Chapter 16 Majorization Theory and Applications

Jiaheng Wang[‡] and Daniel Palomar[♯]
[‡]KTH Royal Institute of Technology, Stockholm, Sweden
[♯]Hong Kong University of Science and Technology

In this chapter we introduce a useful mathematical tool, namely Majorization Theory, and illustrate its applications in a variety of scenarios in signal processing and communication systems. Majorization is a partial ordering and precisely defines the vague notion that the components of a vector are "less spread out" or "more nearly equal" than the components of another vector. Functions that preserve the ordering of majorization are said to be Schur-convex or Schur-concave. Many problems arising in signal processing and communications involve comparing vector-valued strategies or solving optimization problems with vector- or matrix-valued variables. Majorization theory is a key tool that allows us to solve or simplify these problems.

The goal of this chapter is to introduce the basic concepts and results on majorization that serve mostly the problems in signal processing and communications, but by no means to enclose the vast literature on majorization theory. A complete and superb reference on majorization theory is the book by Marshall and Olkin [1]. The building blocks of majorization can be found in [2], and [3] also contains significant material on majorization. Other textbooks on matrix and multivariate analysis, e.g., [4] and [5], may also include a part on majorization. Recent applications of majorization theory to signal processing and communication problems can be found in two good tutorials [6] and [7].

The chapter contains two parts. The first part is devoted to building the framework of majorization theory. The second part focuses on applying the concepts and results introduced in the first part to several problems arising in signal processing and communication systems.

16.1 Majorization Theory

16.1.1 Basic Concepts

To explain the concept of majorization, let us first define the following notations for increasing and decreasing orders of a vector.

Definition 16.1.1. For any vector $\mathbf{x} \in \mathbb{R}^n$, let

$$x_{[1]} \geq \cdots \geq x_{[n]}$$

denote its components in decreasing order, and let

$$x_{(1)} \leq \cdots \leq x_{(n)}$$

denote its components in increasing order.

Majorization[1] defines a partial ordering between two vectors, say $\mathbf{x}$ and $\mathbf{y}$, and precisely describes the concept that the components of $\mathbf{x}$ are "less spread out" or "more nearly equal" than the components of $\mathbf{y}$.

[1]The majorization ordering given in Definition 16.1.2 is also called additive majorization, to distinguish it from multiplicative majorization (or log-majorization) introduced in Section 16.1.4.

Definition 16.1.2. (**Majorization** [1, 1.A.1]) For any two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, we say $\mathbf{x}$ is majorized by $\mathbf{y}$ (or $\mathbf{y}$ majorizes $\mathbf{x}$), denoted by $\mathbf{x} \prec \mathbf{y}$ (or $\mathbf{y} \succ \mathbf{x}$), if

$$\begin{aligned} \sum_{i=1}^{k} x_{[i]} &\leq \sum_{i=1}^{k} y_{[i]}, \qquad 1 \leq k < n \\ \sum_{i=1}^{n} x_{[i]} &= \sum_{i=1}^{n} y_{[i]}. \end{aligned}$$

Alternatively, the previous conditions can be rewritten as

$$\begin{aligned} \sum_{i=1}^{k} x_{(i)} &\geq \sum_{i=1}^{k} y_{(i)}, \qquad 1 \leq k < n \\ \sum_{i=1}^{n} x_{(i)} &= \sum_{i=1}^{n} y_{(i)}. \end{aligned}$$

There are several equivalent characterizations of the majorization relation $\mathbf{x} \prec \mathbf{y}$ in addition to the conditions given in Definition 16.1.2. One alternative definition of majorization given in [2] is that $\mathbf{x} \prec \mathbf{y}$ if

$$\sum_{i=1}^{n} \phi(x_i) \leq \sum_{i=1}^{n} \phi(y_i) \tag{16.1}$$

for all continuous convex functions ϕ. Another interesting characterization of $\mathbf{x} \prec \mathbf{y}$, also from [2], is that $\mathbf{x} = \mathbf{P}\mathbf{y}$ for some doubly stochastic matrix[2] $\mathbf{P}$. In fact, the latter characterization implies that the set of vectors $\mathbf{x}$ that satisfy $\mathbf{x} \prec \mathbf{y}$ is the convex hull spanned by the $n!$ points formed from the permutations of the elements of $\mathbf{y}$.[3] Yet another interesting definition of $\mathbf{y} \succ \mathbf{x}$ is given in the form of waterfilling as

$$\sum_{i=1}^{n} (x_i - a)^+ \leq \sum_{i=1}^{n} (y_i - a)^+ \tag{16.2}$$

for any $a \in \mathbb{R}$ and $\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$, where $(u)^+ \triangleq \max(u, 0)$. The interested reader is referred to [1, Ch. 4] for more alternative characterizations.

Observe that the original order of the elements of $\mathbf{x}$ and $\mathbf{y}$ plays no role in the definition of majorization. In other words, $\mathbf{x} \prec \mathbf{\Pi}\mathbf{x}$ for all permutation matrices $\mathbf{\Pi}$.

Example 16.1.1. The following are simple examples of majorization:

$$\begin{aligned} \left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right) &\prec \left(\frac{1}{n-1}, \frac{1}{n-1}, \ldots, \frac{1}{n-1}, 0\right) \\ &\prec \cdots \prec \left(\frac{1}{2}, \frac{1}{2}, 0, \ldots, 0\right) \prec (1, 0, \ldots, 0). \end{aligned}$$

More generally

$$\left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right) \prec (x_1, x_2, \ldots, x_n) \prec (1, 0, \ldots, 0)$$

whenever $x_i \geq 0$ and $\sum_{i=1}^{n} x_i = 1$. □

[2] A square matrix $\mathbf{P}$ is said to be stochastic if either its rows or columns are probability vectors, i.e., if its elements are all nonnegative and either the rows or the columns sums are one. If both the rows and columns are probability vectors, then the matrix is called doubly stochastic. Stochastic matrices can be considered representations of the transition probabilities of a finite Markov chain.

[3] The permutation matrices are doubly stochastic and, in fact, the convex hull of the permutation matrices coincides with the set of doubly stochastic matrices [1, 2].

It is worth pointing out that majorization provides only a partial ordering, meaning that there exist vectors that can not be compared within the concept of majorization. For example, given $\mathbf{x} = (0.6, 0.2, 0.2)$ and $\mathbf{y} = (0.5, 0.4, 0.1)$, we have neither $\mathbf{x} \prec \mathbf{y}$ nor $\mathbf{x} \succ \mathbf{y}$.

To extend Definition 16.1.2, which is only applicable to vectors with the same sum, the following definition provides two partial orderings between two vectors with different sums.

Definition 16.1.3. (**Weak majorization** [1, 1.A.2]) For any two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, we say $\mathbf{x}$ is weakly submajorized by $\mathbf{y}$ (or $\mathbf{y}$ submajorizes $\mathbf{x}$), denoted by $\mathbf{x} \prec_w \mathbf{y}$ (or $\mathbf{y} \succ_w \mathbf{x}$), if

$$\sum_{i=1}^{k} x_{[i]} \leq \sum_{i=1}^{k} y_{[i]}, \qquad 1 \leq k \leq n.$$

We say $\mathbf{x}$ is weakly supermajorized by $\mathbf{y}$ (or $\mathbf{y}$ supermajorizes $\mathbf{x}$), denoted by $\mathbf{x} \prec^w \mathbf{y}$ (or $\mathbf{y} \succ^w \mathbf{x}$), if

$$\sum_{i=1}^{k} x_{(i)} \geq \sum_{i=1}^{k} y_{(i)}, \qquad 1 \leq k \leq n.$$

In either case, we say $\mathbf{x}$ is weakly majorized by $\mathbf{y}$ (or $\mathbf{y}$ weakly majorizes $\mathbf{x}$).

For nonnegative vectors, weak majorization can be alternatively characterized in terms of linear transformation by doubly substochastic and superstochastic matrices (see [1, Ch. 2]). Note that $\mathbf{x} \prec \mathbf{y}$ implies $\mathbf{x} \prec_w \mathbf{y}$ and $\mathbf{x} \prec^w \mathbf{y}$, but the inverse does not hold. In other words, majorization is a more restrictive definition than weak majorization. A useful connection between majorization and weak majorization is given as follows.

Lemma 16.1.1. ([1, 5.A.9, 5.A.9.a]) If $\mathbf{x} \prec_w \mathbf{y}$, then there exist vectors $\mathbf{u}$ and $\mathbf{v}$ such that

$$\mathbf{x} \leq \mathbf{u} \text{ and } \mathbf{u} \prec \mathbf{y}, \qquad \mathbf{v} \leq \mathbf{y} \text{ and } \mathbf{x} \prec \mathbf{v}.$$

If $\mathbf{x} \prec^w \mathbf{y}$, then there exist vectors $\mathbf{u}$ and $\mathbf{v}$ such that

$$\mathbf{x} \geq \mathbf{u} \text{ and } \mathbf{u} \prec \mathbf{y}, \qquad \mathbf{u} \geq \mathbf{y} \text{ and } \mathbf{x} \prec \mathbf{v}.$$

The notation $\mathbf{x} \leq \mathbf{u}$ means the component–wise ordering $x_i \leq u_i$, $i = 1, \ldots, n$, for all entries of vectors $\mathbf{x}, \mathbf{u}$.

