Joint Bit Allocation and Precoding for MIMO Systems With Decision Feedback Detection

Svante Bergman, *Student Member, IEEE*, Daniel P. Palomar, *Senior Member, IEEE*, and Björn Ottersten, *Fellow, IEEE*

Abstract—This paper considers the joint design of bit loading, precoding and receive filters for a multiple-input multiple-output (MIMO) digital communication system employing decision feedback (DF) detection at the receiver. Both the transmitter as well as the receiver are assumed to know the channel matrix perfectly. It is well known that, for linear MIMO transceivers, a diagonal transmission (i.e., orthogonalization of the channel matrix) is optimal for some criteria. Surprisingly, it was shown five years ago that for the family of Schur-convex functions an additional rotation of the symbols is necessary. However, if the bit loading is optimized jointly with the linear transceiver, then this rotation is unnecessary. Similarly, for DF MIMO optimized transceivers a rotation of the symbols is sometimes needed. The main result of this paper shows that for a DF MIMO transceiver where the bit loading is jointly optimized with the transceiver filters, the rotation of the symbols becomes unnecessary, and because of this, also the DF part of the receiver is not required. The proof is based on a relaxation of the available bit rates on the individual substreams to the set of positive real numbers. In practice, the signal constellations are discrete and the optimal relaxed bit loading has to be rounded. It is shown that the loss due to rounding is small, and an upper bound on the maximum loss is derived. Numerical results are presented that confirm the theoretical results and demonstrate that orthogonal transmission and the truly optimal DF design perform almost equally well.

Index Terms—Channel coding, communication systems, fading channels, MIMO systems, precoding.

I. INTRODUCTION

MULTIPLE-INPUT multiple-output (MIMO) systems have received much attention in recent years due to the tremendous potential for high data throughput [1]. Under the assumption that the transmitter knows the channel perfectly, the capacity-optimal precoding strategy is to linearly orthogonalize the channel matrix using the singular value decomposition (SVD). Information is optimally conveyed over the orthogonal subchannels using infinitely long and Gaussian distributed codewords, with data rates assigned to the subchannels given by the so called waterfilling solution [1].

One downside with the capacity-optimal transmission scheme is the (infinitely) long codewords and the delay this brings to the system. Clearly, long delay has some practical disadvantages, especially when considering time-varying channels, systems with packet retransmission, or delay-sensitive applications. In most cases it is mathematically intractable to provide a global performance analysis of the delay-sensitive system as a whole, let alone to jointly optimize the system. By only optimizing the lowest layer of the communication chain, our hope is that the overall performance also can be brought close to the optimum. This motivates the separate analysis of the modulation part of the physical layer—before we apply outer error-correcting codes. When not considering error correcting codes, the system will suffer from an inherent non-zero probability of detection error. Thus, not only must the optimal design tradeoff uncoded data rate against power usage, but it needs to consider the bit error rate as well. Although SVD based, orthogonal, transmission is optimal in the sense of maximizing the mutual information, it is not necessarily optimal in this uncoded case.

In [2] and [3], it was shown that given a single-user link with a linear receiver, for the class of Schur-concave objective functions of the mean-square errors the orthogonal transmission is indeed optimal, while for the class of Schur-convex objectives it is not (an additional rotation of the transmitted symbols is required). This conclusion, however, is for systems with fixed signal constellations. Arguably the most common approach to multi-channel digital communication is to adapt the constellations for the subchannels according to their respective channel gains. This adaptation, sometimes referred to as adaptive modulation or bit loading [4], [5], will subsequently affect the objective function that we use in the precoder design. In the case of linear detection at the receiver, optimizing the constellations leads to a Schur-concave objective function [6]. The conclusion is that if the optimal constellations are used and if the receiver is linear, then the optimal precoder will orthogonalize the subchannels just as the case was for the capacity-optimal precoder.

Interestingly, the same conclusion does not hold when using an optimal maximum-likelihood (ML) detector at the receiver. A joint constellation-precoder design was proposed in [7] that demonstrated the suboptimality of orthogonal transmission for this specific case. Another design was proposed in [8] that gives the minimum BER solution for a 2×2 MIMO system with quadrature phase-shift keying modulation. Both of these designs include a rotation of the precoder such that the effective channel is not orthogonalized. Apparently the ML detector can resolve very complex multi-dimensional constellations that are more power efficient than the linearly separable constellations of the SVD-based precoder.

An intermediate solution between the linear receiver and the ML receiver is the decision feedback (DF) receiver [9]–[17]. In general, the DF receiver has low decoding complexity compared to the ML receiver [18], and for channels with inter-symbol interference the DF receiver outperforms the linear receiver in terms of error performance (the linear receiver is a special case of the DF receiver). In [16] and [17], it was shown that for multiplicative Schur-concave objective functions orthogonal transmission is optimal, while for multiplicative Schur-convex objectives a rotation of the signal vector is needed (similar to the linear case). This conclusion is for systems with fixed constellations.

This work investigates the problem of jointly optimizing the precoder, receiver filters, and bit loading when using the DF receiver on a point-to-point communication system. The main result is that the optimal bit loading will result in an orthogonalizing precoder design. Decision feedback becomes superfluous when the signal constellations are chosen properly. That said, DF may still be advantageous since it allows us to redistribute the bit loading on high-rate subchannels at a very low cost in terms of reduced performance, which in turn means that a suboptimal bit loading will perform almost as well as the optimal one. Another reason for using DF detection is that perfect transmitter-side CSI may be an unrealistic assumption. Imperfect transmitter-side CSI inevitably cause intersymbol interference that can be reduced using DF (note that perfect CSI is required for orthogonal transmission).

In addition to the results regarding bit loading, we show that the problem of computing the optimal precoder and receiver filters for a fixed bit loading can be posed as a convex problem. An algorithm that solves the convex problem with linear computational complexity is provided. Because of the low computational complexity of the filter optimization, an exhaustive search for the optimal bit loading becomes practically feasible—although the main result of the paper suggests that an exhaustive search is not necessary in most cases.

A. Notation

For an $N \times N$ matrix $\mathbf{X}$, denote the vector of the diagonal elements $\mathbf{d}(\mathbf{X}) = ([\mathbf{X}]_{1,1}, \ldots, [\mathbf{X}]_{N,N})^{\mathrm{T}}$. For a vector, $\mathbf{x}$, of length N, denote the diagonal $N \times N$ matrix with diagonal $\mathbf{x}$ as $\mathbf{D}(\mathbf{x})$. For notational simplicity, define also $\mathbf{D}(\mathbf{X}) = \mathbf{D}(\mathbf{d}(\mathbf{X}))$. The i'th element of the vector $\mathbf{x}$ is denoted x_i. The vector of dimension N with all ones is denoted $\mathbf{1}_N$, where the N may be scrapped if the dimension is clearly given from the context. The function $(x)^+$ is defined as $(x)^+ = x$ if $x > 0$, and $(x)^+ = 0$ if $x \le 0$. Rounding a real number x to the closest integer is denoted $\lfloor x \rceil$. The vector of element-wise absolute values of a vector $\mathbf{x}$ is denoted $|\mathbf{x}|$. The expected value of a random matrix $\mathbf{X}$ is denoted $\mathrm{E}[\mathbf{X}]$.

II. SYSTEM MODEL AND PROBLEM FORMULATION

Consider the discrete-time flat-fading linear model of a $N_r \times N_t$ MIMO communication system

$$\mathbf{y} = \mathbf{HFs} + \mathbf{v} \tag{1}$$

where $\mathbf{y} \in \mathbb{C}^{N_r}$ is the received signal, $\mathbf{H} \in \mathbb{C}^{N_r \times N_t}$ is the channel matrix, $\mathbf{F} \in \mathbb{C}^{N_t \times N}$ is a precoding matrix, $\mathbf{s} \in \mathbb{C}^N$ is
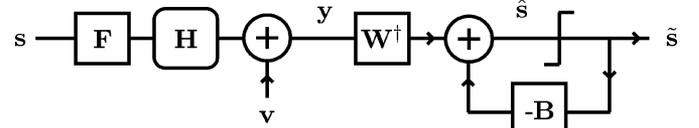

Fig. 1. Schematic view of the MIMO communication system with DF detection.

the data-symbols vector, and $\mathbf{v} \in \mathbb{C}^{N_r}$ is circularly symmetric additive white Gaussian noise. The data symbols and the noise are assumed to be normalized as $\mathrm{E}[\mathbf{ss}^\dagger] = \mathbf{I}$, and $\mathrm{E}[\mathbf{vv}^\dagger] = \mathbf{I}$. The average transmitted power is limited such that $\mathrm{Tr}\{\mathbf{FF}^\dagger\} \le P$, is satisfied.

A. Decision Feedback Receiver

In this work we assume that the receiver employs DF equalization, a schematic view of the system is depicted in Fig. 1. The received signal is linearly equalized using a forward filter, $\mathbf{W}^\dagger$, and subsequently passed to an elementwise detector of the data symbols. From the outcome of the detection, we reconstruct the transmitted data symbols, $\tilde{\mathbf{s}}$, then use the reconstructed symbols to remove interference between the symbols in the equalized signal. In order to ensure that the DF detection is sequential, i.e., that we do not feedback symbols that has not yet been detected, we enforce the feedback matrix $\mathbf{B}$ to be strictly lower triangular. The signal after the interference subtraction, $\hat{\mathbf{s}}$, is then passed on to the detector again. Taking the feedback into account, the error prior detection is

$$\mathbf{e} = \hat{\mathbf{s}} - \mathbf{s} = (\mathbf{W}^\dagger \mathbf{HF} - \mathbf{I})\mathbf{s} - \mathbf{B}\tilde{\mathbf{s}} + \mathbf{W}^\dagger \mathbf{v}. \tag{2}$$

If the probability of detection error is small, one can assume that $\tilde{\mathbf{s}}$ and $\mathbf{s}$ are zero mean and have close to identical correlation and cross-correlation matrices, and that $\tilde{\mathbf{s}}$ is uncorrelated with the noise $\mathbf{v}$. Using these approximations the error covariance matrix is given by

$$\begin{aligned}
\mathbf{R}_{\mathrm{MSE}} &= \mathrm{E}[\mathbf{ee}^\dagger] \\
&= (\mathbf{W}^\dagger \mathbf{HF} - \mathbf{B} - \mathbf{I})(\mathbf{W}^\dagger \mathbf{HF} - \mathbf{B} - \mathbf{I})^\dagger + \mathbf{W}^\dagger \mathbf{W}. \tag{3}
\end{aligned}$$

Since the detection of the symbols $\hat{\mathbf{s}} = \mathbf{s} + \mathbf{e}$ is made elementwise, we can regard the detection problem simply as detecting a scalar signal in additive complex Gaussian noise.[1] We denote each virtual transfer function $\hat{s}_i = s_i + e_i$ as a subchannel, for which the performance is determined by its virtual noise power, $[\mathbf{R}_{\mathrm{MSE}}]_{i,i}$.

