

Precoder Optimization for Nonlinear MIMO Transceiver Based on Arbitrary Cost Function

Yi Jiang* Daniel P. Palomar† Mahesh K. Varanasi*

Abstract—Assuming full channel state information (CSI) at both transmitter (CSIT) and receiver (CSIR), we consider optimizing a nonlinear MIMO transceiver with (nonlinear) decision feedback equalizer (DFE) with respect to some global cost function f_0 . Setting the receiver to be a minimum mean-squared error (MMSE) DFE, the MIMO transceiver optimization problem reduces to optimizing a linear precoder. Based on the generalized triangular decomposition (GTD) and majorization theory, we prove that for *any* cost function f_0 the optimum precoder is of the same special structure and hence the original complicated matrix optimization problem can be significantly simplified to an optimization problem with scalar-valued variables. Furthermore, if the cost function is specialized to the cases where the composite function $f_0 \circ \text{exp}$ is either Schur-convex or Schur-concave, then the nonlinear transceiver design becomes exceedingly simple. In particular, when $f_0 \circ \text{exp}$ is Schur-convex, the optimum nonlinear transceiver design turns out to be the uniform channel decomposition (UCD) scheme; when $f_0 \circ \text{exp}$ is Schur-concave, the optimum nonlinear design degenerates to linear diagonal transmission.

Index Terms—MIMO transceiver optimization, generalized triangular decomposition, majorization theory, Schur-convex.

I. INTRODUCTION

In a slowly varying channel, the channel state information (CSI) can be accurately estimated at the receiver (CSIR), and then the CSI is made available to the transmitter (CSIT) via feedback or the reciprocal principle when time division duplex (TDD) is used. Using the full CSI, one may jointly optimize transmitter and receiver for high rate and reliable communication. The research on this topic can be traced back to the seminal paper [1], which dealt with the joint transceiver design for a single-input single-output (SISO) inter-symbol interference (ISI) channel with pulse amplitude modulation. The paper [2] investigated the joint optimization of transmitter and DF receiver for SISO systems. The papers [3] and [4] are arguably the pioneering work on transceiver design for MIMO ISI channels, where the transceiver is optimized to minimize the trace and determinant of the mean-squared error (MSE) matrix based on a *linear* receiver [3] and a *nonlinear* DF receiver [4], respectively.

Along with the increased interest in MIMO communications, much research effort has been made on the MIMO transceiver designs since the late 1990s. Two paradigms of transceiver designs have been developed, namely, the *linear* designs (see

[5] and the references therein) and the *nonlinear* designs (see [6] and the references therein), corresponding to using a linear receiver or a (nonlinear) DF receiver.¹ Although the recent work are certainly related to [3][4], they are based different mathematical tools.

While the DFE based nonlinear designs were originally developed from a channel decomposition perspective [6], in this paper we derive the nonlinear MIMO transceivers by solving an optimization problem with respect to some global cost function. In doing so, we build interesting connections between the linear [5] and nonlinear designs within the unifying framework of majorization theory [7]. With a minimum mean-squared error (MMSE) DFE as the receiver, the optimum linear precoder is highly structured for an arbitrary cost function f_0 , based on which the original complicated matrix optimization problem can be significantly simplified to an optimization problem with scalar-valued variables. Furthermore, we consider the special cost functions for which the composite function $f_0 \circ \text{exp}$ is either Schur-convex or Schur-concave. We show that when $f_0 \circ \text{exp}$ is Schur-convex, the optimum nonlinear transceiver design turns out to be the uniform channel decomposition (UCD) scheme [8]; when $f_0 \circ \text{exp}$ is Schur-concave, the optimum nonlinear design degenerates to linear diagonal transmission.

The rest of this paper is organized as follows. Section II introduces the channel model and the closed-form representation of MMSE-DFE followed by the formulation of the precoder optimization problem. Section III deals with the precoder optimization with respect to an arbitrary cost function f_0 . The specialization of the cost function to the cases where $f_0 \circ \text{exp}$ is either Schur-convex or Schur-concave is discussed in Section IV. Section V presents some interesting examples of cost function. Section VI gives the conclusion of this paper.

II. CHANNEL MODEL AND PROBLEM FORMULATION

A. Channel Model

Consider a communication system with n_T transmit and n_R receive antennas in a frequency flat fading channel. The information symbol vector $\mathbf{x} \in \mathbb{C}^{L \times 1}$ is left-multiplied by a linear precoder $\mathbf{P} \in \mathbb{C}^{n_T \times L}$ to obtain $\mathbf{s} = \mathbf{P}\mathbf{x} \in \mathbb{C}^{n_T \times 1}$ which is transmitted through the n_T transmit antennas. At the receiver, the sampled baseband signal is

$$\mathbf{y} = \mathbf{H}\mathbf{P}\mathbf{x} + \mathbf{n}, \quad (1)$$

where $\mathbf{H} \in \mathbb{C}^{n_R \times n_T}$ is the channel matrix with rank K , and the additive noise \mathbf{n} is zero-mean circularly symmetric Gaussian with variance a scaled identity matrix. As the received signal

This work is supported in part by NSF Grants CCF-0423842 and CCF-0434410.

* The authors are with the Dept. of ECE, University of Colorado, Boulder, CO 80309-0425, USA (yjiang@dsp.ufl.edu; varanasi@colorado.edu).