16.1.2 Schur-Convex/Concave Functions

Functions that are monotonic with respect to the ordering of majorization are called Schur-convex or Schur-concave functions. This class of functions are of particular importance in this chapter, as it turns out that many design objectives in signal processing and communication systems are Schur-convex or Schur-concave functions.

Definition 16.1.4. (**Schur-convex/concave functions** [1, 3.A.1]) A real-valued function ϕ defined on a set $\mathfrak{A} \subseteq \mathbb{R}^n$ is said to be Schur-convex on $\mathfrak{A}$ if

$$\mathbf{x} \prec \mathbf{y} \text{ on } \mathfrak{A} \quad \Rightarrow \quad \phi(\mathbf{x}) \leq \phi(\mathbf{y}).$$

If, in addition, $\phi(\mathbf{x}) < \phi(\mathbf{y})$ whenever $\mathbf{x} \prec \mathbf{y}$ but $\mathbf{x}$ is not a permutation of $\mathbf{y}$, then ϕ is said to be strictly Schur-convex on $\mathfrak{A}$. Similarly, ϕ is said to be Schur-concave on $\mathfrak{A}$ if

$$\mathbf{x} \prec \mathbf{y} \text{ on } \mathfrak{A} \quad \Rightarrow \quad \phi(\mathbf{x}) \geq \phi(\mathbf{y}),$$

and ϕ is strictly Schur-concave on $\mathfrak{A}$ if strict inequality $\phi(\mathbf{x}) > \phi(\mathbf{y})$ holds when $\mathbf{x}$ is not a permutation of $\mathbf{y}$.

Clearly, if ϕ is Schur-convex on $\mathfrak{A}$, then $-\phi$ is Schur-concave on $\mathfrak{A}$, and vice versa.

It is important to remark that the sets of Schur-convex and Schur-concave functions do no form a partition of the set of all functions from $\mathfrak{A} \subseteq \mathbb{R}^n$ to $\mathbb{R}$. In fact, neither are the two sets disjoint (i.e., the intersection is not empty), unless we consider strictly Schur-convex/concave functions, nor do they cover the entire set of all functions as illustrated in Fig. 16.1.

Example 16.1.2. The simplest example of a Schur-convex function, according to the definition, is $\phi(\mathbf{x}) = \max_k\{x_k\} = x_{[1]}$, which is also strictly Schur-convex. □

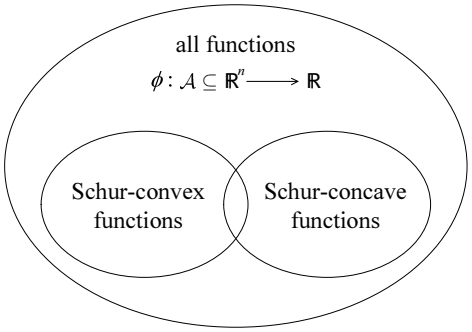

Figure 16.1: Illustration of the sets of Schur-convex and Schur-concave functions within the set of all functions $\phi : \mathfrak{A} \subseteq \mathbb{R}^n \to \mathbb{R}$.

Example 16.1.3. The function $\phi(\mathbf{x}) = \sum_{i=1}^{n} x_i$ is both Schur-convex and Schur-concave since $\phi(\mathbf{x}) = \phi(\mathbf{y})$ for any $\mathbf{x} \prec \mathbf{y}$. However, it is neither strictly Schur-convex nor strictly Schur-concave. □

Example 16.1.4. The function $\phi(\mathbf{x}) = x_1 + 2x_2 + x_3$ is neither Schur-convex nor Schur-concave, as can be seen from the counterexample that for $\mathbf{x} = (2, 1, 1)$, $\mathbf{y} = (2, 2, 0)$ and $\mathbf{z} = (4, 0, 0)$, we have $\mathbf{x} \prec \mathbf{y} \prec \mathbf{z}$ but $\phi(\mathbf{x}) < \phi(\mathbf{y}) > \phi(\mathbf{z})$. □

To distinguish Schur-convexity/concavity from common monotonicity, we also define increasing and decreasing functions that will be frequently used later.

Definition 16.1.5. (**Increasing/Decreasing functions**) A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be increasing if it is increasing in each argument, i.e.,

$$\mathbf{x} \leq \mathbf{y} \quad \Rightarrow \quad f(\mathbf{x}) \leq f(\mathbf{y}),$$

and to be decreasing if it is decreasing in each argument, i.e.,

$$\mathbf{x} \leq \mathbf{y} \quad \Rightarrow \quad f(\mathbf{x}) \geq f(\mathbf{y}).$$

Using directly Definition 16.1.4 to check Schur-convexity/concavity of a function may not be easy. In the following, we present some immediate results to determine whether a function is Schur-convex or Schur-concave.

Theorem 16.1.1. *([1, 3.A.3]) Let the function $\phi : \mathfrak{D}_n \to \mathbb{R}$ be continuous on $\mathfrak{D}_n \triangleq \{\mathbf{x} \in \mathbb{R}^n : x_1 \geq \cdots \geq x_n\}$ and continuously differentiable on the interior of $\mathfrak{D}_n$. Then ϕ is Schur-convex (Schur-concave) on $\mathfrak{D}_n$ if and only if $\frac{\partial \phi(\mathbf{x})}{\partial x_i}$ is decreasing (increasing) in $i = 1, \ldots, n$.*

Theorem 16.1.2. ***(Schur's condition** [1, 3.A.4]) Let $\mathfrak{I} \subseteq \mathbb{R}$ be an open interval and the function $\phi : \mathfrak{I}^n \to \mathbb{R}$ be continuously differentiable. Then ϕ is Schur-convex on $\mathfrak{I}^n$ if and only if ϕ is symmetric[4] on $\mathfrak{I}^n$ and*

$$(x_i - x_j)\left(\frac{\partial \phi}{\partial x_i} - \frac{\partial \phi}{\partial x_j}\right) \geq 0, \qquad 1 \leq i, j \leq n. \tag{16.3}$$

ϕ is Schur-concave on on $\mathfrak{I}^n$ if and only if ϕ is symmetric and the inequality (16.3) *is reversed.*

In fact, to prove Schur-convexity/concavity of a function using Theorem 16.1.1 and Theorem 16.1.2, one can take $n = 2$ without loss of generality (w.l.o.g.), i.e., check only the two-argument case [1, 3.A.5]. Based on Theorem 16.1.1 and Theorem 16.1.2, it is possible to obtain some sufficient conditions guaranteeing Schur-convexity/concavity of different composite functions.

Proposition 16.1.1. (**Monotonic composition** [1, 3.B.1]) Consider the composite function $\phi(\mathbf{x}) = f(g_1(\mathbf{x}), \ldots, g_k(\mathbf{x}))$, where f is a real-valued function defined on $\mathbb{R}^k$. Then, it follows that

- f is increasing and g_i is Schur-convex $\Rightarrow$ ϕ is Schur-convex;
- f is decreasing and g_i is Schur-convex $\Rightarrow$ ϕ is Schur-concave;

[4] A function is said to be symmetric if its arguments can be arbitrarily permuted without changing the function value.

- f is increasing and g_i is Schur-concave $\Rightarrow$ ϕ is Schur-concave;
- f is decreasing and g_i is Schur-concave $\Rightarrow$ ϕ is Schur-convex.

Proposition 16.1.2. (**Convex**[5] **composition** [1, 3.B.2]) Consider the composite function $\phi(\mathbf{x}) = f\left(g(x_1), \ldots, g(x_n)\right)$, where f is a real-valued function defined on $\mathbb{R}^n$. Then, it follows that

- f is increasing Schur-convex and g convex $\Rightarrow$ ϕ is Schur-convex;
- f is decreasing Schur-convex and g concave $\Rightarrow$ ϕ is Schur-convex.

For some special forms of functions, there exist simple conditions to check whether they are Schur-convex or Schur-concave.

Proposition 16.1.3. (**Symmetric convex functions** [1, 3.C.2]) If ϕ is symmetric and convex (concave), then ϕ is Schur-convex (Schur-concave).

Corollary 16.1.1. ([1, 3.C.1]) Let $\phi(\mathbf{x}) = \sum_{i=1}^{n} g\left(x_i\right)$, where g is convex (concave). Then ϕ is Schur-convex (Schur-concave).