In general, sequential decision feedback does not restrict us to use only lower-triangular feedback matrices, any joint row-column permutation is also possible. However, in this case where we are free to design both the precoder and the bit loading, such permutation loses its purpose since it can be absorbed into the other optimization parameters.

[1]Strictly speaking the interference part of the error is not complex Gaussian distributed. However, combined with the noise we can tightly approximate the interference as such by the law of large numbers.

B. Cost Functions Based on the Weighted Mean-Square Error

A general framework was presented in [17] for optimizing the DF system (i.e., the filters $\mathbf{F}$, $\mathbf{W}^\dagger$, and $\mathbf{B}$) based on monotonic cost functions of the MSEs of the subchannels. Our goal here is to optimize not only these DF filters, but also the signal constellations that are used on the subchannels. For mathematical tractability in the later analysis we narrow down the class of cost functions to p-norms of weighted MSEs. More precisely, consider the cost function $\|\mathbf{d}(\mathbf{R}_{\mathrm{MSE}}\mathbf{D}_w)\|_p$, where $\mathbf{D}_w$ is a weighting matrix assumed to be diagonal and non-negative, and the function

$$\|\mathbf{d}(\mathbf{X})\|_p = \left(\sum_{i=1}^N [\mathbf{X}]_{i,i}^p\right)^{1/p} \tag{4}$$

is the p-norm of the diagonal elements of $\mathbf{X}$. The p-norm is defined for $p \geq 1$.

To illustrate how to apply the cost function (4) in practice, consider minimizing the probability of detection error. Using the Gaussian-tail function

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-t^2/2}\, dt \tag{5}$$

the probability of error of subchannel i can be approximated as

$$P_{e,i} \simeq 4\, Q\left(\sqrt{\frac{d_{\min}^2(b_i)}{2[\mathbf{R}_{\mathrm{MSE}}]_{i,i}}}\right) \tag{6}$$

where $d_{\min}^2(b_i)$ denotes the squared minimum distance of a signal constellation that was modulated using b_i information bits and has been normalized to unit variance [19]. Equation (6) allows us to relate the MSE with the performance in terms of error probability. It also indicates how we should chose the MSE weighting matrix $\mathbf{D}_w$ in our cost function. Namely, in order to have symmetry among the subchannels, the weights should be inversely proportional to the squared minimum distance as

$$[\mathbf{D}_w]_{i,i} = d_{\min}^{-2}(b_i) \quad \forall i = 1, \ldots, N. \tag{7}$$

This will make the subchannels (approximately) symmetric with respect to symbol error rate (SER), which is a relevant measure, for example, if we want to minimize the joint probability of detection error. In the case when outer error correcting codes are used it may be more relevant to have symmetric bit error rates (BER) rather than SERs. Assuming Gray coded bit mapping the BERs can be approximated as

$$\mathrm{BER}_i \approx \frac{1}{b_i} P_{e,i}. \tag{8}$$

For moderately low, to low BERs, it can be shown that the dependency on b_i in the BER expression is dominated by the SER factor, $P_{e,i}$. Hence, symmetric SERs can serve as a good approximation to attain symmetric BERs as well.

For most classes of constellations used in practice, the minimum distance typically decreases exponentially with the number of bits b_i. For example, quadrature amplitude modulated (QAM) constellations with even bit loading has minimum distance

$$d_{\min}^2(b_i) = \frac{6}{2^{b_i} - 1} \tag{9}$$

resulting in a weighting matrix (disregarding constant factors)

$$\mathbf{D}_w = \mathbf{D}(2^{b_1} - 1, \ldots, 2^{b_N} - 1). \tag{10}$$

One objective could be to minimize the maximum error probability, $P_{e,i}$, of the subchannels. Under the high-SNR assumption, this objective translates into a cost function

$$\|\mathbf{d}(\mathbf{R}_{\mathrm{MSE}}\mathbf{D}_w)\|_\infty = \max_i [\mathbf{R}_{\mathrm{MSE}}\mathbf{D}_w]_{i,i} \tag{11}$$

corresponding to the $p = \infty$ norm.

Summarizing, if the SNR is high, the probability of error on a subchannel depend to a large extent on the minimum distance of the signal constellations. The minimum distance scales the MSE of the subchannels, which leads to imbalances when different types of signal constellations are used on different subchannels. Using a cost function with weighted MSE this imbalance can be compensated for. The parameter p of the cost function can be used to control how 'flexible' the system is in terms of the spread of the error rates among the subchannels. Low p results in more spread, which may be disadvantageous since the worst subchannel typically dominates. In general, and specifically for high SNRs, the infinity norm seems to translate into the lowest SERs in most cases.

C. Problem Formulation

With the definitions of the MSE matrix (3) and the cost function (4) in place, our problem can be mathematically formulated as

$$\underset{\mathbf{F},\mathbf{B},\mathbf{W}^\dagger,\mathbf{b}}{\text{minimize}} \quad \|\mathbf{d}(\mathbf{R}_{\mathrm{MSE}}(\mathbf{F},\mathbf{B},\mathbf{W}^\dagger)\mathbf{D}_w)\|_p \tag{12a}$$

$$\text{subject to} \quad \mathrm{Tr}\{\mathbf{F}\mathbf{F}^\dagger\} \leq P, \tag{12b}$$

$$[\mathbf{D}_w]_{i,i} = d_{\min}^{-2}(b_i) \quad \forall i = 1, \ldots, N, \tag{12c}$$

$$b_i \in \mathcal{B} \quad \forall i = 1, \ldots, N, \tag{12d}$$

$$\sum_{i=1}^N b_i = R \tag{12e}$$

where the vector $\mathbf{b}$ is the bit loading vector, and the set $\mathcal{B}$ denotes the set of feasible bit rates which is determined by the available signal constellations. Typically, due to the discrete nature of bits, this set is equal to the set of positive integers.

For a fixed bit loading, $\mathbf{b}$, the problem of designing the DF filters can be reformulated as a convex problem that can be solved numerically. In Section III, it is shown how to apply the framework in [17] to the particular problem considered here. In fact, the problem can be solved very efficiently with linear computational complexity. Once the optimal bit rates are known, the remaining problem is therefore fairly simple.

As for the optimization of the bit loading, the set $\mathcal{B}$ is discrete and it is possible to numerically try out all feasible bit loading combinations in order to find the global optimum. An alternative to such an exhaustive search is to relax the problem and extend the set $\mathcal{B}$ to allow for arbitrary positive bit rates. We do this by using (9) to relate these real-valued bit rates to virtual minimum-distance weights. This relaxation allows us to optimize the bit loading for any given MSE matrix. In Section IV, the optimal relaxed bit loading is derived. The remaining problem of jointly optimizing the DF filters is characterized in Section V, together with a discussion on the loss due to rounding of the bit rates. Finally, in Sections VI and VII, various practical strategies for solving the joint problem are presented and evaluated numerically.

III. DESIGN OF OPTIMAL FILTERS

Because the set of feasible bit loads is discrete, its optimization is not easy to combine with the filter and precoder design. This section treats the filter design problem alone, assuming a fixed bit loading. The problem is formulated as

$$\underset{\mathbf{F},\mathbf{B},\mathbf{W}^\dagger}{\text{minimize}} \quad \|\mathbf{d}(\mathbf{R}_{\text{MSE}}\mathbf{D}_w)\|_p \tag{13a}$$

$$\text{subject to} \quad \text{Tr}\{\mathbf{F}\mathbf{F}^\dagger\} \leq P. \tag{13b}$$

In [17, Theorem 4.3], it was shown how problems of this type (with monotonic cost functions of the MSEs) can be reduced to a power-loading problem involving a multiplicative majorization constraint.[2] In Sections III-A, III-B, and III-C, we will, for completeness, apply this procedure to problem (13), and give the filter expressions that are obtained in the process. In Section III-D, we then show how the resulting problem with a majorization constraint can be replaced with a convex problem that is easy to solve numerically.

A. Optimal Forward Receiver Filter

For a given transmit matrix, $\mathbf{F}$, and receive DF matrix $\mathbf{B}$ the optimal forward filter $\mathbf{W}^\dagger$ in the receiver is the well known minimum MSE (MMSE) equalizer

$$\mathbf{W}^\dagger = (\mathbf{B}+\mathbf{I})(\mathbf{F}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{F}+\mathbf{I})^{-1}\mathbf{F}^\dagger\mathbf{H}^\dagger. \tag{14}$$

The proof is given by completing the squares of (3) and then applying the matrix inversion lemma.[3] Using the optimal forward filter, the resulting vector of weighted MSEs is

$$\mathbf{d}(\mathbf{R}_{\text{MSE}}\mathbf{D}_w) = \mathbf{d}\left(\left(\tilde{\mathbf{B}}+\mathbf{D}_w^{1/2}\right)\right.$$
$$\left.\times (\mathbf{F}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{F}+\mathbf{I})^{-1}\left(\tilde{\mathbf{B}}+\mathbf{D}_w^{1/2}\right)^\dagger\right) \tag{15}$$

where $\tilde{\mathbf{B}} \triangleq \mathbf{D}_w^{1/2}\mathbf{B}$. Note that the MMSE equalizer minimizes any monotonic increasing function of the MSEs and is thus optimal for a wider class of cost functions than considered here.

[2] A short recapitulation on vector majorization is provided in Appendix A.

