized triangular decomposition,” *Mathematics of computation.,* Oct. 2007.

[11] Y. Jiang, W. W. Hager, and J. Li “Tunable channel decomposition for MIMO communications using channel state information,” *IEEE Trans. Sig. Proc.,* pp. 4405 - 4418, Nov. 2006.

[12] Y. Jiang, D. P. Palomar, and M. K. Varanasi, “Precoder Optimization for Nonlinear MIMO Transceiver Based on Arbitrary Cost Function,” *Proc. Conference on Information Sciences and Systems.* March, 2007.

[13] Y. P. Lin, and S. M. Phoong, “Optimal ISI-Free DMT transceivers for distorted channels with colored noise,” *IEEE Trans. Sig. Proc.* pp. 2702 - 2712, Nov. 2001.

[14] A. W. Marshall, I. Olkin, *Inequalities: Theory of majorization and its applications,* Academic Press. 1979.

[15] M. B. Shenouda, T. N. Davidson, “A framework for designing MIMO systems with decision feedback equalization or Tomlinson-Harashima precoding,” *Proc. of the ICASSP,* pp. 209-212, Apirl. 2007.

[16] H. Weyl “Inequalities between the two kinds of eigenvalues of linear transformations,” *Proc. Nat. Acad. Sci.* Vol. 35, pp. 408 - 411, July 1949.

[17] F. Xu, T. N. Davidson, J. K. Zhang, S. S. Chan, and K. M. Wong, “Design of block transcivers with MMSE decision feedback detection,” *Proc. of the ICASSP,* pp. 1109 - 1112, March. 2004.

[18] J. Zhang, A. Kavcic, and K. M. Wong, “Equal-diagonal QR decomposition and its application to precoder design for successive-cancellation detection,” *IEEE Trans. Info. Theory,* pp. 154 - 172, Jan. 2005.

[19] http://www.systems.caltech.edu/ccw/asilomar2008.html