Proposition 16.1.3 can be generalized to the case of quasi-convex functions.[6]

Proposition 16.1.4. (**Symmetric quasi-convex functions** [1, 3.C.3]) If ϕ is symmetric and quasi-convex, then ϕ is Schur-convex.

Schur-convexity/concavity can also be extended to weak majorization through the following fact.

Theorem 16.1.3. *([1, 3.A.8]) A real-valued function ϕ defined on a set $\mathfrak{A} \subseteq \mathbb{R}^n$ satisfies*

$$\mathbf{x} \prec_w \mathbf{y} \text{ on } \mathfrak{A} \quad \Rightarrow \quad \phi(\mathbf{x}) \leq \phi(\mathbf{y})$$

if and only if ϕ is increasing and Schur-convex on $\mathfrak{A}$. Similarly, ϕ satisfies

$$\mathbf{x} \prec^w \mathbf{y} \text{ on } \mathfrak{A} \quad \Rightarrow \quad \phi(\mathbf{x}) \leq \phi(\mathbf{y})$$

if and only if ϕ is decreasing and Schur-convex on $\mathfrak{A}$.

By using the above results, we are now able to find various Schur-convex/concave functions. Several such examples are provided in the following, while the interested reader can find more Schur-convex/concave functions in [1].

Example 16.1.5. Consider the l_p norm $|\mathbf{x}|_p = (\sum_i |x_i|^p)^{1/p}$, which is symmetric and convex when $p \geq 1$. Thus, from Proposition 16.1.3, $|\mathbf{x}|_p$ is Schur-convex for $p \geq 1$. □

Example 16.1.6. Suppose that $x_i > 0$. Since x^a is convex when $a \geq 1$ and $a \leq 0$ and concave when $0 \leq a < 1$, from Corollary 16.1.1, $\phi(\mathbf{x}) = \sum_i x_i^a$ is Schur-convex for $a \geq 1$ and $a \leq 0$, and Schur-concave for $0 \leq a < 1$. Similarly, $\phi(\mathbf{x}) = \sum_i \log x_i$ and $\phi(\mathbf{x}) = -\sum_i x_i \log x_i$ are both Schur-concave, since $\log x$ and $-x \log x$ are concave. □

Example 16.1.7. Consider $\phi : \mathbb{R}_+^2 \to \mathbb{R}$ with $\phi(\mathbf{x}) = -x_1 x_2$, which is symmetric and quasi-convex. Thus, from Proposition 16.1.4, it is Schur-convex. □

16.1.3 Relation to Matrix Theory

There are many interesting results that connect majorization theory to matrix theory, among which a crucial finding by Schur is that the diagonal elements of a Hermitian matrix are majorized by its eigenvalues. This fact has been frequently used to simplify optimization problems with matrix-valued variables.

Theorem 16.1.4. *(**Schur's inequality** [1, 9.B.1]) Let $\mathbf{A}$ be a Hermitian matrix with diagonal elements denoted by the vector $\mathbf{d}$ and eigenvalues denoted by the vector $\boldsymbol{\lambda}$. Then $\boldsymbol{\lambda} \succ \mathbf{d}$.*

[5] A function $f : \mathfrak{X} \to \mathbb{R}$ is convex if $\mathfrak{X}$ is a convex set and for any $x, y \in \mathfrak{X}$ and $0 \leq \alpha \leq 1$, $f\left(\alpha x + (1-\alpha)\, y\right) \leq \alpha f(x) + (1-\alpha)\, f(y)$. f is concave if $-f$ is convex.

[6] A function $f : \mathfrak{X} \to \mathbb{R}$ is quasiconvex if $\mathfrak{X}$ is a convex set and for any $x, y \in \mathfrak{X}$ and $0 \leq \alpha \leq 1$, $f(\alpha x + (1-\alpha)y) \leq \max\{f(x), f(y)\}$. A convex function is also quasi-convex, but the converse is not true.

Theorem 16.1.4 provides an "upper bound" on the diagonal elements of a Hermitian matrix in terms of the majorization ordering. From Exercise 16.1.1, a natural "lower bound" of a vector $\mathbf{x} \in \mathbb{R}^n$ would be $\mathbf{1} \prec \mathbf{x}$, where $\mathbf{1} \in \mathbb{R}^n$ denote the vector with equal elements given by $1_i \triangleq \sum_{j=1}^{n} x_j/n$. Therefore, for any Hermitian matrix we have

$$\mathbf{1} \prec \mathbf{d} \prec \boldsymbol{\lambda} \tag{16.4}$$

which is formally described in the following corollary.

Corollary 16.1.2. Let $\mathbf{A}$ be a Hermitian matrix and $\mathbf{U}$ a unitary matrix. Then

$$\mathbf{1}(\mathbf{A}) \prec \mathbf{d}\left(\mathbf{U}^\dagger \mathbf{A}\mathbf{U}\right) \prec \boldsymbol{\lambda}(\mathbf{A})$$

where $\mathbf{1}(\mathbf{A})$ denotes the vector of equal elements whose sum equal to $\mathrm{tr}(\mathbf{A})$, $\mathbf{d}(\mathbf{A})$ is the vector of the diagonal elements of $\mathbf{A}$, and $\boldsymbol{\lambda}(\mathbf{A})$ is the vector of the eigenvalues of $\mathbf{A}$.

Proof: It follows directly from (16.4), as well as the fact that $\mathbf{1}\left(\mathbf{U}^\dagger \mathbf{A}\mathbf{U}\right) = \mathbf{1}(\mathbf{A})$ and $\boldsymbol{\lambda}\left(\mathbf{U}^\dagger \mathbf{A}\mathbf{U}\right) = \boldsymbol{\lambda}(\mathbf{A})$. □

Corollary 16.1.2 "bounds" the diagonal elements of $\mathbf{U}^\dagger \mathbf{A}\mathbf{U}$ for any unitary matrix $\mathbf{U}$. However, it does not specify what can be achieved. The following result will be instrumental for that purpose.

Theorem 16.1.5. *([1, 9.B.2]) For any two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ satisfying $\mathbf{x} \prec \mathbf{y}$, there exists a real symmetric (and therefore Hermitian) matrix with diagonal elements given by $\mathbf{x}$ and eigenvalues given by $\mathbf{y}$.*

Corollary 16.1.3. For any vector $\boldsymbol{\lambda} \in \mathbb{R}^n$, there exists a real symmetric (and therefore Hermitian) matrix with equal diagonal elements and eigenvalues given by $\boldsymbol{\lambda}$.

Corollary 16.1.4. Let $\mathbf{A}$ be a Hermitian matrix and $\mathbf{x} \in \mathbb{R}^n$ be a vector satisfying $\mathbf{x} \prec \boldsymbol{\lambda}(\mathbf{A})$. Then, there exists a unitary matrix $\mathbf{U}$ such that

$$\mathbf{d}\left(\mathbf{U}^\dagger \mathbf{A}\mathbf{U}\right) = \mathbf{x}.$$

Proof: The proofs of Corollary 16.1.3 and Corollary 16.1.4 are straightforward from Corollary 16.1.2 and Theorem 16.1.5. □

Theorem 16.1.5 is the converse of Theorem 16.1.4 (in fact it is stronger than the converse since it guarantees the existence of a real symmetric matrix instead of just a Hermitian matrix). Now, we can provide the converse of Corollary 16.1.2.

Corollary 16.1.5. Let $\mathbf{A}$ be a Hermitian matrix. There exists a unitary matrix $\mathbf{U}$ such that

$$\mathbf{d}\left(\mathbf{U}^\dagger \mathbf{A}\mathbf{U}\right) = \mathbf{1}(\mathbf{A}),$$

and also another unitary matrix $\mathbf{U}$ such that

$$\mathbf{d}\left(\mathbf{U}^\dagger \mathbf{A}\mathbf{U}\right) = \boldsymbol{\lambda}(\mathbf{A}).$$

We now turn to the important algorithmic aspect of majorization theory which is necessary, for example, to compute a matrix with given diagonal elements and eigenvalues. The following definition is instrumental in the derivation of transformations that relate vectors that satisfy the majorization relation.