† The author is with Dept. of ECE, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong (Palomar@ust.hk)

¹Besides the nonlinear design based on DF receiver, there is also the dual form using dirty paper codes [6].

can always be scaled, we assume without loss of generality that $\mathbb{E}[\mathbf{nn}^\dagger] = \mathbf{I}$, where $\mathbb{E}[\cdot]$ is the expectation, $(\cdot)^\dagger$ stands for conjugate transpose, and \mathbf{I} is an identity matrix. We also assume that $\mathbb{E}[\mathbf{xx}^\dagger] = \mathbf{I}$. Hence the input power of the system is

$$P_T = \mathbb{E}[\|\mathbf{s}\|^2] = \text{Tr}(\mathbf{P}^\dagger \mathbf{P}). \quad (2)$$

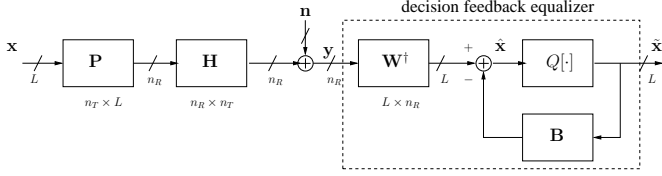


Fig. 1. Scheme of a MIMO Communication system with decision feedback equalizer (DFE) receiver

B. Nonlinear MIMO Transceiver Based on MMSE-DFE

We consider a MIMO transceiver structure given in Figure 1, where the transceiver structure applies DFE to detect the substreams *successively*. The DFE consists of a feed-forward matrix \mathbf{W} and a *strictly* upper triangular feed-backward matrix \mathbf{B} . The block $Q[\cdot]$ stands for mapping the “analog” estimate $\hat{\mathbf{x}}$ to the closest constellation point which yields the “digital” estimate $\tilde{\mathbf{x}}$. The “analog” estimate $\hat{\mathbf{x}}$ can be written as

$$\hat{\mathbf{x}} = \mathbf{W}^\dagger \mathbf{y} - \mathbf{B} \tilde{\mathbf{x}} \quad (3)$$

To simplify the system design and performance analysis, we invoke the usual assumption that $\tilde{\mathbf{x}} = \mathbf{x}$. Then the error vector

$$\hat{\mathbf{x}} - \mathbf{x} = (\mathbf{W}^\dagger \mathbf{H} \mathbf{P} - (\mathbf{B} + \mathbf{I})) \mathbf{x} + \mathbf{W}^\dagger \mathbf{n}, \quad (4)$$

and the MSE matrix is $\mathbf{E} = \mathbb{E}[(\hat{\mathbf{x}} - \mathbf{x})(\hat{\mathbf{x}} - \mathbf{x})^\dagger]$, for which the following lemma holds. (see, e.g., [8].)

Lemma 2.1: Let the QR decomposition of the augmented matrix be ²

$$\mathbf{G}_a \triangleq \begin{bmatrix} \mathbf{H} \mathbf{P} \\ \mathbf{I}_L \end{bmatrix}_{(n_R+L) \times L} = \mathbf{Q} \mathbf{R}. \quad (5)$$

Partition \mathbf{Q} into $\mathbf{Q} = \begin{bmatrix} \bar{\mathbf{Q}} \\ \underline{\mathbf{Q}} \end{bmatrix}$ where $\bar{\mathbf{Q}} \in \mathbb{C}^{n_R \times L}$ and $\underline{\mathbf{Q}} \in \mathbb{C}^{L \times L}$. The MSE matrix satisfies³

$$\mathbf{E} \geq \mathbf{D}_R^{-2} \quad (6)$$

where \mathbf{D}_R is a diagonal matrix with the same diagonal as \mathbf{R} and the equality of (6) holds if the feed-forward and feed-backward matrices are chosen to be

$$\mathbf{W} = \bar{\mathbf{Q}} \mathbf{D}_R^{-1} \quad \text{and} \quad \mathbf{B} = \mathbf{D}_R^{-1} \mathbf{R} - \mathbf{I}. \quad (7)$$

The DFE corresponding to the matrices given in (7) is called MMSE-DFE as it minimizes the MSE matrix. Indeed, the MMSE-DFE decomposes a MIMO channel into L subchannels with output SINRs [9]

$$\text{SINR}_i = \frac{1}{\text{MSE}_i} - 1 = [\mathbf{R}]_{ii}^2 - 1, \quad 1 \leq i \leq L, \quad (8)$$

where $\text{MSE}_i = [\mathbf{E}]_{ii}$ is the i th diagonal entry of \mathbf{E} . With Gaussian codebooks, each subchannels are subject to Gaussian interference-plus-noise, which have capacities

$$R_i = \log(1 + \text{SINR}_i) = \log[\mathbf{R}]_{ii}^2. \quad (9)$$

²To make the QR decomposition unique, the diagonal of \mathbf{R} is set positive.

³By $\mathbf{A} \geq \mathbf{B}$, we mean that $\mathbf{A} - \mathbf{B}$ is positive semidefinite.

C. Problem Formulation

In this paper, we shall consider optimizing the precoder with respect to a global performance measure subject to the overall input power constraint, which can be represented in the following generic form:

$$\begin{aligned} & \underset{\mathbf{P}}{\text{minimize}} && f_0(\{\text{MSE}_i\}) \\ & \text{subject to} && \text{Tr}(\mathbf{P} \mathbf{P}^\dagger) \leq P_0 \end{aligned} \quad (10)$$

where the cost function $f_0(\{\text{MSE}_i\})$ is chosen based on some criterion of practical significance. By Lemma 2.1, $\text{MSE}_i = [\mathbf{R}]_{ii}^{-2}$. Hence we can reformulate (10) to be

$$\begin{aligned} & \underset{\mathbf{P}}{\text{minimize}} && f_0(\{[\mathbf{R}]_{ii}^{-2}\}) \\ & \text{subject to} && \begin{pmatrix} \mathbf{H} \mathbf{P} \\ \mathbf{I} \end{pmatrix} = \mathbf{Q} \mathbf{R} \\ & && \text{Tr}(\mathbf{P} \mathbf{P}^\dagger) \leq P_0. \end{aligned} \quad (11)$$

To solve the above matrix variable optimization problem, we need the following mathematical concepts.

D. Majorization, Schur-convex Function, and GTD

Definition 2.2: [7, 1.A.1] The vector $\mathbf{a} \in \mathbb{R}^n$ is said to be *additively majorized* by $\mathbf{b} \in \mathbb{R}^n$, denoted by $\mathbf{a} \prec_+ \mathbf{b}$, if

$$\begin{aligned} \sum_{i=1}^k a_{[i]} &\leq \sum_{i=1}^k b_{[i]}, \quad 1 \leq k < n, \\ \sum_{i=1}^n a_i &= \sum_{i=1}^n b_i, \end{aligned} \quad (12)$$

where $a_{[i]}$ ($b_{[i]}$) is the i th largest element of \mathbf{a} (\mathbf{b}).