[3] $(\mathbf{A}+\mathbf{B}\mathbf{C}\mathbf{D})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{D}\mathbf{A}^{-1}\mathbf{B}+\mathbf{C}^{-1})^{-1}\mathbf{D}\mathbf{A}^{-1}$

B. Optimal Feedback Receiver Filter

Consider the Cholesky factorization of

$$(\mathbf{F}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{F}+\mathbf{I})^{-1} = \mathbf{L}\mathbf{L}^\dagger \tag{16}$$

where $\mathbf{L}$ is lower triangular. Inserting (16) into (15) we obtain the weighted MSE of subchannel i as

$$[\mathbf{R}_{\text{MSE}}\mathbf{D}_w]_{i,i} = \sum_{j=1}^i \left|\left[\left(\tilde{\mathbf{B}}+\mathbf{D}_w^{1/2}\right)\mathbf{L}\right]_{i,j}\right|^2. \tag{17}$$

Since $\tilde{\mathbf{B}}\mathbf{L}$ is a strictly lower-triangular matrix (zero diagonal) it can only affect the non-diagonal elements of the lower-triangular matrix $\tilde{\mathbf{B}}\mathbf{L}+\mathbf{D}_w^{1/2}\mathbf{L}$. Hence, the feedback matrix, $\mathbf{B}$, that minimizes the MSEs in (17) satisfies

$$\tilde{\mathbf{B}}\mathbf{L} = -\mathcal{L}_{\text{strict}}\left(\mathbf{D}_w^{1/2}\mathbf{L}\right) \tag{18}$$

where $\mathcal{L}_{\text{strict}}(\mathbf{X})$ the matrix that contains the strictly lower-diagonal part of $\mathbf{X}$. The optimal $\tilde{\mathbf{B}}$ is then

$$\tilde{\mathbf{B}} = -\mathbf{D}_w^{1/2}(\mathbf{L}-\mathbf{D}(\mathbf{L}))\mathbf{L}^{-1} \tag{19}$$

and the resulting minimum MSE of subchannel i is

$$[\mathbf{R}_{\text{MSE}}\mathbf{D}_w]_{i,i} = \left[\mathbf{D}_w^{1/2}\mathbf{D}(\mathbf{L})\right]_{i,i}^2 = [\mathbf{D}_w]_{i,i}[\mathbf{L}]_{i,i}^2. \tag{20}$$

C. Optimal Precoder: Left and Right Unitary Matrices

The remaining filter to optimize is the precoder, $\mathbf{F}$. Consider the SVD of the precoder matrix

$$\mathbf{F} = \mathbf{U}_\mathbf{F}\boldsymbol{\Sigma}_\mathbf{F}^{1/2}\mathbf{V}_\mathbf{F}^\dagger \tag{21}$$

where $\mathbf{U}_\mathbf{F}$ and $\mathbf{V}_\mathbf{F}$ are unitary matrices, and $\boldsymbol{\Sigma}_\mathbf{F}^{1/2}$ is non-negative and diagonal with decreasing diagonal. For arbitrary monotonic increasing objective functions of the MSEs, the optimal left unitary matrix has been shown to match the matrix of eigenvectors of $\mathbf{H}^\dagger\mathbf{H} = \mathbf{U}_\mathbf{H}\boldsymbol{\Lambda}_\mathbf{H}\mathbf{U}_\mathbf{H}^\dagger$ [2], [17, Theorem 4.3]:

$$\mathbf{U}_\mathbf{F} = \mathbf{U}_\mathbf{H} \tag{22}$$

where it is assumed that the diagonal of $\boldsymbol{\Lambda}_\mathbf{H}$ is decreasing. Using (22), the right unitary matrix, $\mathbf{V}_\mathbf{F}$, can readily be obtained from the SVD of $\mathbf{L}$ as

$$\mathbf{F}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{F}+\mathbf{I} = \mathbf{V}_\mathbf{F}(\boldsymbol{\Sigma}_\mathbf{F}\boldsymbol{\Lambda}_\mathbf{H}+\mathbf{I})\mathbf{V}_\mathbf{F}^\dagger \Longrightarrow$$
$$\mathbf{L} = \mathbf{V}_\mathbf{F}(\boldsymbol{\Sigma}_\mathbf{F}\boldsymbol{\Lambda}_\mathbf{H}+\mathbf{I})^{-1/2}\mathbf{Q}^\dagger \tag{23}$$

where $\mathbf{Q}$ is a unitary matrix that makes $\mathbf{L}$ lower triangular. Note here that the diagonal matrix with the singular values $\boldsymbol{\Lambda}_\mathbf{L} \triangleq (\boldsymbol{\Sigma}_\mathbf{F}\boldsymbol{\Lambda}_\mathbf{H}+\mathbf{I})^{-1/2}$ is in this case ordered with increasing diagonal. Finally, also assuming $\mathbf{L}$ is known, the singular values of $\mathbf{F}$ are obtained from $\boldsymbol{\Lambda}_\mathbf{L}$ and $\boldsymbol{\Lambda}_\mathbf{H}$ as

$$\boldsymbol{\Sigma}_\mathbf{F} = \boldsymbol{\Lambda}_\mathbf{H}^{-1}\left(\boldsymbol{\Lambda}_\mathbf{L}^{-2}-\mathbf{I}\right). \tag{24}$$

With the optimal $\mathbf{W}^\dagger$, $\mathbf{B}$, and $\mathbf{U_F}$, the remaining optimization problem is

$$\underset{\mathbf{V_F},\mathbf{Q},\boldsymbol{\sigma}}{\text{minimize}} \quad \||\mathbf{D}_w|\mathbf{d}(\mathbf{L})|^2\|_p \tag{25a}$$

$$\text{subject to} \quad \mathbf{L} = \mathbf{V_F}(\mathbf{D}(\boldsymbol{\sigma})\mathbf{\Lambda_H}+\mathbf{I})^{-1/2}\mathbf{Q}^\dagger \tag{25b}$$

$$[\mathbf{L}]_{i,j} = 0 \quad \forall i < j \tag{25c}$$

$$\mathbf{V_F},\mathbf{Q} \in \mathcal{U} \tag{25d}$$

$$\boldsymbol{\sigma} \geq 0 \tag{25e}$$

$$\mathbf{1}^\mathrm{T}\boldsymbol{\sigma} \leq P \tag{25f}$$

where we introduced the power vector $\boldsymbol{\sigma} = \mathbf{d}(\mathbf{\Sigma_F})$, that represents the power assigned to each spatial channel, and where $\mathcal{U}$ is the set of all N-dimensional unitary matrices.

Now, optimizing the unitary matrices $\mathbf{V_F}$, $\mathbf{Q}$, directly is very difficult. From the expressions of the optimal DF filters we see that the filters either implicitly or explicitly depend on the lower triangular matrix $\mathbf{L}$. By optimizing $\mathbf{L}$ instead of $\mathbf{V_F}$, $\mathbf{Q}$, we can avoid the unitary constraints. In order to do this we need a way to specify the singular values of $\mathbf{L}$ as an optimization constraint, fortunately this is possible. The diagonal elements of a triangular matrix represents the eigenvalues of the same matrix. It is known that the absolute values of the eigenvalues are always multiplicatively majorized by the singular values (cf. [20]). Interestingly, this necessary condition is also a sufficient condition on the triangular matrix $\mathbf{L}$: For a given power load, $\mathbf{\Sigma_F}$, and a specified vector $\mathbf{d}(\mathbf{L})$, one can uniquely determine (using generalized triangular decomposition [21]) the lower triangular matrix $\mathbf{L}$ if and only if

$$|\mathbf{d}(\mathbf{L})|^{-2} \preceq_\times \mathbf{d}(\mathbf{\Sigma_F}\mathbf{\Lambda_H}+\mathbf{I}) \tag{26}$$

where $\preceq_\times$ denotes multiplicative majorization [22], [17]. The proof of this statement was given in [21]. With this necessary and sufficient condition on the diagonal elements of a triangular matrix we can replace the constraints (25b), (25c), and (25d), in problem (25). Matrix notation becomes tedious (and unnecessary) to work with at this point, instead define the vectors $\mathbf{w} = \log\mathbf{d}(\mathbf{D}_w)$, $\mathbf{y} = \log|\mathbf{d}(\mathbf{L})|^{-2}$, and use (26) to pose the equivalent optimization problem in vector notation as

$$\underset{\mathbf{y},\boldsymbol{\sigma}}{\text{minimize}} \quad \|\exp(\mathbf{w}-\mathbf{y})\|_p \tag{27a}$$

$$\text{subject to} \quad \mathbf{y} \preceq \log(\mathbf{\Lambda_H}\boldsymbol{\sigma}+\mathbf{1}) \tag{27b}$$

$$\boldsymbol{\sigma} \geq 0 \tag{27c}$$

$$\mathbf{1}^\mathrm{T}\boldsymbol{\sigma} \leq P \tag{27d}$$

where $\preceq$ denotes additive majorization. The vector $\mathbf{w}$ relates to the MSE weights of the problem and we call it the log-weights vector. The vector $\mathbf{y}$ has the interpretation of data rate [see (27b)], and we name it rate vector.[4]

D. Optimal Precoder: Power Allocation

The optimization problem (27) is somewhat difficult to work with due to the majorization constraint. Fortunately a simpler convex problem can be considered instead. Because the p-norm

---

[4]Note also, that $\mathbf{y}$ is the logarithm of the inverse of the MSEs, which corresponds to the mutual information over AWGN channels using Gaussian codebooks.
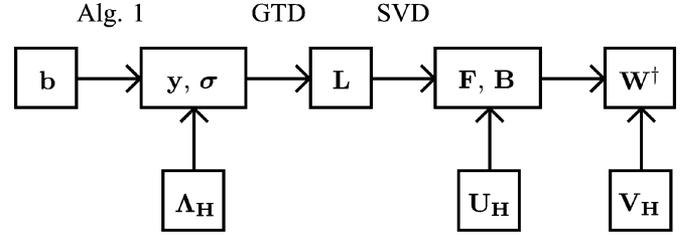

Fig. 2. Flow chart for the calculation of the optimal DF filters given a bit loading vector $\mathbf{b}$. Algorithm 1 denotes the algorithm in Appendix C, GTD denotes generalized triangular decomposition.

is symmetric and because a majorization inequality is invariant to permutations of the vector elements, we can without loss of generality assume the subchannels are ordered such that both $\mathbf{w}$ and $\boldsymbol{\lambda} = \mathbf{d}(\mathbf{\Lambda_H})$ are decreasing.

Now that we have fixed the ordering of $\mathbf{w}$ and $\boldsymbol{\lambda}$, we may consider the following optimization problem:

$$\underset{\mathbf{y},\boldsymbol{\sigma}}{\text{minimize}} \quad \|\exp(\mathbf{w}-\mathbf{y})\|_p \tag{28a}$$

$$\text{subject to} \quad \sum_{j=1}^{i} y_j - \log(1+\sigma_j\lambda_j) \leq 0 \quad \forall i \tag{28b}$$

$$\boldsymbol{\sigma} \geq 0 \tag{28c}$$

$$\mathbf{1}^\mathrm{T}\boldsymbol{\sigma} \leq P. \tag{28d}$$

Note that problem (27) differs from (28), in that the majorization constraint has been replaced with another (similar) constraint that does not include the monotonic rearrangement of the vector elements. Although the two problems seem to be different, the following theorem states that the both problems share the same optimal solution.

Theorem 3.1: Given that $\boldsymbol{\lambda}$, and $\mathbf{w}$ are decreasing, the optimum solutions of problems (27) and (28) coincide.

Proof: See Appendix B. ∎

Problem (28) is convex and can be solved numerically with relative ease using standard tools for convex optimization [23]. In Appendix C, we present an algorithm that solves the problem exactly with only $O(N)$ complexity.

To summarize, Fig. 2 shows the flow chart for the procedure of calculating the optimal DF filters. The first step computes the power loading, $\boldsymbol{\sigma}$, and rate vector, $\mathbf{y}$, e.g., by using the algorithm in Appendix C. The second step uses the generalized triangular decomposition to compute the Cholesky factor, $\mathbf{L}$. The third step computes both the precoder, $\mathbf{F}$, and the feedback filter, $\mathbf{B}$ from $\mathbf{L}$. Finally, the feed forward filter, $\mathbf{W}^\dagger$, is computed from $\mathbf{F}$, and $\mathbf{B}$. Note that after the first step we can already evaluate the objective value. It is therefore less computationally demanding to evaluate the performance than it is to compute the optimal filters for a particular bit loading. This fact allows us to reduce the complexity when an exhaustive search for the jointly optimal bit loading is performed (as was proposed in Section II-C).