Definition 16.1.6. (***T-transform*** [1, p. 21]) A *T-transform* is a matrix of the form

$$\mathbf{T} = \alpha \mathbf{I} + (1-\alpha)\,\boldsymbol{\Pi} \tag{16.5}$$

for some $\alpha \in [0,1]$ and some $n \times n$ permutation matrix $\boldsymbol{\Pi}$ with $n-2$ diagonal entries equal to 1. Let $[\boldsymbol{\Pi}]_{ij} = [\boldsymbol{\Pi}]_{ji} = 1$ for some indices $i < j$, then

$$\boldsymbol{\Pi}\mathbf{y} = [y_1, \ldots, y_{i-1}, y_j, y_{i+1}, \ldots, y_{j-1}, y_i, y_{j+1}, \ldots, y_n]^T$$

and hence

$$\begin{aligned}\mathbf{T}\mathbf{y} = [&y_1, \ldots, y_{i-1}, \alpha\, y_i + (1-\alpha)\, y_j, y_{i+1}, \ldots, \\ &y_{j-1}, \alpha\, y_j + (1-\alpha)\, y_i, y_{j+1}, \ldots, y_n]^T.\end{aligned}$$

Lemma 16.1.2. ([1, 2.B.1]) For any two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ satisfying $\mathbf{x} \prec \mathbf{y}$, there exists a sequence of T-transforms $\mathbf{T}^{(1)}, \ldots, \mathbf{T}^{(K)}$ such that $\mathbf{x} = \mathbf{T}^{(K)} \cdots \mathbf{T}^{(1)} \mathbf{y}$ and $K < n$.

An algorithm to obtain such a sequence of T-transforms is introduced next.

Algorithm 14. *([1, 2.B.1]) Algorithm to obtain a sequence of T-transforms such that* $\mathbf{x} = \mathbf{T}^{(K)} \cdots \mathbf{T}^{(1)} \mathbf{y}$.
Input: *Vectors* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ *satisfying* $\mathbf{x} \prec \mathbf{y}$ *(it is assumed that the components of* $\mathbf{x}$ *and* $\mathbf{y}$ *are in decreasing order and that* $\mathbf{x} \neq \mathbf{y}$*).*
Output: *Set of T-transforms* $\mathbf{T}^{(1)}, \ldots, \mathbf{T}^{(K)}$.

0. *Let* $\mathbf{y}^{(0)} = \mathbf{y}$ *and* $k = 1$ *be the iteration index.*

1. *Find the largest index* i *such that* $y_i^{(k-1)} > x_i$ *and the smallest index* j *greater than* i *such that* $y_j^{(k-1)} < x_j$.

2. *Let* $\delta = \min\left(x_j - y_j^{(k-1)}, y_i^{(k-1)} - x_i\right)$ *and* $\alpha = 1 - \delta / \left(y_i^{(k-1)} - y_j^{(k-1)}\right)$.

3. *Use* α *to compute* $\mathbf{T}^{(k)}$ *as in* (16.5) *and let* $\mathbf{y}^{(k)} = \mathbf{T}^{(k)} \mathbf{y}^{(k-1)}$.

4. *If* $\mathbf{y}^{(k)} \neq \mathbf{x}$*, then set* $k = k + 1$ *and go to step 1; otherwise, finish.*

Recursive algorithms to obtain a matrix with given eigenvalues and diagonal elements are provided in [1, 9.B.2] and [8]. Here, we introduce the practical and simple method proposed in [8] as follows.

Algorithm 15. *([8]) Algorithm to obtain a real symmetric matrix* $\mathbf{A}$ *with diagonal values given by* $\mathbf{x}$ *and eigenvalues given by* $\mathbf{y}$ *provided that* $\mathbf{x} \prec \mathbf{y}$.
Input: *Vectors* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ *satisfying* $\mathbf{x} \prec \mathbf{y}$ *(it is assumed that the components of* $\mathbf{x}$ *and* $\mathbf{y}$ *are in decreasing order and that* $\mathbf{x} \neq \mathbf{y}$*).*
Output: *Matrix* $\mathbf{A}$.

1. *Using Algorithm 14, obtain a sequence of T-transforms such that* $\mathbf{x} = \mathbf{T}^{(K)} \cdots \mathbf{T}^{(1)} \mathbf{y}$.

2. *Define the Givens rotation* $\mathbf{U}^{(k)}$ *as*

$$\left[\mathbf{U}^{(k)}\right]_{ij} = \begin{cases} \sqrt{\left[\mathbf{T}^{(k)}\right]_{ij}}, & \text{for } i < j \\ -\sqrt{\left[\mathbf{T}^{(k)}\right]_{ij}}, & \text{otherwise.} \end{cases}$$

3. *Let* $\mathbf{A}^{(0)} = \operatorname{diag}(\mathbf{y})$ *and* $\mathbf{A}^{(k)} = \mathbf{U}^{(k)T} \mathbf{A}^{(k-1)} \mathbf{U}^{(k)}$. *The desired matrix is given by* $\mathbf{A} = \mathbf{A}^{(K)}$. *Define the unitary matrix* $\mathbf{U} = \mathbf{U}^{(1)} \cdots \mathbf{U}^{(K)}$ *and the desired matrix is given by* $\mathbf{A} = \mathbf{U}^T \operatorname{diag}(\mathbf{y}) \mathbf{U}$.

Algorithm 15 obtains a real symmetric matrix $\mathbf{A}$ with given eigenvalues and diagonal elements. For the interesting case in which the diagonal elements must be equal and the desired matrix is allowed to be complex, it is possible to obtain an alternative much simpler solution in closed form as given next.

Lemma 16.1.3. ([9]) Let $\mathbf{U}$ be a unitary matrix satisfying the condition $|[\mathbf{U}]_{ik}| = |[\mathbf{U}]_{il}| \ \forall i, k, l$. Then, the matrix $\mathbf{A} = \mathbf{U}^H \operatorname{diag}(\lambda) \mathbf{U}$ has equal diagonal elements (and eigenvalues given by λ). Two examples of $\mathbf{U}$ are the unitary DFT matrix and the Hadamard matrix (when the dimensions are appropriate such as a power of two).

Nevertheless, Algorithm 15 has the nice property that the obtained matrix $\mathbf{U}$ is real-valued and can be naturally decomposed (by construction) as the product of a series of rotations. This simple structure plays a key role for practical implementation. Interestingly, an iterative approach to construct a matrix with equal diagonal elements and with a given set of eigenvalues was obtained in [10], based also on a sequence of rotations.

16.1.4 Multiplicative Majorization

Parallel to the concept of majorization introduced in Section 16.1.1, which is often called additive majorization, is the notion of multiplicative majorization (also termed log-majorization) defined as follows.

Definition 16.1.7. The vector $\mathbf{x} \in \mathbb{R}_+^n$ is multiplicatively majorized by $\mathbf{y} \in \mathbb{R}_+^n$, denoted by $\mathbf{x} \prec_\times \mathbf{y}$, if

$$\begin{aligned} \prod_{i=1}^{k} x_{[i]} &\leq \prod_{i=1}^{k} y_{[i]}, \qquad 1 \leq k < n \\ \prod_{i=1}^{n} x_{[i]} &= \prod_{i=1}^{n} y_{[i]}. \end{aligned}$$

To differentiate the two types of majorization, we sometimes use the symbol $\prec_+$ rather than $\prec$ to denote (additive) majorization. It is easy to see the relation between additive majorization and multiplicative majorization: $\mathbf{x} \prec_+ \mathbf{y}$ if and only if $\exp(\mathbf{x}) \prec_\times \exp(\mathbf{y})$.[7]

Example 16.1.8. Given $\mathbf{x} \in \mathbb{R}_+^n$, let $\mathbf{g}$ denote the vector of equal elements given by $g_i \triangleq (\prod_{j=1}^n x_j)^{1/n}$, i.e., the geometric mean of $\mathbf{x}$. Then, $\mathbf{g} \prec_\times \mathbf{x}$.

Similar to the definition of Schur-convex/concave functions, it is also possible to define multiplicatively Schur-convex/concave functions.

Definition 16.1.8. A function $\phi : \mathcal{A} \to \mathbb{R}$ is said to be multiplicatively Schur-convex on $\mathcal{A} \in \mathbb{R}^n$ if

$$\mathbf{x} \prec_\times \mathbf{y} \text{ on } \mathcal{A} \quad \Rightarrow \quad \phi(\mathbf{x}) \leq \phi(\mathbf{y}),$$

and multiplicatively Schur-concave on $\mathcal{A}$ if

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

However, considering the correspondence between additive and multiplicative majorization, it may not be necessary to use the notion of multiplicatively Schur-convex/concave functions. Instead, the so-called multiplicatively Schur-convex/concave functions in Definition 16.1.8 can be equivalently referred to as functions such that $\phi \circ \exp$ is Schur-convex and Schur-concave, respectively, where the composite function is defined as $\phi \circ \exp(\mathbf{x}) \triangleq \phi(e^{x_1}, \ldots, e^{x_n})$.

The following two lemmas relate Schur-convexity/concavity of a function f with that of the composite function $\phi \circ \exp$.

Lemma 16.1.4. If ϕ is increasing and Schur-convex, then $\phi \circ \exp$ is Schur-convex.

Proof: It is an immediate result from Proposition 16.1.2. □

Lemma 16.1.5. If the composite function $\phi \circ \exp$ is Schur-concave on $\mathcal{D}_n \triangleq \{\mathbf{x} \in \mathbb{R}^n : x_1 \geq \cdots \geq x_n\}$, then ϕ is Schur-concave on $\mathcal{D}_n$ if it is increasing.