Definition 2.3: The positive vector $\mathbf{a} \in \mathbb{R}_+^n$ ⁴ is *multiplicatively majorized* by $\mathbf{b} \in \mathbb{R}_+^n$, denoted by $\mathbf{a} \prec_\times \mathbf{b}$, if

$$\begin{aligned} \prod_{i=1}^k a_{[i]} &\leq \prod_{i=1}^k b_{[i]}, \quad 1 \leq k < n, \\ \prod_{i=1}^n a_i &= \prod_{i=1}^n b_i. \end{aligned} \quad (13)$$

Clearly, $\mathbf{a} \prec_+ \mathbf{b}$ if and only if $\exp(\mathbf{a}) \prec_\times \exp(\mathbf{b})$.

Definition 2.4: [7, 3.A.1] A real-valued function ϕ defined on a set $\mathcal{A} \subseteq \mathbb{R}^n$ is said to be Schur-convex on \mathcal{A} if

$$\mathbf{x} \prec_+ \mathbf{y} \text{ on } \mathcal{A} \Rightarrow \phi(\mathbf{x}) \leq \phi(\mathbf{y}).$$

Similarly, ϕ is said to be Schur-concave on \mathcal{A} if

$$\mathbf{x} \prec_+ \mathbf{y} \text{ on } \mathcal{A} \Rightarrow \phi(\mathbf{x}) \geq \phi(\mathbf{y}).$$

The so-called generalized triangular decomposition (GTD) theorem and its fast algorithm are established in [10].

Theorem 2.5 (GTD Theorem): Let $\mathbf{H} \in \mathbb{C}^{m \times n}$ be a matrix with rank K and singular values $\sigma_{H,1} \geq \sigma_{H,2} \geq \dots \geq \sigma_{H,K} > 0$. There exists an upper triangular matrix $\mathbf{R} \in \mathbb{C}^{K \times K}$ and semi-unitary matrices \mathbf{Q} and \mathbf{P} such that $\mathbf{H} = \mathbf{Q} \mathbf{R} \mathbf{P}^\dagger$ if and only if the diagonal elements of \mathbf{R} satisfy $|\mathbf{r}| \prec_\times \boldsymbol{\sigma}_H$.⁵ Here $|\mathbf{r}|$ is a vector with the absolute values of \mathbf{r} element-wise.

An interesting special case of the GTD is the geometric mean decomposition (GMD) which yields the upper triangular matrix \mathbf{R} with equal diagonal elements:

$$[\mathbf{R}]_{ii} = \bar{\sigma}_H \triangleq \left(\prod_{i=1}^K \sigma_{H,i} \right)^{\frac{1}{K}}, \quad 1 \leq i \leq K \quad (14)$$

which is the *geometric mean* of the nonzero singular values of \mathbf{H} . The existence of the GMD follows immediately from Theorem 2.5 since $\bar{\sigma}_H \mathbf{1} \prec_\times \boldsymbol{\sigma}_H$.

⁴We denote $x \in \mathbb{R}_+$ if $x \in [0, \infty)$.

⁵We have adopted the notation used in the original paper [6]. The notations \mathbf{Q} , \mathbf{R} , and \mathbf{P} are certainly *not* the ones given in (5) or the precoder matrix.

III. PRECODER OPTIMIZATION FOR ARBITRARY COST FUNCTION

In this section, we derive a general solution to the optimization problem (11) as summarized in the following theorem.

Theorem 3.1: The solution to (11) has the form $\mathbf{P} = \mathbf{V}_H \boldsymbol{\Sigma} \boldsymbol{\Omega}^\dagger$ where $\mathbf{V}_H \in \mathbb{C}^{n_T \times K}$ appears in the SVD $\mathbf{H} = \mathbf{U}_H \boldsymbol{\Sigma}_H \mathbf{V}_H^\dagger$, and $\boldsymbol{\Sigma} = \text{diag}(\sqrt{\mathbf{p}})$, where $\mathbf{p} \in \mathbb{R}_+^K$ is the solution to the optimization problem:

$$\begin{aligned} & \underset{\mathbf{p}, \{[\mathbf{R}]_{ii}^2\}}{\text{minimize}} && f_0(\{[\mathbf{R}]_{ii}^{-2}\}) \\ & \text{subject to} && ([\mathbf{R}]_{11}^2, \dots, [\mathbf{R}]_{LL}^2) \prec_{\times} \\ & && \left(\{1 + \sigma_{H,i}^2 p_i\}_{i=1}^K, 1, \dots, 1 \right) \\ & && \mathbf{1}^T \mathbf{p} \leq P_0 \\ & && \mathbf{p} \geq \mathbf{0}, \end{aligned} \quad (15)$$

where $(\{1 + \sigma_{H,i}^2 p_i\}_{i=1}^K, 1, \dots, 1)$ denotes an L -dimensional vector whose first K elements are $1 + \sigma_{H,i}^2 p_i$, $i = 1, \dots, K$ followed by $L - K$ ones. The semi-unitary matrix $\boldsymbol{\Omega} \in \mathbb{C}^{L \times K}$ is obtained such that the matrix \mathbf{R} in the QR decomposition

$$\begin{pmatrix} \mathbf{H}\mathbf{P} \\ \mathbf{I} \end{pmatrix} = \begin{pmatrix} \mathbf{U}_H \boldsymbol{\Sigma}_H \boldsymbol{\Sigma} \boldsymbol{\Omega}^\dagger \\ \mathbf{I} \end{pmatrix} = \mathbf{Q}\mathbf{R} \quad (16)$$

has diagonal $\{[\mathbf{R}]_{ii}\}$ being the solution to (15).

Proof: Let us denote the singular values of the effective channel $\mathbf{G} \triangleq \mathbf{H}\mathbf{P} \in \mathbb{C}^{n_R \times L}$ as $\sigma_{G,1} \geq \dots \geq \sigma_{G,K} \geq \sigma_{G,K+1} = \dots = \sigma_{G,L} = 0$. The singular values of the augmented matrix $\mathbf{G}_a = \begin{pmatrix} \mathbf{H}\mathbf{P} \\ \mathbf{I} \end{pmatrix} \in \mathbb{C}^{(n_R+L) \times L}$ are

$$\sigma_{G_a,i} = \begin{cases} \sqrt{1 + \sigma_{G,i}^2}, & 1 \leq i \leq K, \\ 1, & K < i \leq L. \end{cases} \quad (17)$$

Let $\boldsymbol{\sigma}_{G_a} \in \mathbb{R}_+^L$ be the vector consisting of $\sigma_{G_a,i}$. Clearly $\boldsymbol{\sigma}_{G_a}$ is invariant to the choice of $\boldsymbol{\Omega}$. However, $\boldsymbol{\Omega}$ determines the diagonal of \mathbf{R} as shown in the following lemma.