IV. OPTIMAL BIT LOADING

This section switches focus to the bit loading problem, i.e., the problem of computing the optimal $\mathbf{b}$ in (12). In the previous section it was shown that for any weighting vector $\mathbf{w}$, problem (12) can be simplified to the form (27). Because the bit loading

and the DF filters are coupled only through the cost function (there are no common constraints), the results from Section III can be incorporated into the original problem formulation (12) as

$$\underset{\mathbf{y},\boldsymbol{\sigma},\mathbf{b}}{\text{minimize}} \quad \|\exp(\mathbf{w}-\mathbf{y})\|_p \tag{29a}$$

$$\text{subject to} \quad \mathbf{y} \preceq \log(\boldsymbol{\Lambda}_{\mathbf{H}}\boldsymbol{\sigma}+\mathbf{1}) \tag{29b}$$

$$\boldsymbol{\sigma} \geq 0 \tag{29c}$$

$$\mathbf{1}^{\mathrm{T}}\boldsymbol{\sigma} \leq P \tag{29d}$$

$$w_i = -\log d_{\min}^2(b_i) \quad \forall i \tag{29e}$$

$$b_i \in \mathcal{B} \quad \forall i \tag{29f}$$

$$\mathbf{1}^{\mathrm{T}}\mathbf{b} = R. \tag{29g}$$

In general, the set of available constellations, $\mathcal{B}$, is discrete. In particular, if QAM constellations are used the bit rates are restricted to positive, even integers. The set of feasible bit rates is then

$$b_i \in \{0, 2, 4, \ldots\} \quad \forall i \tag{30a}$$

$$\mathbf{1}^{\mathrm{T}}\mathbf{b} = R. \tag{30b}$$

Clearly, the feasible set is finite and it is possible to compute the optimal bit loading numerically by trying out all possible candidates. This observation does however provide little insight into the overall behavior of the system. It is also questionable whether it is worth the computational burden to globally search all possible bit loading candidates. In order to gain more insight and to find heuristics for computing the bit rates more efficiently, the following subsection considers optimizing $\mathbf{b}$ while the vectors $\mathbf{y}$ and $\boldsymbol{\sigma}$ remain fixed (recall that Section III was optimizing $\mathbf{y}$ and $\boldsymbol{\sigma}$ for a fixed $\mathbf{b}$). Then, later, Section V will be devoted to the joint optimization of all three vectors $\mathbf{y}$, $\boldsymbol{\sigma}$, and $\mathbf{b}$.

A. Continuous Bit Loading Relaxation

Optimization of the discrete valued bit loading is difficult in closed form. One way to approach this optimization problem is to relax the set of bit rates to the continuous domain [by ignoring the constraint (30a)] to allow for us to analytically optimize $\mathbf{b}$. This leads to the continuous relaxation of problem (29), where $\mathcal{B} = \mathbb{R}_+$. In order to specify constraint (29e), we assume for simplicity that the constellations are QAM. Then, using (10), the log-weights, $\mathbf{w}$, depend on the bit allocations as

$$e^{\mathbf{w}} = \mathbf{d}(\mathbf{D}_w) = e^{\mathbf{b}} - \mathbf{1} \tag{31}$$

where the unit of the rate has been changed to nats (rather than bits) in order to simplify the notation below. For a given $\mathbf{y}$ and $\boldsymbol{\sigma}$, using weights defined by (31), the continuous relaxation of problem (29) with $\mathbf{y}$, $\boldsymbol{\sigma}$ fixed, is then formulated as

$$\underset{b_1,\ldots,b_N}{\text{minimize}} \quad \sum_{i=1}^{N}(e^{b_i}-1)^p e^{-p y_i} \tag{32a}$$

$$\text{subject to} \quad \sum_{i=1}^{N} b_i = R \tag{32b}$$

$$b_i \geq 0 \quad \forall i \tag{32c}$$

where we use the fact that $\|\cdot\|_p$ and $\|\cdot\|_p^p$ are minimized simultaneously.[5] Note (by inspection) that the problem is convex.

Theorem 4.1: The optimum bit allocation in (32) is given by

$$b_i = g(\mu + y_i) \quad \forall i = 1, \ldots, N \tag{33}$$

where the function $g(x)$ is defined as

$$\begin{cases} x = (1 - p^{-1})\log\left(1 - e^{-g(x)}\right) + g(x), & \text{if } p > 1 \\ g(x) = (x)^+, & \text{if } p = 1 \end{cases} \tag{34}$$

and where μ is chosen so that

$$\sum_{i=1}^{N} g(\mu + y_i) = R. \tag{35}$$

Proof: In the following proof we assume $p \in (1,\infty)$. The proofs for $p = 1$ and $p \to \infty$ are similar but needs to be treated separately. Disregarding for a while the constraint that $\mathbf{b}$ must be positive, minimizing the Lagrangian cost function of (32) yields the optimal solution to the problem as

$$p e^{b_i}(e^{b_i}-1)^{p-1}e^{-p y_i} = \theta \quad \forall i = 1, \ldots, N \tag{36}$$

where θ is the dual variable such that constraint (32b) is satisfied. Equation (36) contains multiple roots. However, if there exists a root with strictly positive b_i's, then it must also be a global optimum to the convex problem (32) (a convex problem does not have local optima unless they also are global optima).

Taking the logarithm of (36) and performing some rearrangements yields

$$f(b_i) \triangleq (1 - p^{-1})\log(1 - e^{-b_i}) + b_i$$
$$= p^{-1}\log(\theta\, p^{-1}) + y_i. \tag{37}$$

Note that the function $f(b_i)$ is real valued when all b_i's are positive. By inspection, $f(b_i)$ is strictly increasing, concave, and it maps the set $(0,\infty)$ to the set $(-\infty,\infty)$. Because $f(b_i)$ is strictly increasing and concave, the inverse function $g(x)$ exists and is strictly increasing and convex. Following (37), the function $g(x)$ must satisfy

$$x = (1 - p^{-1})\log\left(1 - e^{-g(x)}\right) + g(x). \tag{38}$$

This implies that for any vector $\mathbf{y}$ there exists one and only one solution to (36) with strictly positive bit rates, $\mathbf{b}$, given by

$$b_i = g(\mu + y_i) \quad \forall i = 1, \ldots, N \tag{39}$$

where $\mu \triangleq p^{-1}\log(\theta\, p^{-1})$. ∎

Given a rate vector $\mathbf{y}$, (33) and (35) uniquely determines the optimal bit loading (as well as μ). These equations will later be applied to eliminate $\mathbf{b}$ from the joint optimization problem.

The observant reader may have noticed that the optimal relaxed bit loading will never be exactly zero on any subchannel. Instead of switching off weak subchannels with zero bit loading,

---

[5]The infinity norm is not well defined at this point and has to be treated separately. The conclusions of the following discussion are however valid for the infinity norm as well.

it turns out that it is more favorable to use an infinitesimal (positive) bit rate. Of course there is no such thing as infinitesimal bit rates in practice, recall that the relaxation is merely a tool that we can use to obtain practically implementable bit-loading candidates by means of rounding. Fortunately, as will be shown in Section V-B, the impact that a low-rate subchannel has on the rest of the system is limited, i.e., the performance will remain close to optimal if we turn off the low-rate subchannels. After the bit loading these low-rate subchannels will in any case be rounded to zero when we perform the rounding. The next subsection contains a comment on the sensitivity of the relaxed optimum towards rounding.

B. Rounded Bit Loading

In practice, arbitrary real-valued bit rates are not implementable and the impact of rounding or quantization of the bit rates has to be considered. Assume $\tilde{\mathbf{b}}$, is the optimal solution to the relaxed bit loading problem for some given rate vector $\mathbf{y}$, and assume that $\mathbf{b}'$ is the rounded or quantized version of $\tilde{\mathbf{b}}$ such that the sum rate is R. Denote the logarithm of the objective function (29a) as

$$ J(\mathbf{b}, \mathbf{y}) = p^{-1} \log \left(\sum_{i=1}^{N} (e^{b_i} - 1)^p e^{-p y_i} \right) \quad (40) $$

where we have chosen the weights according to (31). The first order Taylor expansion of (40) around the optimal bit loading $\tilde{\mathbf{b}}$ yields

$$ J(\mathbf{b}', \mathbf{y}) \approx J(\tilde{\mathbf{b}}, \mathbf{y}) + \frac{\sum_{i=1}^{N} (e^{\tilde{b}_i} - 1)^{p-1} e^{\tilde{b}_i} e^{-p y_i} \delta_i}{e^{p J(\tilde{\mathbf{b}}, \mathbf{y})}} \quad (41) $$

where $\delta = \mathbf{b}' - \tilde{\mathbf{b}}$. Now, using (36), and assuming both $\mathbf{b}'$ and $\tilde{\mathbf{b}}$ satisfy (32b) so that $\mathbf{1}^T \delta = 0$, the first order term in the expansion sums to zero

$$ J(\mathbf{b}', \mathbf{y}) \approx J(\tilde{\mathbf{b}}, \mathbf{y}) + \frac{\theta \sum_{i=1}^{N} \delta_i}{p \, e^{p J(\tilde{\mathbf{b}}, \mathbf{y})}} = J(\tilde{\mathbf{b}}, \mathbf{y}). \quad (42) $$

This result indicates that rounding of the optimal relaxed bit loading can be performed without too much loss in performance, although it is not clear how to quantify the loss. In the following section the loss is quantified for the joint optimum by making a distinction between low-rate and high-rate subchannels.

V. Joint Optimization of Bit Loading and Filters

Now that we know how to obtain the optimal DF filters (via $\mathbf{y}$ and $\boldsymbol{\sigma}$) for a given bit allocation (cf. Section III), and how to optimize the bit allocation $\mathbf{b}$ given vectors $\mathbf{y}$ and $\boldsymbol{\sigma}$ (cf. Section IV), our next step is to combine these results into a joint optimization problem.

A. The Bit-Loading Optimized Objective

The optimal relaxed bit allocation from Section IV depends on the rate vector $\mathbf{y}$. By inserting the optimal bit loading into the refined transceiver problem (27), we obtain a new objective with a dependence on $\mathbf{y}$ that is not as easily characterized as before. This subsection analyzes the behavior of this bit-loading optimized objective with respect to $\mathbf{y}$. As it turns out, even though the dependency on $\mathbf{y}$ is complicated, it is still possible to determine the optimal $\mathbf{y}$ as a function of $\boldsymbol{\sigma}$.