Proof: It can be easily proved using Theorem 16.1.1. □

The following two examples show that the implication in Lemmas 16.1.4 and 16.1.5 does not hold in the opposite direction.

Example 16.1.9. The function $\phi(\mathbf{x}) = \prod_{i=1}^n x_i$ is Schur-concave on $\mathcal{D}_n$ since $\frac{\partial \phi(\mathbf{x})}{\partial x_i} = \frac{\phi(\mathbf{x})}{x_i}$ is increasing in i on $\mathcal{D}_n$ (see Theorem 16.1.1). However, the composite function $\phi \circ \exp(\mathbf{x}) = \exp(\sum_i x_i)$ is Schur-convex (and Schur-concave as well).

Example 16.1.10. The function $\phi(\mathbf{x}) = \sum_{i=1}^n \alpha_i x_i$ with $\alpha_1 \leq \cdots \leq \alpha_n$ is Schur-concave on $\mathcal{D}_n$. The composite function is $\phi \circ \exp(\mathbf{x}) = \sum_{i=1}^n \alpha_i \exp(x_i)$. For $\alpha_i \leq \alpha_{i+1}$, the derivative $\frac{\partial \phi \circ \exp(\mathbf{x})}{\partial x_i} = \alpha_i \exp(x_i)$ is not always monotonic in $i = 1, \ldots, n$ for any $\mathbf{x} \in \mathcal{D}_n$. Hence according to Theorem 16.1.1, although ϕ is Schur-concave, $\phi \circ \exp$ is not Schur-concave (neither Schur-convex) on $\mathcal{D}_n$.

[7] Indeed, using the language of group theory, we say that the groups $(\mathbb{R}, +)$ and $(\mathbb{R}_+, \times)$ are isomorphic since there is a bijection function $\exp : \mathbb{R} \to \mathbb{R}_+$ such that $\exp(x + y) = \exp(x) \times \exp(y)$ for $\forall\, x, y \in \mathbb{R}$.

In contrast with (additive) majorization that leads to some important relations between eigenvalues and diagonal elements of a Hermitian matrix (see Section 16.1.3), multiplicative majorization also brings some insights to matrix theory but mostly on the relation between singular values and eigenvalues of a matrix. In the following, we introduce one recent result that will be used later.

Theorem 16.1.6. *(**Generalized triangular decomposition (GTD)** [11]) Let $\mathbf{H} \in \mathbb{C}^{m\times n}$ be a matrix with rank k and singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k$. 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 $\mathbf{H} = \mathbf{Q}\mathbf{R}\mathbf{P}^H$ if and only if the diagonal elements of $\mathbf{R}$ satisfy $|\mathbf{r}| \prec_\times \sigma$, where $|\mathbf{r}|$ is a vector with the absolute values of $\mathbf{r}$ element-wise. Vectors σ and $\mathbf{r}$ stand for the singular values of $\mathbf{H}$ and diagonal entries of $\mathbf{R}$, respectively.*

The GTD is a generic form including many well-known matrix decompositions such as the SVD, the Schur decomposition, and the QR factorization [4]. Theorem 16.1.6 implies that, given $|\mathbf{r}| \prec_\times \sigma$, there exists a matrix $\mathbf{H}$ with its singular values and eigenvalues being $\mathbf{r}$ and σ, respectively. A recursive algorithm to find such a matrix was proposed in [11].

16.1.5 Stochastic Majorization

A comparison of some kind between two random variables X and Y is called stochastic majorization if the comparison reduces to the ordinary majorization $x \prec y$ in case X and Y are degenerate at x and y, i.e., $\Pr(X = x) = 1$ and $\Pr(Y = y) = 1$. Random vectors to be compared by stochastic majorization often have distributions belonging to the same parametric family, where the parameter space is a subset of $\mathbb{R}^n$. In this case, random variables X and Y with corresponding distributions F_θ and $F_{\theta'}$ are ordered by stochastic majorization if and only if the parameters θ and θ' are ordered by ordinary majorization.

Specifically, let $\mathcal{A} \subseteq \mathbb{R}^n$ and $\{F_\theta : \theta \in \mathcal{A}\}$ be a family of n-dimensional distribution functions indexed by a vector-valued parameter θ. Let

$$E\{\phi(X)\} = \int_{\mathbb{R}^n} \phi(x) dF_\theta(x) \tag{16.6}$$

denote the expectation of $\phi(X)$ when X has distribution F_θ, and let

$$\Pr(\phi(X) \leq t) = \int_{\phi(x)\leq t} dF_\theta(x) \tag{16.7}$$

denote the tail probability that $\phi(X)$ is less than or equal to t when X has distribution F_θ. We are particularly interested in investigating whether or in what conditions $E\{\phi(X)\}$ and $\Pr(\phi(X) \leq t)$ are Schur-convex/concave in θ.

The following results provide the conditions in which $E\{\phi(X)\}$ is Schur-convex in θ for exchangeable random variables.[8]

Proposition 16.1.5. ([1, 11.B.1]) Let $X_1, \ldots, X_n$ be exchangeable random variables and suppose that $\Phi : \mathbb{R}^{2n} \to \mathbb{R}$ satisfies: (i) $\Phi(\mathbf{x}, \theta)$ is convex in θ for each fixed $\mathbf{x}$; (ii) $\Phi(\mathbf{\Pi x}, \mathbf{\Pi}\theta) = \Phi(\mathbf{x}, \theta)$ for all permutations $\mathbf{\Pi}$; (iii) $\Phi(\mathbf{x}, \theta)$ is Borel measurable in $\mathbf{x}$ for each fixed θ. Then,

$$\psi(\theta) = E\{\Phi(X_1, \ldots, X_n, \theta)\}$$

is symmetric and convex (and thus Schur-convex).

Corollary 16.1.6. ([1, 11.B.2, 11.B.3]) Let $X_1, \ldots, X_n$ be exchangeable random variables and $\phi : \mathbb{R}^n \to \mathbb{R}$ be symmetric and convex. Then,

$$\psi(\theta) = E\{\phi(\theta_1 X_1, \ldots, \theta_n X_n)\}$$

and

$$\psi(\theta) = E\{\phi(X_1 + \theta_1, \ldots, X_n + \theta_n)\}$$

are symmetric and convex (and thus Schur-convex).

Corollary 16.1.7. ([1, 11.B.2.c]) Let $X_1, \ldots, X_n$ be exchangeable random variables and g be a continuous convex function. Then,

$$\psi(\theta) = E\left[g\left(\sum_i \theta_i X_i\right)\right]$$

is symmetric and convex (and thus Schur-convex).

[8] $X_1, \ldots, X_n$ are exchangeable random variables if the distribution of $X_{\pi(1)}, \ldots, X_{\pi(n)}$ does not depend on the permutation π. In other words, the joint distribution of $X_1, \ldots, X_n$ is invariant under permutations of its arguments. For example, independent and identically distributed random variables are exchangeable.

Compared to the expectation form, there are only a limited number of results on the Schur-convexity/concavity of the tail probability $P\{\phi(X) \leq t\}$ in terms of θ, which are usually given in some specific form of ϕ or distribution of X. In the following is one important result concerning linear combinations of some random variables.

Theorem 16.1.7. *([12]) Let $\theta_i \geq 0$ for all i, and $X_1, \ldots, X_n$ be independent and identically distributed (iid) random variables following a Gamma distribution with the density*

$$f(x) = \frac{x^{k-1}\exp(-x)}{\Gamma(k)}$$

where $\Gamma(k) = (k-1)!$. Suppose that $g : \mathbb{R} \to \mathbb{R}$ is nonnegative and the inverse function g^{-1} exists. Then,

$$P\left\{g\left(\sum_i \theta_i X_i\right) \leq t\right\}$$

is Schur-concave in θ for $t \geq g(2)$ and Schur-convex in θ for $t \leq g(1)$.

Corollary 16.1.8. Let $\theta_i \geq 0$ for all i, and $X_1, \ldots, X_n$ be iid exponential random variables with the density $f(x) = \exp(-x)$. Then, $P\left\{\sum_i \theta_i X_i \leq t\right\}$ is Schur-concave in θ for $t \geq 2$ and Schur-convex in θ for $t \leq 1$.

For more examples of stochastic majorization in the form of expectations or tail probabilities, we refer the interested reader to [1] and [7].

16.2 Applications of Majorization Theory

16.2.1 CDMA Sequence Design

The code division multiple access (CDMA) system is an important multiple-access technique in wireless networks. In a CDMA system, all users share the same bandwidth and they are distinguished from each other by their signature sequences or codes. A fundamental problem in CDMA systems is to optimally design signature sequences so that the system performance, such as the sum capacity, is maximized.