Lemma 3.1: There exists an $\boldsymbol{\Omega}$ in (16) such that $\mathbf{G}_a = \mathbf{Q}\mathbf{R}$ with $[\mathbf{R}]_{ii}$, $1 \leq i \leq L$, being the diagonal of \mathbf{R} if and only if

$$\{ |[\mathbf{R}]_{ii}|^2 \} \prec_{\times} \boldsymbol{\sigma}_{G_a}^2. \quad (18)$$

In other words, when $\boldsymbol{\Omega}$ goes over the whole Stiefel manifold

$$\mathcal{S}(L; K) \triangleq \{ \mathbf{Q} \in \mathbb{C}^{L \times K} : \mathbf{Q}^\dagger \mathbf{Q} = \mathbf{I} \}, \quad (19)$$

the achievable set of the diagonal of \mathbf{R} given in (16) is

$$\{ \mathbf{r} \in \mathbb{C}^L : |\mathbf{r}| \prec_{\times} \boldsymbol{\sigma}_{G_a} \}. \quad (20)$$

Proof: This lemma is a straightforward corollary of Theorem III.1 in [6]. \blacksquare

Using the above lemma, we can replace the constraint $\begin{pmatrix} \mathbf{H}\mathbf{P} \\ \mathbf{I} \end{pmatrix} = \mathbf{Q}\mathbf{R}$ in (11) by $([\mathbf{R}]_{11}^2, \dots, [\mathbf{R}]_{LL}^2) \prec_{\times} \boldsymbol{\sigma}_{G_a}^2$. Let $\mathbf{P} = \mathbf{U}\boldsymbol{\Sigma}\boldsymbol{\Omega}^\dagger$ be the SVD, where $\boldsymbol{\Sigma} = \text{diag}(\sqrt{\mathbf{p}})$. Note that $\mathbf{P}\mathbf{P}^\dagger$ is invariant to $\boldsymbol{\Omega}$. We can at this time remove $\boldsymbol{\Omega}$ from the optimization problem and simplify (11) as

$$\begin{aligned} & \underset{\mathbf{U}, \mathbf{p}, [\mathbf{R}]_{ii}^2}{\text{minimize}} && f_0(\{[\mathbf{R}]_{ii}^{-2}\}) \\ & \text{subject to} && ([\mathbf{R}]_{11}^2, \dots, [\mathbf{R}]_{LL}^2) \prec_{\times} \boldsymbol{\sigma}_{G_a}^2 \\ & && \mathbf{1}^T \mathbf{p} \leq P_0 \\ & && \mathbf{p} \geq \mathbf{0}. \end{aligned} \quad (21)$$

Note that $\boldsymbol{\sigma}_{G_a}$ depends on \mathbf{p} and \mathbf{U} . If we denote $r_{[i]}$ the i th largest element of $\{[\mathbf{R}]_{ii}^2\}_{i=1}^L$, (21) can be rewritten more explicitly as

$$\begin{aligned} & \underset{\mathbf{U}, \mathbf{p}, [\mathbf{R}]_{ii}^2}{\text{minimize}} && f_0(\{[\mathbf{R}]_{ii}^{-2}\}) \\ & \text{subject to} && \prod_{i=1}^k r_{[i]} \leq \prod_{i=1}^k \sigma_{G_a,i}^2, \quad 1 \leq i \leq K-1 \\ & && \prod_{i=1}^L r_{[i]} = \prod_{i=1}^K \sigma_{G_a,i}^2 \\ & && \mathbf{1}^T \mathbf{p} \leq P_0 \\ & && \mathbf{p} \geq \mathbf{0}. \end{aligned} \quad (22)$$

Next, we show that the solution to (21) occurs when $\mathbf{U} = \mathbf{V}_H$. Denote $\{\sigma_{H,i}\}$ and $\{\sqrt{p_i}\}$ as the singular values of \mathbf{H} and \mathbf{P} , and both are in non-increasing ordering. By [7, Theorem 3.3.14], the singular values of the product of two matrices \mathbf{H} and \mathbf{P} are multiplicatively majorized by the product of the singular values of \mathbf{H} and \mathbf{P} . Thus,

$$\begin{aligned} \prod_{i=1}^k \sigma_{G,i}^2 &\leq \prod_{i=1}^k \sigma_{H,i}^2 p_i, \quad 1 \leq i \leq K-1 \\ \prod_{i=1}^K \sigma_{G,i}^2 &= \prod_{i=1}^K \sigma_{H,i}^2 p_i, \end{aligned} \quad (23)$$

where the equality holds if and only if $\mathbf{U} = \mathbf{V}_H$. Denote $\mathbf{x}, \mathbf{y} \in \mathbb{R}^K$ by

$$x_i \triangleq \log(\sigma_{G,i}^2), \quad y_i \triangleq \log(\sigma_{H,i}^2 p_i), \quad 1 \leq i \leq K.$$

Then $\mathbf{x} \prec_+ \mathbf{y}$, and $\mathbf{x} = \mathbf{y}$ if and only if $\mathbf{U} = \mathbf{V}_H$. It is easy to prove that $\log(1 + \exp(x))$ is an increasing convex function. Hence $\sum_{i=1}^K \log(1 + \exp(x_i))$ is a Schur-convex function of \mathbf{x} , and we obtain

$$\sum_{i=1}^k \log(1 + \exp(x_i)) \leq \sum_{i=1}^k \log(1 + \exp(y_i)), \quad 1 \leq k \leq K, \quad (24)$$

and equivalently

$$\prod_{i=1}^k (1 + \sigma_{G,i}^2) \leq \prod_{i=1}^k (1 + \sigma_{H,i}^2 p_i), \quad 1 \leq k \leq K. \quad (25)$$