The optimal relaxed bit-loading vector $\tilde{\mathbf{b}}$, obtained from (32) in the original objective, yields the following objective function:

$$ J(y_1, \ldots, y_N) = p^{-1} \log \left(\sum_{i=1}^{N} (e^{\tilde{b}_i} - 1)^p e^{-p y_i} \right) \quad (43) $$

where the logarithm is introduced for later mathematical simplicity. The optimal bit allocation must satisfy (33), which can be reformulated to

$$ (e^{\tilde{b}_i} - 1)^p e^{-p y_i} = e^{p\mu} (1 - e^{-\tilde{b}_i}) \quad \forall i. \quad (44) $$

Using $\tilde{b}_i = g(\mu + y_i)$, the bit-loading optimized cost function is then formulated without the vector $\tilde{\mathbf{b}}$ as

$$ J(y_1, \ldots, y_N) = \mu + p^{-1} \log \sum_{i=1}^{N} \left(1 - e^{-g(\mu + y_i)} \right) \quad (45) $$

where μ is chosen such that

$$ \sum_{i=1}^{N} g(\mu + y_i) = R. \quad (46) $$

Note that (45) is also valid for the ∞-norm when $p^{-1} = 0$.

Using the new bit-optimized cost function, the remaining optimization problem (that determines the DF filters) is

$$
\begin{aligned}
\underset{\mathbf{y}, \boldsymbol{\sigma}}{\text{minimize}} \quad & J(y_1, \ldots, y_N) & (47a) \\
\text{subject to} \quad & \mathbf{y} \preceq \log(\boldsymbol{\Lambda}_{\mathbf{H}} \boldsymbol{\sigma} + \mathbf{1}) & (47b) \\
& \boldsymbol{\sigma} \geq 0, \quad \mathbf{1}^T \boldsymbol{\sigma} \leq P. & (47c)
\end{aligned}
$$

Although the problem is non-convex and perhaps difficult to solve, it turns out the cost function is symmetric and concave which enables us to solve at least parts of the problem with relative ease.

Theorem 5.1: The function $J(y_1, \ldots, y_N)$ is Schur-concave with respect to $y_1, \ldots, y_N$.

Proof: See Appendix D. ∎

A direct consequence of Theorem 5.1 is the following important corollary.

Corollary 5.2: Orthogonal SVD-based transmission with no decision feedback is always an optimal solution to the decision feedback problem given that the optimal relaxed bit loading is used.

Proof: Because the objective is Schur-concave (see Appendix A for definition), the optimal vector $\mathbf{y}$ must satisfy the majorization constraint with equality (cf. [24]), i.e., we have $\mathbf{y} = \log(\mathbf{1} + \boldsymbol{\Lambda}_{\mathbf{H}} \boldsymbol{\sigma})$. This means that $\mathbf{V}_{\mathbf{F}}$ in (21) can be chosen as the identity matrix so that the subchannels are orthogonalized. Orthogonal subchannels implies that $\mathbf{L}$ is diagonal and that the optimal feedback matrix $\mathbf{B}$ is zero (see Section III-B). ∎

Although the remaining problem of computing the optimal power load, $\boldsymbol{\sigma}$, is a non-convex problem with a non-trivial solution, the result above shows that it suffices to use state of the art SVD-based bit and power loading schemes (e.g., [6]) to compute a close-to-optimal bit loading.

Theorem 5.1 relies on the continuous relaxation, and the behavior of $J(\mathbf{y})$ for discrete constellation sets is not clear at this point. On the other hand, as was shown in Section IV-B, small perturbations of the optimal relaxed bit load will not significantly alter the value of the cost function. So, rounding the optimal bit loading should still remain close to optimal. In the next subsection, an upper bound on the loss due to rounding of the bit rates is derived.

B. Turning Off Low-Rate Subchannels

Essentially, rounding the bit loading corresponds to turning off low-rate subchannels and slightly perturbing the bit rates on the remaining high-rate subchannels. In this subsection we will show that the loss by turning off low-rate subchannels is relatively small, and then, that a system with no active low-rate subchannels is insensitive to reallocations of the bit loading.

An interesting property of $g(x)$ is that its asymptotes[6] coincide with the function $(x)^+$. Therefore, by analyzing (45) and (46), we see that weak subchannels with values of x that are negative or close to zero will have almost no impact on μ or on $J(y_1, \ldots, y_N)$. These subchannels can consequently be turned off at a very low cost in terms of performance. To formalize this, assume that $\mathbf{y}$ is decreasing and that all $N - K$ weakest subchannels with indexes $i > K$ are turned off. This will result in a new dual variable $\tilde{\mu}$ and cost function $\tilde{J}(y_1, \ldots, y_K)$ as

$$\tilde{J}(y_1, \ldots, y_K) = \tilde{\mu} + p^{-1} \log \sum_{i=1}^{K} \left(1 - e^{-g(\tilde{\mu}+y_i)} \right) \quad (48)$$

$$\sum_{i=1}^{K} g(\tilde{\mu} + y_i) = R. \quad (49)$$

The following theorem quantifies the loss.

Theorem 5.3: The loss when turning off the $N - K$ weakest subchannels can be upper bounded as

$$\tilde{J}(y_1, \ldots, y_K) - J(y_1, \ldots, y_N)$$
$$\leq \frac{\sum_{i=K+1}^{N} g(\mu + y_i)}{K - \sum_{i=1}^{K} \frac{1 - p^{-1}}{e^{g(\mu+y_i)} - p^{-1}}}. \quad (50)$$

Proof: See Appendix E.   ∎

In order to get a sense of how this bound behaves, denote the sum rate of the truncated subchannels

$$\Delta_R = \sum_{i=K+1}^{N} g(\mu + y_i). \quad (51)$$

Assuming the active subchannels satisfy $e^{g(\mu+y_i)} \gg 1$, then the denominator in (50) can be approximated with K, and the bound becomes

$$\tilde{J} - J \leq \frac{\Delta_R}{R} \cdot \frac{R}{K}. \quad (52)$$

As an example, typical figures for Δ_R/R could be on the order of 10% while the average data rate per active subchannel is typically less than, lets say, 3 nats. This would correspond to a maximum loss of around 1 dB.

The next step is to see how the cost function behaves when all low-rate subchannels have been turned off. Given that all

---

6First note that the range of $g(x)$ is non-negative, then from (34) we see that $g(x) \gg 1 \Rightarrow x \approx g(x)$ and $g(x) \approx 0 \Rightarrow x \approx (1 - p^{-1}) \log g(x)$.

---

subchannels are high rate we can apply $e^{g(\tilde{\mu}+y_i)} \gg 1$ to the definition (34) and obtain

$$g(\tilde{\mu} + y_i) \approx \tilde{\mu} + y_i. \quad (53)$$

By applying the asymptote to (49), the cost function (48) tends to

$$\tilde{J}(y_1, \ldots, y_K) \approx \frac{R}{K} - \frac{1}{K} \sum_{i=1}^{K} y_i + p^{-1} \log(K). \quad (54)$$

Given that the majorization constraint (47b) is satisfied, the following equality holds

$$\sum_{i=1}^{K} y_i = \sum_{i=1}^{K} \log(\lambda_i \sigma_i + 1) \quad (55)$$

and we can eliminate $\mathbf{y}$ completely from (54). Interestingly, any $\mathbf{y}$ that satisfies (47b) can be used and still be optimal. Since there is a direct connection between the optimal $\mathbf{b}$ and $\mathbf{y}$, this result implies that we can redistribute the bit allocations at a very low cost, provided the resulting $\mathbf{y}$ satisfies (47b) and the data rates on the active subchannels remain sufficiently high.

The results in this subsection predicts very limited losses when rounding the relaxed bit loading. This fact is further motivated by the numerical results in Section VII where almost identical performance of the truly optimal bit loading (achieved by a global search) and the rounded optimal relaxed bit loading is shown.

VI. TRANSMISSION SCHEMES

Due to the potential high complexity of the truly optimal joint bit loading and filter design, this section defines (in addition to the optimal design) three suboptimal schemes. Two of which, in theory, should perform very close to optimal.

A. Optimal Design

This transmission scheme is optimal in terms of (12). The strategy is to exhaustively search all combinations of bit loading allocations. For each bit loading candidate, optimize the rate vector, $\mathbf{y}$, and the power loading vector, $\boldsymbol{\sigma}$, by solving problem (28). Compute the weighted MSEs, and use the bit loading with the least weighted MSE.

B. Suboptimal Bit Loading

As Theorem 5.1 shows, the optimal relaxed bit loading will make the DF optimized system orthogonal. In [6], the so called gap approximation was used for determining the constellations of an orthogonal system. The gap approximation is close to optimal for the orthogonal system, and since the optimal bit loading with DF results in an orthogonal system it must be approximately optimal in this case as well.

In short, using the gap approximation the bit loading can be computed as

$$b_i = 2 \left\lfloor \frac{1}{2} \log_2(\tilde{\sigma}_i \lambda_i \Gamma^{-1} + 1) \right\rceil \quad (56)$$

where the gap, Γ, is chosen such that $\sum_i b_i = R_{\text{tot}}$, and where $\tilde{\sigma}_i$ is the waterfilling power allocation, given by

$$\tilde{\sigma}_i = \left(\Phi - \Gamma \lambda_i^{-1} \right)^+ \quad (57)$$

where the water level Φ is chosen such that $\sum_i \tilde{\sigma}_i = P$. Insertion of the waterfilling power allocation yields

$$b_i = 2 \left\lfloor \frac{1}{2} \log_2(\lambda_i) + \alpha \right\rceil^+ \qquad (58)$$

where α is a constant such that $\sum_i b_i = R$. The precoder, forward filter, and feedback filter are then optimized for this particular bit loading.

C. Orthogonal Design

As was shown in Section V, the optimal relaxed bit loading makes the subchannels orthogonal. It was also shown that the first order Taylor expansion around the optimal relaxed bit loading is constant. Hence, an optimal design under the constraint that the subchannels are forced to be orthogonal should perform almost as well as the two schemes above.

Use the gap approximation to compute the bit rates, then design the optimal *orthogonalizing* precoder and forward filter for this particular bit loading. That is, design the optimal precoder such that the interference among the subchannels is zero. Since the subchannels are orthogonal for this scheme, the optimal feedback matrix will be zero.

D. Equal Rate Design

The bit rates are distributed uniformly among all available subchannels. Again, the precoder, the forward filter, and the feedback filter are subsequently optimized for this particular bit loading.

VII. NUMERICAL RESULTS

In this section we numerically compare the schemes that was introduced in Section VI. For simplicity only the infinity norm has been considered as cost function. Fig. 3 shows a comparison of the scheme over an 8×8 Rayleigh-fading MIMO channel. The data rate is set to 24 bits per channel use. The optimal design and the suboptimal bit loading design have almost identical performance. This confirms that the rounding of the bits does not affect the overall performance significantly, and the DF filters compensate for small deviations from the optimal bit loading. The orthogonal design performs only slightly worse, which is an indication that if the appropriate bit loading is used, the importance of DF is very limited. The final scheme, equal bit loading, shows that the bit loading is important for achieving optimal performance. The difference in performance is however less than one dB and this indicates that DF can partly compensate for suboptimal bit loading. In Fig. 4, a 4×4 Rayleigh-fading channel was simulated with a data rate at 12 bits per channel use with similar results.