Consider the uplink of a single-cell synchronous CDMA system with K users and processing gain N. In the presence of additive white Gaussian noise, the sampled baseband received signal vector in one symbol interval is

$$\mathbf{r} = \sum_{i=1}^{K} \mathbf{s}_i \sqrt{p_i} b_i + \mathbf{n} \tag{16.8}$$

where, for each user i, p_i is the received power, b_i is the transmitted symbol, and $\mathbf{s}_i \in \mathbb{R}^N$ is the unit-energy signature sequence, i.e., $\|\mathbf{s}_i\| = 1$, and $\mathbf{n}$ is a zero-mean Gaussian random vector with covariance matrix $\sigma^2\mathbf{I}_N$, i.e., $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_N)$. Introduce an $N \times K$ signature sequence matrix $\mathbf{S} \triangleq [\mathbf{s}_1, \ldots, \mathbf{s}_K]$ and let $\mathbf{P}^{1/2} \triangleq \operatorname{diag}\left\{\sqrt{p_1}, \ldots, \sqrt{p_K}\right\}$ and $\mathbf{b} \triangleq [b_1, \ldots, b_K]^T$. Then (16.8) can be compactly expressed as

$$\mathbf{r} = \mathbf{S}\mathbf{P}^{1/2}\mathbf{b} + \mathbf{n}. \tag{16.9}$$

There are different criteria to measure the performance of a CDMA system, among which the most commonly used one may be the sum capacity given by [13]

$$C_{\text{sum}} = \frac{1}{2}\log\det\left(\mathbf{I}_N + \sigma^{-2}\mathbf{S}\mathbf{P}\mathbf{S}^T\right). \tag{16.10}$$

In practice, the system performance may also be measured by the total MSE of all users, which, assuming that each uses a LMMSE filter at his receiver, is given by [14]

$$\text{MSE} = K - \text{Tr}\left[\mathbf{S}\mathbf{P}\mathbf{S}^T\left(\mathbf{S}\mathbf{P}\mathbf{S}^T + \sigma^2\mathbf{I}_N\right)^{-1}\right]. \tag{16.11}$$

Another important global quantity that measures the total interference in the CDMA system is the total weighted square correlation (TWSC), which is given by [15]

$$\text{TWSC} = \sum_{i=1}^{K}\sum_{j=1}^{K} p_i p_j \left(\mathbf{s}_i^T \mathbf{s}_j\right) = \text{Tr}\left[\left(\mathbf{S}\mathbf{P}\mathbf{S}^T\right)^2\right]. \tag{16.12}$$

The goal of the sequence design problem is to optimize the system performance, e.g., maximize C_{sum} or minimize MSE or TWSC, by properly choosing the signature sequences for all users or, equivalently, by choosing the optimal signature sequence matrix $\mathbf{S}$.

Observe that the aforementioned three performance measures are all determined by the eigenvalues of the matrix $\mathbf{S}\mathbf{P}\mathbf{S}^T$. To be more exact, denoting the eigenvalues of $\mathbf{S}\mathbf{P}\mathbf{S}^T$ by $\lambda \triangleq (\lambda_i)_{i=1}^N$, it follows that

$$\begin{aligned} C_{\text{sum}} &= \frac{1}{2}\sum_{i=1}^{N} \log\left(1+\frac{\lambda_i}{\sigma^2}\right) \\ \text{MSE} &= K - \sum_{i=1}^{N} \frac{\lambda_i}{\lambda_i+\sigma^2} \\ \text{TWSC} &= \sum_{i=1}^{N} \lambda_i^2. \end{aligned}$$

Now, we can apply majorization theory. Indeed, since $\log(1+\sigma^{-2}x)$ is a concave function, it follows from Corollary 16.1.1 that C_{sum} is a Schur-concave function with respect to λ. Similarly, given that $-x/(x+\sigma^2)$ and x^2 are convex functions, both MSE and TWSC are Schur-convex in λ. Therefore, if one can find a signature sequence matrix yielding the Schur-minimal eigenvalues, i.e., one with eigenvalues that are majorized by all other feasible eigenvalues, then the resulting signature sequences will not only maximize C_{sum} but also minimize MSE and TWSC at the same time.

To find the optimal signature sequence matrix $\mathbf{S}$, let us first define the set of all feasible $\mathbf{S}$

$$\mathcal{S} \triangleq \{\mathbf{S} \in \mathbb{R}^{N\times K} : \|\mathbf{s}_i\| = 1,\ i = 1,\ldots,K\} \tag{16.13}$$

and correspondingly the set of all possible λ

$$\mathcal{L} \triangleq \left\{\lambda(\mathbf{S}\mathbf{P}\mathbf{S}^T) : \mathbf{S} \in \mathcal{S}\right\}. \tag{16.14}$$

Now the question is how to find a Schur-minimal vector within $\mathcal{L}$, which is, however, not easy to answer given the form of $\mathcal{L}$ in (16.14). To overcome this difficulty, we transform $\mathcal{L}$ to a more convenient equivalent form by utilizing the majorization relation.

Lemma 16.2.1. When $K \le N$, $\mathcal{L}$ is equal to

$$\mathcal{M} \triangleq \{\lambda \in \mathbb{R}^N : (\lambda_1,\ldots,\lambda_K) \succ (p_1,\ldots,p_K),\ \lambda_{K+1} = \cdots = \lambda_N = 0\}.$$

When $K > N$, $\mathcal{L}$ is equal to

$$\mathcal{N} \triangleq \{\lambda \in \mathbb{R}^N : (\lambda_1,\ldots,\lambda_N,\underbrace{0,\ldots,0}_{K-N}) \succ (p_1,\ldots,p_K)\}.$$

Proof: Consider the case $K \le N$. We first show that if $\lambda \in \mathcal{L}$ then $\lambda \in \mathcal{M}$. Since $K \le N$, $\lambda(\mathbf{S}\mathbf{P}\mathbf{S}^T)$ has at most K nonzero elements, which are denoted by λ_a. Let $\mathbf{p} \triangleq (p_i)_{i=1}^K$. Observe that, for $\mathbf{S} \in \mathcal{S}$, $\mathbf{p}$ and λ_a are the diagonal elements and the eigenvalues of the matrix $\mathbf{P}^{1/2}\mathbf{S}^T\mathbf{S}\mathbf{P}^{1/2}$, respectively. From Theorem 16.1.4 we have $\lambda_a \succ \mathbf{p}$, implying that $\lambda \in \mathcal{M}$.

To see the other direction, let $\lambda \in \mathcal{M}$ and thus $\lambda_a \succ \mathbf{p}$. According to Theorem 16.1.5, there exists a symmetric matrix $\mathbf{Z}$ with eigenvalues λ_a and diagonal elements $\mathbf{p}$. Denote the eigenvalue decomposition (EVD) of $\mathbf{Z}$ by $\mathbf{Z} = \mathbf{U}_z\mathbf{\Lambda}_z\mathbf{U}_z^T$ and introduce $\mathbf{\Lambda} = \text{diag}\{\mathbf{\Lambda}_z, \mathbf{0}_{(N-K)\times(N-K)}\}$ and $\mathbf{U} = [\mathbf{U}_z\, \mathbf{0}_{K\times(N-K)}]$. Then we can choose $\mathbf{S} = \mathbf{\Lambda}^{1/2}\mathbf{U}^T\mathbf{P}^{-1/2}$. It is easy to check that the eigenvalues of $\mathbf{S}\mathbf{P}\mathbf{S}^T = \mathbf{\Lambda}$ are λ, and $\|\mathbf{s}_i\|^2$, $i = 1,\ldots,K$, coincide with the diagonal elements of $\mathbf{S}^T\mathbf{S} = \mathbf{P}^{-1/2}\mathbf{Z}\mathbf{P}^{-1/2}$ and thus are all ones, so we have $\lambda \in \mathcal{L}$.

The equivalence between $\mathcal{L}$ and $\mathcal{N}$ when $K > N$ can be obtained in a similar way, for which a detailed proof was provided in [8]. □

When $K \leq N$, the Schur-minimal vector in $\mathcal{L}$ (or $\mathcal{M}$), from Lemma 16.2.1, is $\lambda^\star = (p_1, \ldots, p_L, 0, \ldots, 0)$, which can be achieved by choosing arbitrary K orthonormal sequences, i.e., $\mathbf{S}^T\mathbf{S} = \mathbf{I}_K$.

When $K > N$, the problem of finding a Schur-minimal vector in $\mathcal{L}$ (or $\mathcal{N}$) is, however, not straightforward. It turns out that in this case the Schur-minimal vector is given in a complicated form based on the following definition.

Definition 16.2.1. (**Oversized users** [8]) User i is defined to be oversized if

$$p_i > \frac{\sum_{i=1}^{K} p_j 1_{\{p_i > p_j\}}}{N - \sum_{j=1}^{K} 1_{\{p_j \geq p_i\}}}$$

where $1_{\{\cdot\}}$ is the indication function. Intuitively, a user is oversized if his power is large relative to those of the others.