Since $\sigma_{G_a,i}^2 = 1 + \sigma_{G,i}^2$ (see (17)), it follows from (25) and (22) that given \mathbf{p} , the feasible set of $\{[\mathbf{R}]_{ii}^2\}_{i=1}^L$ is relaxed to the maximal extent when $\mathbf{U} = \mathbf{V}_H$. More precisely, if \mathbf{p} and $[\mathbf{R}]_{ii}^2$'s are feasible for some $\mathbf{U} \neq \mathbf{V}_H$, they must also be feasible for $\mathbf{U} = \mathbf{V}_H$. Hence the optimum solution occurs when $\mathbf{U} = \mathbf{V}_H$. With $\mathbf{P} = \mathbf{V}_H \boldsymbol{\Sigma} \boldsymbol{\Omega}^\dagger$,

$$\sigma_{G,i}^2 = \sigma_{H,i}^2 p_i, \quad \text{for } 1 \leq i \leq K \quad (26)$$

and (22) is simplified to be (15). The theorem is proven. \blacksquare

It is worth emphasizing that the above proof does not make any assumption on f_0 . Indeed, given the MMSE-DFE, the precoder matrix has the structure $\mathbf{P} = \mathbf{V}_H \boldsymbol{\Sigma} \boldsymbol{\Omega}^\dagger$ for any cost function f_0 .

Given \mathbf{p} and the diagonal of \mathbf{R} as the solution to (15), we only need to calculate $\boldsymbol{\Omega}$ such that the QR decomposition in (16) yields \mathbf{R} with such a diagonal. Denoting $\boldsymbol{\Omega}_0$ the unitary matrix whose first K columns form $\boldsymbol{\Omega}$, we rewrite the second term of (16) as

$$\begin{bmatrix} \mathbf{I}_{n_R} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Omega}_0 \end{bmatrix} \begin{bmatrix} \mathbf{U}_H \boldsymbol{\Sigma}_H [\boldsymbol{\Sigma} : \mathbf{0}_{K \times (L-K)}] \\ \mathbf{I}_L \end{bmatrix} \boldsymbol{\Omega}_0^\dagger. \quad (27)$$

Calculate the GTD

$$\mathbf{J} \triangleq \begin{bmatrix} \mathbf{U}_H \boldsymbol{\Sigma}_H [\boldsymbol{\Sigma} : \mathbf{0}_{K \times (L-K)}] \\ \mathbf{I}_L \end{bmatrix} = \mathbf{Q}_J \mathbf{R} \mathbf{P}_J^\dagger, \quad (28)$$

where \mathbf{R} has diagonal $\{[\mathbf{R}]_{ii}\}$ being the solution to (15). The existence of such a decomposition follows from the fact that the singular values of \mathbf{J} ,

$$\sigma_{J,i} = \begin{cases} \sqrt{1 + \sigma_{H,i}^2 p_i} & 1 \leq i \leq K \\ 1 & K+1 \leq i \leq L, \end{cases} \quad (29)$$

majorize $\{[\mathbf{R}]_{ii}\}$ (cf. the constraint of (15)). Letting $\boldsymbol{\Omega}_0 = \mathbf{P}_J^\dagger$ and hence $\boldsymbol{\Omega}^\dagger$ are formed by the first K rows of \mathbf{P}_J and

$$\mathbf{Q} = \begin{bmatrix} \mathbf{I}_{n_R} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Omega}_0 \end{bmatrix} \mathbf{Q}_J, \quad (30)$$

we obtain the QR decomposition (16). According to Lemma 2.1, the DFE feed-forward and feed-backward matrices are:

$$\mathbf{W} = \bar{\mathbf{Q}} \mathbf{D}_R^{-1} \quad \text{and} \quad \mathbf{B} = \mathbf{D}_R^{-1} \mathbf{R} - \mathbf{I}, \quad (31)$$

where $\bar{\mathbf{Q}}$ is the matrix consisting of the first n_R rows of \mathbf{Q} (or \mathbf{Q}_J).

IV. SCHUR-CONVEX AND SCHUR-CONCAVE FUNCTIONS

In this section we further specialize the cost function to the case where f_0 is increasing in each argument and the composite function $f_0 \circ \exp : \mathbb{R}^L \rightarrow \mathbb{R}$ is either Schur-convex or Schur-concave. Here the composite function is defined as

$$f_0 \circ \exp(\mathbf{x}) \triangleq f_0(e^{x_1}, e^{x_2}, \dots, e^{x_L}). \quad (32)$$

Such specialization leads to exceedingly simple solution to (11) as shown in the following theorem.

Theorem 4.1: An optimal solution \mathbf{P} of the problem (11), where $f_0 : \mathbb{R}^L \rightarrow \mathbb{R}$ is a function increasing in each argument, can be characterized as follows:

- If $f_0 \circ \exp$ is Schur-concave on $\mathcal{D}_L \triangleq \{\mathbf{x} \in \mathbb{R}^L : x_1 \geq \dots \geq x_L\}$, then

$$\mathbf{P} = \mathbf{V}_H \boldsymbol{\Sigma} \quad (33)$$

where $\boldsymbol{\Sigma} = \text{diag}(\sqrt{\mathbf{p}})$ with $\mathbf{p} \in \mathbb{R}_+^K$ being the solution to

$$\begin{aligned} & \underset{\mathbf{p}}{\text{minimize}} && f_0 \left(\left\{ \frac{1}{1 + \sigma_{H,i}^2 p_i} \right\} \right) \\ & \text{subject to} && \mathbf{1}^T \mathbf{p} \leq P_0 \\ & && \mathbf{p} \geq \mathbf{0}, \end{aligned} \quad (34)$$

where $\sigma_{H,i}$ is the i th largest singular values of \mathbf{H} . In this case $L \leq K$ otherwise some argument of $f_0 \circ \exp$ should be set to zero.