VIII. CONCLUSION

In this work, we considered the problem of joint optimization of the bit loading, precoder, and receiver filters for a DF-detection system. It was shown that minimizing the probability of detection error can be translated into minimizing a weighted p-norm of the MSEs. Then, by fixing the bit loading, it was shown that the problem of optimizing the precoder and receiver filters may be reduced to a convex optimization problem that
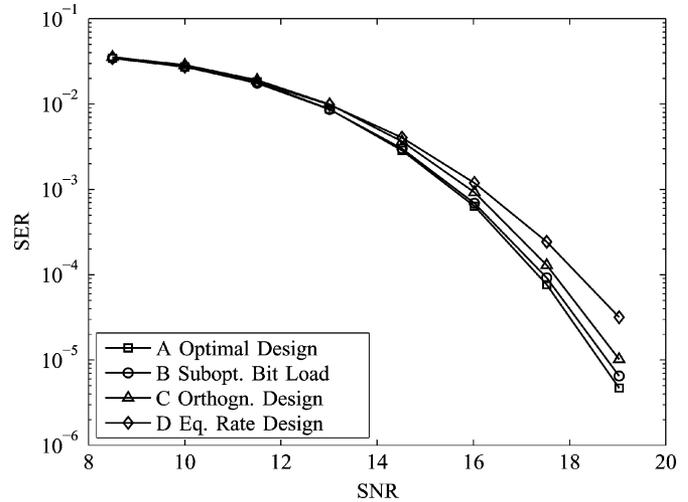

Fig. 3. Monte Carlo simulations of an 8×8 MIMO system. The data rate is set to 24 bits, and the cost function is the infinity norm.
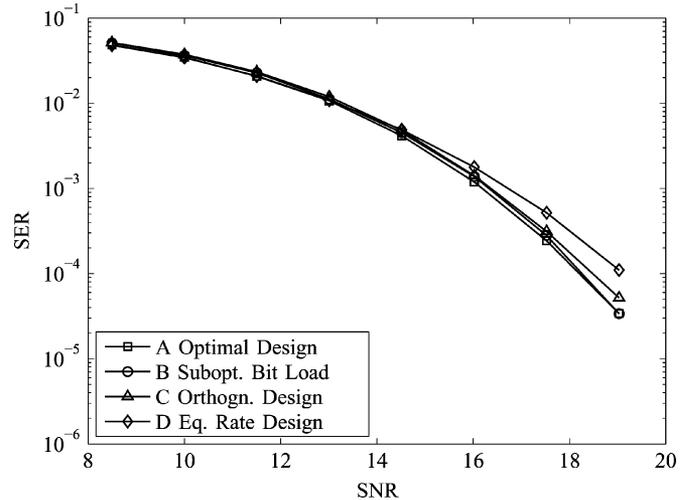

Fig. 4. Monte Carlo simulations of a 4×4 MIMO system. The data rate is set to 12 bits, and the cost function is the infinity norm.

is easy to solve numerically. Due to the low computational complexity of the problem, the task of jointly optimizing the bit loading and filters by exhaustively searching through all possible bit-loading candidates becomes a feasible option in practice.

In another approach to the same problem, by instead fixing the DF filters, we derived the optimal bit loading by relaxing the integer constraints on the subchannel bit rates. It was shown that this optimum is insensitive towards small deviations in the bit loading. When combining the relaxed bit loading with filter optimization, we showed that it is optimal to use orthogonal non-interfering subchannels. Therefore, by jointly optimizing bit loading and filters, the DF part of the receiver becomes superfluous. That said, another conclusion is that the DF receiver makes the system more robust towards rounding of the bit loading. These results were illustrated numerically by comparisons between the truly optimal solution and various suboptimal transmission strategies.

APPENDIX

A. Definitions From Majorization Theory

Denote $x_{[1]}, x_{[2]}, \ldots, x_{[N]}$ as the monotonic rearrangement of a vector $\mathbf{x}$ such that $x_{[1]} \geq x_{[2]} \geq \cdots \geq x_{[N]}$. For two vectors $\mathbf{x}$ and $\mathbf{y}$, additive majorization is defined as

$$\mathbf{x} \preceq \mathbf{y} \Leftrightarrow \begin{cases} \sum_{i=1}^{k} x_{[i]} \leq \sum_{i=1}^{k} y_{[i]} \forall k = 1, \ldots, N-1 \\ \sum_{i=1}^{N} x_{[i]} = \sum_{i=1}^{N} y_{[i]}. \end{cases}$$

Similarly, multiplicative majorization is defined as

$$\mathbf{x} \preceq_{\times} \mathbf{y} \Leftrightarrow \begin{cases} \prod_{i=1}^{k} x_{[i]} \leq \prod_{i=1}^{k} y_{[i]} \forall k = 1, \ldots, N-1 \\ \prod_{i=1}^{N} x_{[i]} = \prod_{i=1}^{N} y_{[i]} \end{cases}.$$

A function $f(\mathbf{x})$ is said to be Schur-convex if

$$\mathbf{x} \preceq \mathbf{y} \implies f(\mathbf{x}) \leq f(\mathbf{y}) \tag{59}$$

similarly it is defined Schur-concave if

$$\mathbf{x} \preceq \mathbf{y} \implies f(\mathbf{x}) \geq f(\mathbf{y}). \tag{60}$$

Multiplicative Schur-convex/concave functions are defined in a similar fashion using $\preceq_{\times}$ instead of $\preceq$. For a more complete introduction to majorization theory, please see [24].

B. Proof of Theorem 3.1

First we show that if $\boldsymbol{\lambda}$ is decreasing, then the optimal power loading, $\boldsymbol{\sigma}$, for problems (27) and (28) must ensure that $\boldsymbol{\Lambda}_{\mathbf{H}}\boldsymbol{\sigma}$ is decreasing: Assume that $\boldsymbol{\Lambda}_{\mathbf{H}}$ has a strictly positive diagonal. Define $\boldsymbol{\alpha} \triangleq \boldsymbol{\Lambda}_{\mathbf{H}}\boldsymbol{\sigma}$. Let $\boldsymbol{\Pi}$ be an arbitrary permutation matrix. Define an alternative power allocation, $\tilde{\boldsymbol{\sigma}}$, that yields a permutation of $\boldsymbol{\alpha}$ as $\boldsymbol{\Lambda}_{\mathbf{H}}\tilde{\boldsymbol{\sigma}} \triangleq \boldsymbol{\Pi}\boldsymbol{\alpha}$. Now, the total power consumption for the alternative power allocation is $\mathbf{1}^{\mathrm{T}}\tilde{\boldsymbol{\sigma}} = \mathbf{1}^{\mathrm{T}}\boldsymbol{\Lambda}_{\mathbf{H}}^{-1}\boldsymbol{\Pi}\boldsymbol{\alpha}$. Because $\boldsymbol{\Lambda}_{\mathbf{H}}^{-1}\mathbf{1}$ is increasing by assumption, the permutation matrix that yields the minimum power consumption is the one that makes $\boldsymbol{\Pi}\boldsymbol{\alpha}$ decreasing. Consequently, if $\boldsymbol{\alpha}$ is not decreasing then it cannot be optimal, and for both problems the optimal solution yields a decreasing vector $\log(\boldsymbol{\Lambda}_{\mathbf{H}}\boldsymbol{\sigma} + \mathbf{1})$.

As a consequence, by forcing the vector $\boldsymbol{\Lambda}_{\mathbf{H}}\boldsymbol{\sigma}$ to be decreasing in both problems we do not change their corresponding optima. The two problems can therefore be reformulated with more strict constraints: The reformulated version of problem (27) is

$$\underset{\mathbf{y}, \boldsymbol{\sigma}}{\text{minimize}} \quad \|\exp(\mathbf{w} - \mathbf{y})\|_p \tag{61a}$$

$$\text{subject to} \quad \sum_{j=1}^{i} y_{[j]} - \log(1 + \lambda_j \sigma_j) \leq 0 \quad \forall i \tag{61b}$$

$$\sum_{i=1}^{N} y_i = \sum_{i=1}^{N} \log(1 + \lambda_i \sigma_i) \tag{61c}$$

$$\lambda_i \sigma_i \geq \lambda_{i+1} \sigma_{i+1} \quad \forall 1 \leq i < N \tag{61d}$$

$$\boldsymbol{\sigma} \geq 0 \tag{61e}$$

$$\mathbf{1}^{\mathrm{T}}\boldsymbol{\sigma} \leq P \tag{61f}$$

and the corresponding reformulation of (28) is

$$\underset{\mathbf{y}, \boldsymbol{\sigma}}{\text{minimize}} \quad \|\exp(\mathbf{w} - \mathbf{y})\|_p \tag{62a}$$

$$\text{subject to} \quad \sum_{j=1}^{i} y_j - \log(1 + \sigma_j \lambda_j) \leq 0 \quad \forall i \tag{62b}$$

$$\sum_{j=1}^{N} y_j - \log(1 + \sigma_j \lambda_j) = 0 \tag{62c}$$

$$\lambda_i \sigma_i \geq \lambda_{i+1} \sigma_{i+1} \quad \forall 0 \leq i < N \tag{62d}$$

$$\boldsymbol{\sigma} \geq 0 \tag{62e}$$

$$\mathbf{1}^{\mathrm{T}}\boldsymbol{\sigma} \leq P. \tag{62f}$$

Note that the equality constraint (62c) is a consequence of the objective (62a) being decreasing with respect to y_N, so that the optimal y_N must hit the upper bound defined by the only inequality containing y_N.

From the definition of monotonic rearrangements we have

$$\sum_{j=1}^{i} y_j \leq \sum_{j=1}^{i} y_{[j]} \tag{63}$$

and we see that problem (62) is a relaxation of (61). Furthermore, since the function $\|\exp(-\mathbf{z})\|_p$ is Schur-convex with respect to $\mathbf{z}$, and because $\mathbf{w}$ and $\log(\boldsymbol{\Lambda}_{\mathbf{H}}\boldsymbol{\sigma} + \mathbf{1})$ are decreasing, it can be shown, due to the constraints (62b), that the optimal $\mathbf{y}$ in (62) will be decreasing (cf. [25]). This means that the optimum of the relaxed problem (62) is also a feasible point given the constraints in (61). Hence, the problems (61), (62), (27), and (28) have equivalent optimal solutions.

C. Algorithm That Solves Problem (28)

Algorithm 1 presented below solves problem (28) exactly with only $O(N)$ complexity. The proof of this statement is available in [26]. The algorithm is a modified version of the algorithm in [15] that solves the quality of service (QoS) constrained MSE optimization problem.