Theorem 16.2.1. *([8]) Assume w.l.o.g. that the users are ordered according to their powers $p_1 \geq \cdots \geq p_K$, and the first L users are oversized. Then, the Schur-minimal vector in $\mathcal{L}$ (or $\mathcal{N}$) is given by*

$$\lambda^\star = \left(p_1, \ldots, p_L, \frac{\sum_{j=L+1}^{K} p_j}{N-L}, \ldots, \frac{\sum_{j=L+1}^{K} p_j}{N-L} \right).$$

The left question is how to find an $\mathbf{S} \in \mathcal{S}$ such that the eigenvalues of $\mathbf{S}\mathbf{P}\mathbf{S}^T$ are $\lambda^\star$. Note that the constraint $\mathbf{S} \in \mathcal{S}$ is equivalent to saying that the diagonal elements of $\mathbf{S}^T\mathbf{S}$ are all equal to 1. Therefore, given the optimal $\mathbf{S}$, the matrix $\mathbf{P}^{1/2}\mathbf{S}^T\mathbf{S}\mathbf{P}^{1/2}$ has the diagonal elements $\mathbf{p} = (p_1, \ldots, p_K)$ and the eigenvalues $\lambda_b = (\lambda^\star, \mathbf{0})$. From Theorem 16.1.5, there exists a $K \times K$ symmetric matrix $\mathbf{M}$ such that its diagonal elements and eigenvalues are given by $\mathbf{p}$ and λ_b, respectively. Denote the (EVD) of $\mathbf{M}$ by $\mathbf{M} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$, where $\mathbf{\Lambda} = \text{diag}\{\lambda^\star\}$ and $\mathbf{U} \in \mathbb{R}^{K \times N}$ contains the N eigenvectors corresponding to $\lambda^\star$. Then, the optimal signature sequence matrix can be obtained as $\mathbf{S} = \mathbf{\Lambda}^{1/2}\mathbf{U}^T\mathbf{P}^{-1/2}$. It can be verified that the eigenvalues of $\mathbf{S}\mathbf{P}\mathbf{S}^T$ are $\lambda^\star$ and $\mathbf{S} \in \mathcal{S}$.

Finally, to construct the symmetric matrix $\mathbf{M}$ with the diagonal elements $\mathbf{p}$ and the eigenvalues λ_b (provided $\mathbf{p} \prec \lambda_b$), one can exploit Algorithm 15 introduced in Section 16.1.3. Interestingly, an iterative algorithm was proposed in [14, 15] to generate the optimal signature sequences. This algorithm updates each user's signature sequence in a sequential way, and was proved to converge to the optimal solution.

16.2.2 Linear MIMO Transceiver Design

MIMO channels, usually arising from using multiple antennas at both ends of a wireless link, have been well recognized as an effective way to improve the capacity and reliability of wireless communications [16]. A low-complexity approach to harvest the benefits of MIMO channels is to exploit linear transceivers, i.e., a linear precoder at the transmitter and a linear equalizer at the receiver. Designing linear transceivers for MIMO channels has a long history, but mostly focused on some specific measure of the global performance. It has recently been found in [9, 17] that the design of linear MIMO transceivers can be unified by majorization theory into a general framework that embraces a wide range of different performance criteria. In the following we briefly introduce this unified framework.

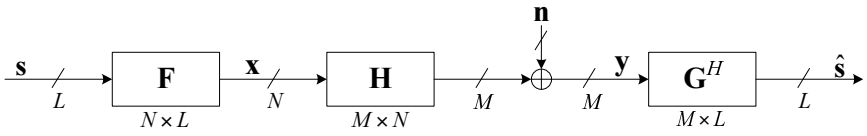

Figure 16.2: Linear MIMO transceiver consisting of a linear precoder and a linear equalizer.

Consider a communication link with N transmit and M receive antennas. The signal model of such a MIMO channel is

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n} \tag{16.15}$$

where $\mathbf{x} \in \mathbb{C}^N$ is the transmitted signal vector, $\mathbf{H} \in \mathbb{C}^{M\times N}$ is the channel matrix, $\mathbf{y} \in \mathbb{C}^M$ is the received signal vector, and $\mathbf{n} \in \mathbb{C}^M$ is a zero-mean circularly symmetric complex Gaussian random vector with covariance matrix $\mathbf{I}$,[9] i.e., $\mathbf{n} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$. In the linear transceiver scheme as illustrated in Fig. 16.2, the transmitted signal $\mathbf{x}$ results from the linear transformation of a symbol vector $\mathbf{s} \in \mathbb{C}^L$ through a linear precoder $\mathbf{F} \in \mathbb{C}^{N\times L}$ and is given by $\mathbf{x} = \mathbf{Fs}$. Assume w.l.o.g. that $L \le \min\{M, N\}$ and $E\left\{\mathbf{ss}^H\right\} = \mathbf{I}$. The total average transmit power is

$$P_T = E\left\{\|\mathbf{x}\|^2\right\} = \mathrm{Tr}(\mathbf{FF}^H). \tag{16.16}$$

At the receiver is a linear equalizer $\mathbf{G}^H \in \mathbb{C}^{L\times M}$ used to estimate $\mathbf{s}$ from $\mathbf{y}$ resulting in $\hat{\mathbf{s}} = \mathbf{G}^H\mathbf{y}$. Therefore, the relation between the transmitted symbols and the estimated symbols can be expressed as

$$\hat{\mathbf{s}} = \mathbf{G}^H\mathbf{HFs} + \mathbf{G}^H\mathbf{n}. \tag{16.17}$$

An advantage of MIMO channels is the support of simultaneously transmitting multiple data streams, leading to significant capacity improvement.[10] Observe from (16.17) that the estimated symbol at the ith data stream is given by

$$\hat{s}_i = \mathbf{g}_i^H\mathbf{Hf}_i s_i + \mathbf{g}_i^H\mathbf{n}_i \tag{16.18}$$

where $\mathbf{f}_i$ and $\mathbf{g}_i$ are the ith columns of $\mathbf{F}$ and $\mathbf{G}$, respectively, and $\mathbf{n}_i = \sum_{j\ne i}\mathbf{Hf}_j s_j + \mathbf{n}$ is the equivalent noise seen by the ith data stream with covariance matrix $\mathbf{R}_{n_i} = \sum_{j\ne i}\mathbf{Hf}_j\mathbf{f}_j^H\mathbf{H}^H + \mathbf{I}$. In practice, the performance of a data stream can be measured by the MSE, signal–to–interference-plus-noise ratio (SINR), or bit error rate (BER), which according to (16.18) are given by

$$\begin{aligned}
\mathrm{MSE}_i &\triangleq E\left\{|\hat{s}_i - s_i|^2\right\} = \left|\mathbf{g}_i^H\mathbf{Hf}_i - 1\right|^2 + \mathbf{g}_i^H\mathbf{R}_{n_i}\mathbf{g}_i \\
\mathrm{SINR}_i &\triangleq \frac{\text{desired component}}{\text{undesired component}} = \frac{\left|\mathbf{g}_i^H\mathbf{Hf}_i\right|^2}{\mathbf{g}_i^H\mathbf{R}_{n_i}\mathbf{g}_i} \\
\mathrm{BER}_i &\triangleq \frac{\#\text{ bits in error}}{\#\text{ transmitted bits}} \approx \varphi_i(\mathrm{SINR}_i)
\end{aligned}$$

where φ_i is a decreasing function relating the BER to the SINR at the ith stream [6, 9]. Any properly designed system should attempt to minimize the MSEs, maximize the SINRs, or minimize the BERs.

Measuring the global performance of a MIMO system with several data streams is tricky as there is an inherent tradeoff among the performance of the different streams. Different applications may require a different balance on the performance of the streams, so there are a variety of criteria in the literature, each leading to a particular design problem (see [6] for a survey). However, in fact, all these particular problems can be unified into one framework using the MSEs as the nominal cost. Specifically, suppose that the system performance is measured by an arbitrary global cost function of the MSEs $f_0\left(\{\mathrm{MSE}_i\}_{i=1}^L\right)$ that is increasing in each argument.[11] The linear transceiver design problem is then formulated as

$$\begin{aligned}
&\underset{\mathbf{F},\mathbf{G}}{\text{minimize}} && f_0\left(\{\mathrm{MSE}_i\}\right) \\
&\text{subject to} && \mathrm{Tr}(\mathbf{FF}^H) \le P
\end{aligned} \tag{16.19}$$

where $\mathrm{Tr}(\mathbf{FF}^H) \le P$ represents the transmit power constraint.