- If $f_0 \circ \exp$ is Schur-convex on \mathbb{R}^L , then

$$\mathbf{P} = \mathbf{V}_H \boldsymbol{\Sigma} \boldsymbol{\Omega}^\dagger \quad (35)$$

where $\boldsymbol{\Sigma} = \text{diag}(\sqrt{\mathbf{p}})$ with $\mathbf{p} \in \mathbb{R}_+^K$ being obtained via standard waterfilling power allocation

$$p_i = \left(\mu - \frac{1}{\sigma_{H,i}^2} \right)^+, \quad 1 \leq i \leq K, \quad (36)$$

where μ is chosen such that $\sum_{i=1}^K p_i = P_0$. The semi-unitary matrix $\boldsymbol{\Omega}$ is chosen so that the QR decomposition

$\begin{bmatrix} \mathbf{H} \mathbf{P} \\ \mathbf{I} \end{bmatrix} = \mathbf{Q} \mathbf{R}$ yields \mathbf{R} with equal diagonal elements.

In this case, L is not limited by the dimensionality of \mathbf{H} .

Proof: Based on Theorem 3.1, we start with the optimization problem (15). Let $R_i \triangleq \log[\mathbf{R}]_{ii}^2$. (In fact R_i stands for the capacity of the i th subchannel. See (9).) Equation (15) can be rewritten as

$$\begin{aligned} & \underset{\mathbf{p}, \{R_i\}}{\text{minimize}} && f_0(\{e^{-R_i}\}) \\ & \text{subject to} && (R_1, \dots, R_L) \prec_+ (\{\log(1 + \sigma_{H,i}^2 p_i)\}_{i=1}^K, 0, \dots, 0) \\ & && \mathbf{1}^T \mathbf{p} \leq P_0 \\ & && \mathbf{p} \geq \mathbf{0}. \end{aligned} \quad (37)$$

Note that $f(\mathbf{x})$ and $f(-\mathbf{x})$ have the same Schur-convexity/concavity. Thus, if $f_0 \circ \exp$ is Schur-concave, then $f_0(\{e^{-R_i}\})$ is a Schur-concave function of (R_1, \dots, R_L) . By Definition 2.4, the cost function is minimized when

$$R_i = \begin{cases} \log(1 + \sigma_{H,i}^2 p_i) & 1 \leq i \leq K \\ 0 & i \geq K, \end{cases} \quad (38)$$

which corresponds to $\boldsymbol{\Omega} = [\mathbf{I}_K : \mathbf{0}_{K \times (L-K)}]^T$. For $L > K$, there are subchannels with capacity $R_i = 0$, which is of course meaningless. Therefore in this case it should be constrained that $L \leq K$. Now (37) can be simplified to be

$$\begin{aligned} & \underset{\mathbf{p}}{\text{minimize}} && f_0 \left(\left\{ \frac{1}{1 + \sigma_{H,i}^2 p_i} \right\} \right) \\ & \text{subject to} && \mathbf{1}^T \mathbf{p} \leq P_0 \\ & && \mathbf{p} \geq \mathbf{0}. \end{aligned} \quad (39)$$

If $f_0 \circ \exp$ is Schur-convex, $f_0(\{e^{-R_i}\})$ is a Schur-convex function of (R_1, \dots, R_L) . Hence the solution to (37) occurs when

$$R_i = \bar{R} \triangleq \frac{1}{L} \sum_{i=1}^K \log(1 + \sigma_{H,i}^2 p_i) \quad \text{for } 1 \leq i \leq L. \quad (40)$$

Since f_0 is an increasing function of each argument, we further simplify (37) to be

$$\begin{aligned} & \underset{\mathbf{p}}{\text{maximize}} && \sum_{i=1}^K \log(1 + \sigma_{H,i}^2 p_i) \\ & \text{subject to} && \mathbf{1}^T \mathbf{p} \leq P_0 \\ & && \mathbf{p} \geq \mathbf{0}. \end{aligned} \quad (41)$$

The solution is the standard waterfilling power allocation given in (36). The semi-unitary matrix $\boldsymbol{\Omega}$ is chosen such that the diagonal elements of \mathbf{R} in (16) are equal, i.e.,

$$[\mathbf{R}]_{ii} = \exp(R_i) = \left(\prod_{k=1}^K (1 + \sigma_{H,i}^2 p_k) \right)^{\frac{1}{L}}, \quad \text{for } 1 \leq i \leq L. \quad (42)$$

According to Lemma 3.1, such an $\boldsymbol{\Omega}$ exists. The theorem is proven. \blacksquare

Some comments on the connections between the nonlinear MIMO transceiver designs and the linear designs are in order. For the first case in Theorem 4.1, the precoder $\mathbf{P} = \mathbf{V}_H \boldsymbol{\Sigma}$ orthogonalizes the MIMO channel into multiple eigen-subchannels and the power allocation applied to the subchannels are given in (39). In this case, the optimal nonlinear transceiver degenerates into *linear* diagonal transmission. In fact this transceiver design is the same as the first case in [5,

Theorem 1] where linear MIMO transceivers are considered. However, the condition here is that the composite function $f_0 \circ \exp$ be Schur-concave, while the condition in [5, Theorem 1] is that f_0 be Schur-concave. The relationship between the two conditions is shown in the following lemma.

Lemma 4.2: If $f_0 \circ \exp$ is Schur-concave on $\mathcal{D}_n = \{\mathbf{x} \in \mathbb{R}^n : x_1 \geq \dots \geq x_n\}$, then f_0 is Schur-concave on \mathcal{D}_n .

But the other direction is not true, i.e., f_0 being Schur-concave does *not* imply the Schur-concavity of $f_0 \circ \exp$, as we will see counter-examples in Section V-A. Therefore, if $f_0 \circ \exp$ is Schur-concave, then the optimum nonlinear transceiver degenerates into a linear one, which performs a diagonal transmission. However, if f_0 is Schur-concave, then the optimum *linear* transceiver must use a diagonal transmission, but for the optimum *nonlinear* transceiver it may not be the case.

Consider now the second case in Theorem 4.1. For *any* cost function such that $f_0 \circ \exp$ is Schur-convex, the optimum *nonlinear* MIMO transceiver design is the same. To see the relationship between Theorem 4.1 and [5, Theorem 1], the following lemma is useful [7, 3.B.2].

Lemma 4.3: The composite function $f_0(g(x_1), \dots, g(x_n))$ is Schur-convex if $f_0 : \mathbb{R}^n \rightarrow \mathbb{R}$ is Schur-convex and $g : \mathbb{R} \rightarrow \mathbb{R}$ is convex.