The input to the algorithm is the vectors $\mathbf{w}$ and $\boldsymbol{\lambda}$ of length N, and the p parameter. In the algorithm, the arrow "$\leftarrow$" denotes assignment of a variable, "&" and "|" denotes the logical operators AND and OR, respectively. When the algorithm has terminated the result set consists of the following variables: $q, \alpha, I_0, \ldots, I_q$, and $C_0, \ldots, C_{q-1}$. The sequence $I_0, \ldots, I_q$ is an ordered subset of the indexes $0, \ldots, N$ and it defines the indexes for which the constraints (28b) are satisfied with equality. Within each interval $I_i + 1, \ldots, I_{i+1}$ the optimal power allocation follows a water-filling-like solution where the water levels are given by $e^{C_i - \alpha}$. From the result set we can calculate the optimal power assignments $\sigma_1, \ldots, \sigma_N$, and MSE exponents $y_1, \ldots, y_N$, as follows.

- For all $j = I_i + 1, \ldots, I_{i+1}$, and for all $i = 0, \ldots, q-1$, assign

$$\sigma_j = \left(e^{C_i - \alpha} - \lambda_j^{-1}\right)^+, \quad y_j = w_j - p^{-1}C_i - \alpha. \tag{64}$$

- For all $j = I_q + 1, \ldots, N$, assign $\sigma_j = 0, y_j = 0$.

Algorithm 1: Power allocation algorithm

1: Allocate memory for the following vectors of length $N + 1$:
   $I[0], \ldots, I[N], C[0], \ldots, C[N], K[0], \ldots, K[N]$.
2: Initialization of variables:
   $I[0] \leftarrow 0, q \leftarrow 0, n \leftarrow 0, k \leftarrow 0$,
   $s \leftarrow 0, A \leftarrow$ true, $B \leftarrow$ true.
   For all $i = 1, \ldots, N$:
   $\tilde{w}_i \leftarrow \sum_{j=1}^{i} w_j, \tilde{g}_i \leftarrow \sum_{j=1}^{i} \lambda_j^{-1}, \tilde{h}_i \leftarrow \sum_{j=1}^{i} \log \lambda_j^{-1}$,
   $\tilde{w}_0 = \tilde{h}_0 = \tilde{g}_0 = 0$.
3: **while** $A \mid B$ **do**
4:    Given $I[q]$ and n, calculate the highest
      $k \in \{I[q]+1, \ldots, n\}$ such that $\gamma \geq \log \lambda_k^{-1} + (1+\beta)\alpha$,
      where

$$\gamma = \frac{\tilde{w}_n - \tilde{w}_{I[q]} + \tilde{h}_k - \tilde{h}_{I[q]}}{k + np^{-1} - (1+p^{-1})I[q]},$$
$$\beta = \frac{(n-k)}{k + np^{-1} - (1+p^{-1})I[q]},$$

and finally α that is evaluated by solving the non-linear equation

$$(P + \tilde{g}_k)e^{\alpha} = s + (k - I[q])e^{\gamma - \alpha\beta}.$$

5:    $K[q] \leftarrow k - I[q]$
6:    $C[q] \leftarrow \gamma - \beta\alpha$
7:    **if** $(q > 0)\&(C[q-1] \leq C[q])$ **then**
8:       $q \leftarrow q - 1$
9:       $s \leftarrow s - K[q]\,e^{C[q]}$
10:   **else**
11:      **if** $n \leq N$ **then**
12:         $A \leftarrow (k = n)\&(w_{n+1} - \alpha > p^{-1}(\log \lambda_{n+1}^{-1} + \alpha))$
13:         $B \leftarrow (p^{-1}C[q] < w_{n+1} - \alpha)$
14:      **else**
15:         $A \leftarrow$ false, $B \leftarrow$ false
16:      **end if**
17:      **if** $A \mid \mathbf{not}(B)$ **then**
18:         $s \leftarrow s + K[q]\,e^{C[q]}$
19:         $q \leftarrow q + 1$
20:         $I[q] \leftarrow n$
21:      **end if**
22:      $n \leftarrow n + 1$
23:   **end if**
24: **end while**

D. Proof of Theorem 5.1

First we will derive the second order derivative of the cost function (45), then, using the second order derivatives, we will show that the cost function is concave.

Let $\phi = p^{-1} \in [0, 1]$. The derivative of $g(x)$ is

$$g' = \frac{\partial g}{\partial x} = \frac{e^g - 1}{e^g - \phi}. \tag{65}$$

Rearranging the equation yields

$$\phi g' e^{-g} = g' + e^{-g} - 1. \tag{66}$$

Define $g_i = g(\mu + y_i)$, and $g_i' = g'(\mu + y_i)$, then the derivative with respect to y_n is

$$\frac{\partial g_i}{\partial y_n} = \frac{\partial g(\mu + y_i)}{\partial y_n} = \left(\frac{\partial \mu}{\partial y_n} + \delta_{n,i}\right) g_i' \tag{67}$$

where $\delta_{n,i} = 1$ if $n = i$, and zero otherwise. Differentiating (46) with respect to y_n results in

$$\frac{\partial \mu}{\partial y_n} = -\frac{g_n'}{\sum_i g_i'}. \tag{68}$$

Using (67), then (66), and finally (68), the derivative of (45) with respect to y_n is

$$\begin{aligned}
\frac{\partial J}{\partial y_n} &= \frac{\partial \mu}{\partial y_n} + \frac{\sum_i \phi e^{-g_i} g_i'}{\sum_i 1 - e^{-g_i}} \frac{\partial \mu}{\partial y_n} + \frac{\phi e^{-g_n} g_n'}{\sum_i 1 - e^{-g_i}} \\
&= \frac{\sum_i g_i'}{\sum_i 1 - e^{-g_i}} \frac{\partial \mu}{\partial y_n} + \frac{g_n' + e^{-g_n} - 1}{\sum_i 1 - e^{-g_i}} \\
&= -\frac{1 - e^{-g_n}}{\sum_i 1 - e^{-g_i}}.
\end{aligned} \tag{69}$$

The second order derivative of $J(\cdot)$ with respect to y_n, y_m, is given by

$$\begin{aligned}
\frac{\partial^2 J}{\partial y_n \partial y_m} = &-\frac{e^{-g_n}}{\sum_i 1 - e^{-g_i}} \frac{\partial g_n}{\partial y_m} \\
&+ \frac{1 - e^{-g_n}}{(\sum_i 1 - e^{-g_i})^2} \left(\sum_i e^{-g_i} \frac{\partial g_i}{\partial y_m}\right).
\end{aligned} \tag{70}$$

Using (67) and (68) we can obtain

$$\frac{\partial g_n}{\partial y_m} = \sqrt{g_n'} \left(\delta_{n,m} - \frac{\sqrt{g_n'}\sqrt{g_m'}}{\sum_j g_j'}\right) \sqrt{g_m'}. \tag{71}$$

In order to show that $J(\mathbf{y})$ is concave, we will compute the Hessian matrix. To do that, we first introduce the following diagonal matrices

$$[\mathbf{A}]_{i,i} = g_i', \quad [\mathbf{B}]_{i,i} = 1 - e^{-g_i}. \tag{72}$$

Note that since $g \geq 0$, and $g' \geq 0$, both $\mathbf{A}$ and $\mathbf{B}$ are positive semi-definite (PSD). With these definitions we can define a matrix $\mathbf{G}$ as

$$[\mathbf{G}]_{n,m} = \frac{\partial g_n}{\partial y_m} \implies$$
$$\mathbf{G} = \mathbf{A}^{1/2} \left(\mathbf{I} - \frac{\mathbf{A}^{1/2}\mathbf{1}\mathbf{1}^{\mathrm{T}}\mathbf{A}^{1/2}}{\mathbf{1}^{\mathrm{T}}\mathbf{A}\mathbf{1}}\right) \mathbf{A}^{1/2} \tag{73}$$

and then, using (70), we can derive the Hessian matrix of $J(\cdot)$ as

$$\mathbf{H} = -\frac{(\mathbf{I} - \mathbf{B})\mathbf{G}}{\mathbf{1}^{\mathrm{T}}\mathbf{B}\mathbf{1}} + \frac{\mathbf{B}\mathbf{1}\mathbf{1}^{\mathrm{T}}(\mathbf{I} - \mathbf{B})\mathbf{G}}{(\mathbf{1}^{\mathrm{T}}\mathbf{B}\mathbf{1})^2}. \tag{74}$$

Because $\mathbf{G}^{\mathrm{T}}\mathbf{1} = \mathbf{0}$, the Hessian can be simplified as

$$\mathbf{H} = -\frac{\mathbf{CG}}{\mathbf{1}^{\mathrm{T}}\mathbf{B1}} \qquad (75)$$

where

$$\mathbf{C} \triangleq \mathbf{I} - \mathbf{B} + \frac{\mathbf{B11}^{\mathrm{T}}\mathbf{B}}{\mathbf{1}^{\mathrm{T}}\mathbf{B1}}. \qquad (76)$$

We will show in a few steps that this Hessian, $\mathbf{H}$, is a negative semi-definite matrix. As a reference regarding the various properties of PSD matrices we refer to [27]. The center factor of $\mathbf{G}$ is a projection matrix (thus PSD), and consequently the entire matrix $\mathbf{G}$ is PSD. By inspection, the matrices $\mathbf{I} - \mathbf{B}$ and $\mathbf{B11}^{\mathrm{T}}\mathbf{B}$ are both PSD, and because the sum of two PSD matrices is also PSD, $\mathbf{C}$ is PSD. The eigenvalues of the product of two PSD matrices are always real and non-negative, and consequently we know that the Hessian has non-positive real eigenvalues. Any real, symmetric matrix with non-positive real eigenvalues is negative semi-definite, hence the Hessian is negative semi-definite.[7] By inspection, the function $J(\cdot)$ is component-wise symmetric and thus, because it is jointly concave, it is also Schur-concave [24].

E. Proof of Theorem 5.3

Due to (46), the dual variable μ will inevitably be affected when reducing the number of subchannels. Denote the new dual variable $\tilde{\mu}$, and (46) gives

$$\sum_{i=1}^{K} g(\tilde{\mu} + y_i) = \sum_{i=1}^{N} g(\mu + y_i). \qquad (77)$$

Since $g(x) \geq 0$ and $g'(x) \geq 0$, we have $\tilde{\mu} \geq \mu$. The convexity of $g(x)$ implies

$$g(\tilde{\mu} + y_i) - g(\mu + y_i) \geq g'(\mu + y_i)(\tilde{\mu} - \mu) \qquad (78)$$

where the derivative is specified in (65). Apply (78) to (77) as

$$\tilde{\mu} - \mu \leq \frac{\sum_{i=K+1}^{N} g(\mu + y_i)}{K - \sum_{i=1}^{K} \frac{1 - p^{-1}}{e^{g(\mu + y_i)} - p^{-1}}}. \qquad (79)$$

Note that, because $\tilde{\mu} \geq \mu$, $g(x)$ is positive and increasing, and because $y_{K+1}, \ldots, y_N$ correspond to the weakest subchannels; the vectors

$$\tilde{\mathbf{b}} = [g(\tilde{\mu} + y_1), \ldots, g(\tilde{\mu} + y_K), 0, \ldots, 0]^{\mathrm{T}}$$
$$\mathbf{b} = [g(\mu + y_1), \ldots, g(\mu + y_N)]^{\mathrm{T}} \qquad (80)$$

of length N satisfy $\mathbf{b} \preceq \tilde{\mathbf{b}}$. Since $\mathbf{1}^{\mathrm{T}}e^{-\mathbf{b}}$ is a Schur-convex function we therefore have $\mathbf{1}^{\mathrm{T}}e^{-\mathbf{b}} \leq \mathbf{1}^{\mathrm{T}}e^{-\tilde{\mathbf{b}}}$, and consequently

$$\log \sum_{i=1}^{K} \left(1 - e^{-g(\tilde{\mu} + y_i)}\right) \leq \log \sum_{i=1}^{N} \left(1 - e^{-g(\mu + y_i)}\right). \qquad (81)$$

This together with (79) proves the theorem.

---

[7]Symmetry is perhaps not apparent from (75). However, all second order derivatives of $J(\cdot)$ are continuous and consequently we know that the Hessian matrix (75) is symmetric. The interested reader can alternatively show symmetry of (75) by applying (66), but this requires a few extra steps of derivations.

REFERENCES

[1] E. Telatar, "Capacity of multi-antenna Gaussian channels," *Trans. Telecommun.*, vol. 10, no. 6, pp. 585–595, Nov./Dec. 1999, 1995 Tech. Memo., Bell Labs..

[2] D. P. Palomar, J. M. Cioffi, and M. A. Lagunas, "Joint Tx-Rx beamforming design for multicarrier MIMO channels: A unified framework for convex optimization," *IEEE Trans. Signal Process.*, vol. 51, no. 9, pp. 2381–2401, Sep. 2003.

[3] Y. Ding, T. N. Davidson, Z. Luo, and K. M. Wong, "Minimum BER block precoders for zero-forcing equalization," *IEEE Trans. Signal Process.*, vol. 51, no. 9, pp. 2410–2423, Sep. 2003.

[4] G. D. Jr. Forney and M. V. Eyuboglu, "Combined equalization and coding using precoding," *IEEE Commun. Mag.*, vol. 29, no. 12, pp. 25–34, Dec. 1991.

[5] A. J. Goldsmith and S. Chua, "Variable-rate variable-power MQAM for fading channels," *IEEE Trans. Commun.*, vol. 45, no. 10, pp. 1218–1230, Oct. 1997.

[6] D. P. Palomar and S. Barbarossa, "Designing MIMO communication systems: Constellation choice and linear transceiver design," *IEEE Trans. Signal Process.*, vol. 53, no. 10, pp. 3804–3818, Oct. 2005.

[7] S. Bergman and B. Ottersten, "Lattice based linear precoding for multicarrier block codes," *IEEE Trans. Signal Process.*, vol. 56, no. 7, pp. 2902–2914, Jul. 2008.

[8] L. Collin, O. Berder, P. Rostaing, and G. Burel, "Optimal minimum distance-based precoder for MIMO spatial multiplexing systems," *IEEE Trans. Signal Process.*, vol. 52, no. 3, pp. 617–627, Mar. 2004.

[9] C. A. Belfiore and J. H. Jr. Park, "Decision feedback equalization," *Proc. IEEE*, vol. 67, no. 8, pp. 1143–1156, Aug. 1979.

[10] G. Ginis and J. M. Cioffi, "On the relation between V-BLAST and the GDFE," *IEEE Commun. Lett.*, vol. 5, no. 9, pp. 364–366, Sep. 2001.

[11] P. W. Wolniansky, G. J. Foschini, G. D. Golden, and R. A. Valenzuela, "V-BLAST: An architecture for realizing very high data rates over the rich-scattering wireless channel," in *Proc. URSI Int. Symp. Signals, Systems, Electronics*, Oct. 1998, pp. 295–300.

[12] T. Guess, "Optimal sequences for CDMA with decision-feedback receivers," *IEEE Trans. Inf. Theory*, vol. 49, no. 4, pp. 886–900, Apr. 2003.

[13] A. Stamoulis, G. B. Giannakis, and A. Scaglione, "Block FIR decision-feedback equalizers for filterbank precoded transmissions with blind channel estimation capabilities," *IEEE Trans. Commun.*, vol. 49, no. 1, pp. 69–83, Jan. 2001.

[14] F. Xu, T. N. Davidson, J. Zhang, and K. M. Wong, "Design of block transceivers with decision feedback detection," *IEEE Trans. Signal Process.*, vol. 54, no. 3, pp. 965–978, Mar. 2006.

[15] Y. Jiang, W. W. Hager, and J. Li, "Tunable channel decomposition for MIMO communications using channel state information," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4405–4418, Nov. 2006.

[16] M. B. Shenouda and T. N. Davidson, "A framework for designing MIMO systems with decision feedback equalization or Tomlinson–Harashima precoding," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 2, pp. 401–411, Feb. 2008.

[17] D. P. Palomar and Y. Jiang, "MIMO transceiver design via majorization theory," *Found. Trends Commun. Inf. Theory*, vol. 3, no. 4–5, pp. 331–551, 2006.

[18] J. Jaldén and B. Ottersten, "On the complexity of sphere decoding in digital communications," *IEEE Trans. Signal Process.*, vol. 53, no. 4, pp. 1474–1484, Apr. 2005.

[19] J. G. Proakis, *Digital Communications*. New York: McGraw-Hill, 2001.

[20] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1991.

[21] Y. Jiang, W. W. Hager, and J. Li, "The generalized triangular decomposition," *Math. Comput.*, vol. 77, no. 262, pp. 1037–1056, Apr. 2008.

[22] A. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Applications*. New York: Academic, 1979.

[23] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[24] E. Jorswieck and H. Boche, "Majorization and matrix-monotone functions in wireless communications," *Found. Trends Commun. Inf. Theory*, vol. 3, pp. 553–701, 2006.

[25] S. Bergman, S. Järmyr, E. Jorswieck, and B. Ottersten, "Optimization with skewed majorization constraints: Application to MIMO systems," in *Proc. IEEE Symp. Personal, Indoor, Mobile Radio Commun. (PIMRC)*, Sep. 2008, pp. 1–6.

[26] S. Bergman, "Bit loading and precoding for MIMO communication systems with various receivers," Ph.D. dissertation, Royal Institute of Technology (KTH), Stockholm, Sweden, 2009, in preparation.

[27] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1985.

Svante Bergman (S'04) was born in Karlstad, Sweden, in 1979. His undergraduate studies where pursued at Chalmers University of Technology (1998–2000), Göteborg, KTH (2001–2002, 2004), and Stanford University (2002–2003). He received the M.S. degree in electrical engineering from the Royal Institute of Technology (KTH), Stockholm, Sweden, in 2004, where he is currently working towards the Ph.D. degree in telecommunications.

During 2000 and 2001, he was a software developer at room33 AB, Stockholm, Sweden. Since April 2004, he has been a member of the Signal Processing Laboratory at KTH. During spring 2008, he was a guest researcher at the ECE Department at the Hong Kong University of Science and Technology. His research interests are linear precoding and constellation adaptation schemes for MIMO and block based communication systems.

Daniel P. Palomar (S'99–M'03–SM'08) received the Electrical Engineering and Ph.D. degrees (both with honors) from the Technical University of Catalonia (UPC), Barcelona, Spain, in 1998 and 2003, respectively.

Since 2006, he has been an Assistant Professor in the Department of Electronic and Computer Engineering at the Hong Kong University of Science and Technology (HKUST), Hong Kong. He has held several research appointments, namely, at King's College London (KCL), London, U.K.; Technical University of Catalonia (UPC), Barcelona; Stanford University, Stanford, CA; Telecommunications Technological Center of Catalonia (CTTC), Barcelona, Spain; Royal Institute of Technology (KTH), Stockholm, Sweden; University of Rome "La Sapienza", Rome, Italy; and Princeton University, Princeton, NJ. His current research interests include applications of convex optimization theory, game theory, and variational inequality theory to signal processing and communications.

Dr. Palomar is an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING, a Guest Editor of the *IEEE Signal Processing Magazine* 2010 Special Issue on Convex Optimization for Signal Processing, was a Guest Editor of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS 2008 Special Issue on Game Theory in Communication Systems, as well as the lead Guest Editor of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS 2007 Special Issue on Optimization of MIMO Transceivers for Realistic Communication Networks. He serves on the IEEE Signal Processing Society Technical Committee on Signal Processing for Communications (SPCOM). He is a recipient of a 2004/06 Fulbright Research Fellowship; the 2004 Young Author Best Paper Award by the IEEE Signal Processing Society; the 2002–2003 best Ph.D. prize in information technologies and communications by the Technical University of Catalonia (UPC); the 2002–2003 Rosina Ribalta first prize for the Best Doctoral Thesis in Information Technologies and Communications by the Epson Foundation; and the 2004 prize for the best Doctoral Thesis in Advanced Mobile Communications by the Vodafone Foundation and COIT.

Björn Ottersten (S'86–M'89–SM'99–F'04) was born in Stockholm, Sweden, in 1961. He received the M.S. degree in electrical engineering and applied physics from Linköping University, Linköping, Sweden, in 1986 and the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA in 1989.

He has held research positions at the Department of Electrical Engineering, Linköping University; the Information Systems Laboratory, Stanford University; and the Katholieke Universiteit Leuven, Leuven. During 1996–1997, he was Director of Research at ArrayComm Inc, San Jose, CA, a start-up company based on his patented technology. In 1991, he was appointed Professor of signal processing at the Royal Institute of Technology (KTH), Stockholm, Sweden. From 2004 to 2008, he was Dean of the School of Electrical Engineering at KTH, and from 1992 to 2004 he was head of the Department for Signals, Sensors, and Systems at KTH. He is also Director of securityandtrust.lu at the University of Luxembourg. His research interests include wireless communications, stochastic signal processing, sensor array processing, and time series analysis.

Dr. Ottersten has coauthored papers that received an IEEE Signal Processing Society Best Paper Award in 1993, 2001, and 2006. He has served as Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING and on the Editorial Board of the *IEEE Signal Processing Magazine*. He is currently Editor-in-Chief of the *EURASIP Signal Processing Journal* and a member of the Editorial Board of the *EURASIP Journal of Advances Signal Processing*. He is a Fellow of the EURASIP. He is a first recipient of the European Research Council advanced research grant.