The other direction of Lemma 4.3 is not true, as we shall see counter examples in Section V-A. An immediate corollary of Lemma 4.3 is that $f_0 \circ \exp$ is Schur-convex if f_0 is Schur-convex. Therefore if f_0 is Schur-convex, which implies that the optimum linear transceiver uses non-diagonal transmission according to [5, Theorem 1], then $f_0 \circ \exp$ is also Schur-convex, but not vice versa. One of the major distinctions between the linear and nonlinear design is on the power allocation applied to the eigen-subchannels. For the nonlinear design, the standard capacity-achieving waterfilling power allocation is applied,

$$p_i = \left(\mu - \sigma_{H,i}^{-2} \right)^+. \quad (43)$$

For the linear design, however, the power allocation is the MMSE waterfilling solution [5]

$$p_i = \left(\mu \sigma_{H,i}^{-1} - \sigma_{H,i}^{-2} \right)^+, \quad (44)$$

which yields suboptimal mutual information.

Figure 2 illustrates the relationship between the sets of Schur-convex/concave functions as well as the sets of functions such that $f \circ \exp$ is Schur-convex/concave.

From (42) and (9) we see that the nonlinear MIMO transceiver design yields L subchannels with the same capacity

$$C_{\text{ucd}} = \frac{1}{L} \sum_{i=1}^K \log(1 + p_i \sigma_{H,i}^2) = \frac{C}{L}. \quad (45)$$

where C is the capacity the MIMO channel. We refer to this MIMO transceiver design as the uniform channel decomposition (UCD) scheme [8], as it uniformly decomposes, in an information lossless manner, a MIMO channel into L *identical* subchannels.

Roughly speaking, for the cost function that $f_0 \circ \exp$ is Schur-concave, the system is optimized by making the MSEs of the substreams as ‘‘spread out’’ as possible, for which we obtain

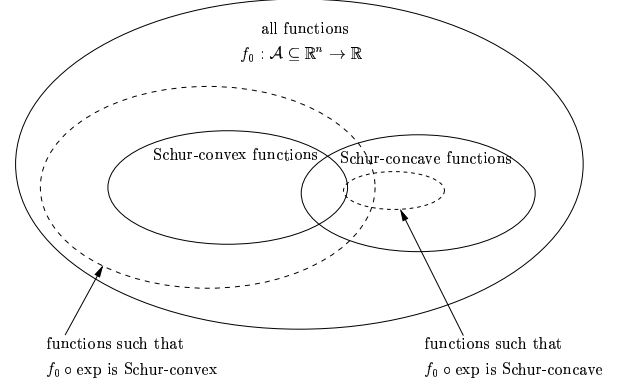


Fig. 2. The illustration of the sets of Schur-convex and Schur-concave functions as well as the functions such that $f \circ \exp$ is Schur-convex and Schur-concave function.

diagonal transmission. For the cost function that $f_0 \circ \exp$ is Schur-convex, however, the system is optimized when the MSEs of substreams are made as ‘‘close’’ to each other as possible, for which case we obtain L substreams with uniform MSEs.

V. EXAMPLES OF COST FUNCTIONS

A. Examples of Cost Function Leading to UCD

We study some examples of cost functions for which UCD is the optimal solution. According to [5], all the following problems can be formulated in terms of a Schur-convex cost function f_0 .

- Minimization of the maximum of the MSEs
- Maximization of the harmonic mean of the SINRs
- Maximization of the minimum of the SINRs
- Minimization of the average BER
- Minimization of the maximum of the BERs

Hence $f_0 \circ \exp$ is also Schur-convex according to Lemma 4.3. Even for some Schur-concave function f_0 , the composite $f_0 \circ \exp$ is Schur-convex. The following are some examples.

1) *Minimization of the sum of MSEs:* The cost function is

$$f_0(\{\text{MSE}_i\}) = \sum_{i=1}^L \text{MSE}_i, \quad (46)$$

which is both Schur-concave and Schur-convex. The composite function $f_0 \circ \exp(\mathbf{x}) = \sum_i e^{x_i}$ is Schur-convex (but not Schur-concave) since e^x is a convex function. Hence according to Theorem 4.1, the optimum solution is the UCD scheme. Indeed, using the relationship $\text{MSE}_i = [\mathbf{R}]_{ii}^{-2} = \exp(-R_i)$, the cost function

$$f_0(\{\exp(-R_i)\}) = \sum_{i=1}^L \exp(-R_i), \quad (47)$$

is minimized when $R_1 = \dots = R_L = C/L$, where C is the MIMO channel capacity.

2) *Maximization of the product of SINRs:* The objective function is to maximize $\prod_{i=1}^L \left(\frac{1}{\text{MSE}_i} - 1 \right)$. Hence the cost function to minimize is

$$f_0(\{\text{MSE}_i\}) = - \prod_{i=1}^L \left(\frac{1}{\text{MSE}_i} - 1 \right), \quad (48)$$

which is Schur-concave (see [5, Lemma 8]). The composite function $f_0 \circ \exp(\mathbf{x}) = -\prod_{i=1}^L (e^{-x_i} - 1)$ is, however, Schur-convex. To prove this point, we note that $\sum_{i=1}^L \log(e^{-x_i} - 1)$ is Schur-concave, as the second order derivative:

$$\frac{\partial^2 \log(e^x - 1)}{\partial x^2} = -\frac{e^x}{(e^x - 1)^2} \leq 0. \quad (49)$$

Hence $\prod_{i=1}^L (e^{-x_i} - 1) = \exp\left(\sum_{i=1}^L \log(e^{-x_i} - 1)\right)$ is Schur-concave. Therefore $f_0 \circ \exp(\mathbf{x}) = -\prod_{i=1}^L (e^{-x_i} - 1)$ is Schur-convex.

B. Examples of Cost Function Leading to Diagonal Transmission

The following are two examples of cost function for which $f_0 \circ \exp$ is Schur-concave and the nonlinear design degenerates to linear diagonal transmission.

1) *Minimization of the exponentially weighted product of MSEs*: The cost function is

$$f_0(\{\text{MSE}_i\}) = \prod_{i=1}^L \text{MSE}_i^{\alpha_i}. \quad (50)$$

Without loss of generality, it is assumed that $0 < \alpha_1 \leq \dots \leq \alpha_L$. The composite function $f_0 \circ \exp$ is

$$f_0 \circ \exp(\mathbf{x}) = \exp\left(\sum_{i=1}^L \alpha_i x_i\right). \quad (51)$$

It is easy to prove that $\sum_{i=1}^L \alpha_i x_i$ (assuming $\alpha_i \leq \alpha_{i+1}$) is a Schur-concave function on $\mathcal{D}_L \triangleq \{\mathbf{x} \in \mathbb{R}^L : x_1 \geq \dots \geq x_L\}$, so is $\exp\left(\sum_{i=1}^L \alpha_i x_i\right)$. In this case, the optimum nonlinear DFE based transceiver design degenerates to (linear) diagonal transmission.

2) *Maximization of the weighted sum of SINRs*: The objective function to maximize is $\sum_{i=1}^L \alpha_i \text{SINR}_i$, where $\alpha_1 \leq \dots \leq \alpha_L$. Then the cost function to minimize is

$$f_0(\{\text{MSE}_i\}) = -\sum_{i=1}^L \alpha_i \left(\frac{1}{\text{MSE}_i} - 1\right), \quad (52)$$

and the composite function $f_0 \circ \exp$ is

$$f_0 \circ \exp(\mathbf{x}) = -\sum_{i=1}^L \alpha_i (\exp(-x_i) - 1). \quad (53)$$

It can be readily shown that the function $f_0 \circ \exp(\mathbf{x})$ (assuming $\alpha_i \leq \alpha_{i+1}$) is minimized when $\mathbf{x} \in \mathcal{D}_L$. Note that for $\mathbf{x} \in \mathcal{D}_L$ and $\alpha_i \leq \alpha_{i+1}$, the derivative $\frac{\partial f_0 \circ \exp(\mathbf{x})}{\partial x_i} = \alpha_i \exp(-x_i)$ is increasing in $i = 1, \dots, L$. By [7, Theorem 3.A.3], $f_0 \circ \exp(\mathbf{x})$ is a Schur-concave function on \mathcal{D}_L .

C. Other Cost Functions

There do exist cost functions which are neither Schur-convex nor Schur-concave. For such cost functions, we have to solve the more complicated problem (15). We present two examples with no proof due to space constraint.

1) *Minimization of the weighted sum of MSEs*: The cost function is $f_0(\{\text{MSE}_i\}) = \sum_{i=1}^L \alpha_i \text{MSE}_i$. The composite function

$$f_0(\{\exp(\mathbf{x})\}) = \sum_{i=1}^L \alpha_i \exp(x_i) \quad (54)$$

is neither Schur-concave nor Schur-convex for general $\{\alpha_i\}$.

2) *Maximization of the exponentially weighted product of SINRs*: The objective function to maximize is $\prod_{i=1}^L \text{SINR}_i^\alpha$. The cost function to minimize is

$$f_0(\{\text{MSE}_i\}) = \sum_{i=1}^L \log((\text{MSE}_i^{-1} - 1)^{-\alpha_i}). \quad (55)$$

The composite function

$$f_0 \circ \exp(\mathbf{x}) = -\sum_{i=1}^L \alpha_i \log(\exp(-x_i) - 1) \quad (56)$$

is neither Schur-concave nor Schur-convex for general $\{\alpha_i\}$.

VI. CONCLUSION

In this paper, we have studied optimizing precoder for the nonlinear MIMO transceiver based on MMSE-DFE. We first consider the general problem of minimizing some arbitrary cost function subject to the overall input power constraint. It is shown that the precoder matrix has some special structure for *any* cost function, based on which the original complicated matrix optimization problem is much simplified. Moreover, the optimum nonlinear design degenerates to linear diagonal transmission if the cost function f_0 is such that the composite $f_0 \circ \exp$ is Schur-concave. If f_0 is increasing in each argument and $f_0 \circ \exp$ is Schur-convex, then the optimum nonlinear transceiver design is the same, i.e., the uniform channel decomposition (UCD) scheme.

REFERENCES

- [1] T. Berger and D. W. Tufts, "Optimum pulse modulation – Part I: Transmitter-receiver design and bounds from information-theory," *IEEE Transactions on Information Theory*, vol. IT-13, Apr. 1967.
- [2] J. Salz, "Optimum mean-square decision-feedback equalization," *Bell Syst. Tech. J.*, pp. 1341–1373, Oct. 1973.
- [3] J. Yang and S. Roy, "On joint transmitter and receiver optimization for multiple-input-multiple-output (MIMO) transmission systems," *IEEE Transactions on Communications*, vol. 42, pp. 3221–3231, December 1994.
- [4] J. Yang and S. Roy, "Joint transmitter-receiver optimization for multi-input multi-output systems with decision feedback," *IEEE Transactions on Information Theory*, vol. 40, pp. 1334–1347, September 1994.
- [5] D. Palomar, J. Cioffi, and M. Lagunas, "Joint Tx-Rx beamforming design for multicarrier MIMO channels: A unified framework for convex optimization," *IEEE Transactions on Signal Processing*, vol. 51, pp. 2381–2401, September 2003.
- [6] Y. Jiang, W. Hager, and J. Li, "Tunable channel decomposition for MIMO communications using channel state information," *IEEE Transactions on Signal Processing*, vol. 54, pp. 4405 – 4418, November 2006.
- [7] A. Marshall and I. Olkin, *Inequalities: Theory of Majorization*. New York: Academic, 1979.
- [8] Y. Jiang, J. Li, and W. Hager, "Uniform channel decomposition for MIMO communications," *IEEE Transactions on Signal Processing*, vol. 53, pp. 4283 – 4294, November 2005.
- [9] M. Varanasi and T. Guess, "Optimum decision feedback multiuser equalization with successive decoding achieves the total capacity of the Gaussian multiple-access channel," *Proceedings of the Thirty-First Asilomar Conference on Signals, Systems and Computers*, vol. 2, pp. 1405 – 1409, Nov 2-5 1997.
- [10] Y. Jiang, W. Hager, and J. Li, "The generalized triangular decomposition," *Mathematics of Computation*, accepted, Nov. 2006.