To solve (16.19), we first find the optimal $\mathbf{G}$ for a fixed $\mathbf{F}$. It turns out that the optimal equalizer is the LMMSE filter, also termed the Wiener filter [19]

$$\mathbf{G}^\star = (\mathbf{HFF}^H\mathbf{H}^H + \mathbf{I})^{-1}\mathbf{HF}. \tag{16.20}$$

To see this, let us introduce the MSE matrix

$$\begin{aligned}
\mathbf{E}(\mathbf{F},\mathbf{G}) &\triangleq E\left[(\hat{\mathbf{s}} - \mathbf{s})(\hat{\mathbf{s}} - \mathbf{s})^H\right] \\
&= (\mathbf{G}^H\mathbf{HF} - \mathbf{I})(\mathbf{F}^H\mathbf{H}^H\mathbf{G} - \mathbf{I}) + \mathbf{G}^H\mathbf{G}
\end{aligned} \tag{16.21}$$

[9] If the noise is not white, say with covariance matrix $\mathbf{R}_n$, one can always whiten it by $\bar{\mathbf{y}} = \mathbf{R}_n^{-1/2}\mathbf{y} = \bar{\mathbf{H}}\mathbf{x} + \bar{\mathbf{n}}$, where $\bar{\mathbf{H}} = \mathbf{R}_n^{-1/2}\mathbf{H}$ is the equivalent channel.

[10] This kind of improvement is often called the multiplexing gain [18].

[11] The increasingness of f is a mild and reasonable assumption: if the performance of one stream improves, the global performance should improve too.

from which the MSE of the ith data stream is given by $\mathrm{MSE}_i = [\mathbf{E}]_{ii}$. It is not difficult to verify that

$$\mathbf{E}(\mathbf{F}, \mathbf{G}^\star) = (\mathbf{I} + \mathbf{F}^H\mathbf{H}^H\mathbf{H}\mathbf{F})^{-1} \preceq \mathbf{E}(\mathbf{F}, \mathbf{G}) \tag{16.22}$$

for any $\mathbf{G}$, meaning that $\mathbf{G}^\star$ simultaneously minimizes all diagonal elements of $\mathbf{E}$ or all MSEs. At the same time, one can verify that $\mathbf{g}_i^\star$, i.e., the ith column of $\mathbf{G}^\star$, also maximizes SINR_i (or equivalently minimizes BER_i) [6, 9]. Hence, the Wiener filter is optimal in the sense of both minimizing all MSEs and maximizing all SINRs (or minimizing all BERs). Observe that the optimality is regardless of the particular choice of the cost function f_0 in (16.19).

Using the Wiener filter as the equalizer, we can easily obtain

$$\begin{aligned} \mathrm{MSE}_i &= [(\mathbf{I} + \mathbf{F}^H\mathbf{H}^H\mathbf{H}\mathbf{F})^{-1}]_{ii} \\ \mathrm{SINR}_i &= \frac{1}{\mathrm{MSE}_i} - 1 \\ \mathrm{BER}_i &= \varphi_i(\mathrm{MSE}_i^{-1} - 1). \end{aligned}$$

This means that different performance measures based on the MSEs, the SINRs, or the BERs can be uniformly represented by the MSE-based criteria, thus indicating the generality of the problem formulation (16.19).

Now, the transceiver design problem (16.19) reduces to the following precoder design problem:

$$\begin{array}{ll} \underset{\mathbf{F}}{\text{minimize}} & f_0\left(\{[(\mathbf{I} + \mathbf{F}^H\mathbf{H}^H\mathbf{H}\mathbf{F})^{-1}]_{ii}\}\right) \\ \text{subject to} & \mathrm{Tr}(\mathbf{F}\mathbf{F}^H) \le P. \end{array} \tag{16.23}$$

Solving such a general problem is very challenging and the solution hinges on majorization theory.

Theorem 16.2.2. *([6, Theorem 3.13]) Suppose that the cost function $f_0 : \mathbb{R}^L \mapsto \mathbb{R}$ is increasing in each argument. Then, the optimal solution to (16.23) is given by*

$$\mathbf{F}^\star = \mathbf{V}_h \mathrm{diag}(\sqrt{\mathbf{p}})\mathbf{\Omega}$$

where

(i) $\mathbf{V}_h \in \mathbb{C}^{N\times L}$ *is a semi-unitary matrix with columns equal to the right singular vectors of* $\mathbf{H}$ *corresponding to the L largest singular values in increasing order;*

(ii) $\mathbf{p} \in \mathbb{R}_+^L$ *is the solution to the following power allocation problem:*[12]

$$\begin{array}{ll} \underset{\mathbf{p},\rho}{\text{minimize}} & f_0\left(\rho_1, \ldots, \rho_L\right) \\ \text{subject to} & \left(\frac{1}{1+p_1\gamma_1}, \ldots, \frac{1}{1+p_L\gamma_L}\right) \succ^w (\rho_1, \ldots, \rho_L) \\ & \mathbf{p} \ge \mathbf{0},\ \mathbf{1}^T\mathbf{p} \le P \end{array} \tag{16.24}$$

where $\{\gamma_i\}_{i=1}^L$ are the L largest eigenvalues of $\mathbf{H}^H\mathbf{H}$ in increasing order;

(iii) $\mathbf{\Omega} \in \mathbb{C}^{L\times L}$ *is a unitary matrix such that $[(\mathbf{I}+\mathbf{F}^{\star H}\mathbf{H}^H\mathbf{H}\mathbf{F}^\star)^{-1}]_{ii} = \rho_i$ for all i, which can be computed with Algorithm 15.*

Proof: We start by rewriting (16.23) into the equivalent form

$$\begin{array}{ll} \underset{\mathbf{F},\rho}{\text{minimize}} & f_0\left(\rho\right) \\ \text{subject to} & \mathbf{d}\left((\mathbf{I} + \mathbf{F}^H\mathbf{H}^H\mathbf{H}\mathbf{F})^{-1}\right) \le \rho \\ & \mathrm{Tr}(\mathbf{F}\mathbf{F}^H) \le P \end{array} \tag{16.25}$$

Note that, given any $\mathbf{F}$, we can always find another $\tilde{\mathbf{F}} = \mathbf{F}\mathbf{\Omega}^H$ with a unitary matrix $\mathbf{\Omega}$ such that $\tilde{\mathbf{F}}^H\mathbf{H}^H\mathbf{H}\tilde{\mathbf{F}} = \mathbf{\Omega}\mathbf{F}^H\mathbf{H}^H\mathbf{H}\mathbf{F}\mathbf{\Omega}^H$ is diagonal with diagonal elements in increasing order. The original MSE matrix is given by $(\mathbf{I} +$

[12] $\succ^w$ denotes weak supermajorization (see Definition 16.1.3)

$\mathbf{F}^H\mathbf{H}^H\mathbf{H}\mathbf{F})^{-1} = \mathbf{\Omega}^H(\mathbf{I}+\tilde{\mathbf{F}}^H\mathbf{H}^H\mathbf{H}\tilde{\mathbf{F}})^{-1}\mathbf{\Omega}$. Thus we can rewrite (16.25) in terms of $\tilde{\mathbf{F}}$ and $\mathbf{\Omega}$ as

$$\begin{array}{ll} \underset{\tilde{\mathbf{F}},\mathbf{\Omega},\rho}{\text{minimize}} & f_0\left(\rho\right) \\ \text{subject to} & \tilde{\mathbf{F}}^H\mathbf{H}^H\mathbf{H}\tilde{\mathbf{F}} \quad \text{diagonal} \\ & \mathbf{d}\left(\mathbf{\Omega}^H(\mathbf{I}+\tilde{\mathbf{F}}^H\mathbf{H}^H\mathbf{H}\tilde{\mathbf{F}})^{-1}\mathbf{\Omega}\right) \leq \rho \\ & \mathrm{Tr}(\tilde{\mathbf{F}}\tilde{\mathbf{F}}^H) \leq P. \end{array} \tag{16.26}$$

It follows from Lemma 16.1.1 and Corollary 16.1.2 that, for a given $\tilde{\mathbf{F}}$, we can always find a feasible $\mathbf{\Omega}$ if and only if

$$\lambda\left((\mathbf{I}+\tilde{\mathbf{F}}^H\mathbf{H}^H\mathbf{H}\tilde{\mathbf{F}})^{-1}\right) \succ^w \rho.$$

Therefore, using the diagonal property of $\tilde{\mathbf{F}}^H\mathbf{H}^H\mathbf{H}\tilde{\mathbf{F}}$, (16.26) is equivalent to

$$\begin{array}{ll} \underset{\tilde{\mathbf{F}},\rho}{\text{minimize}} & f_0\left(\rho\right) \\ \text{subject to} & \tilde{\mathbf{F}}^H\mathbf{H}^H\mathbf{H}\tilde{\mathbf{F}} \quad \text{diagonal} \\ & \mathbf{d}\left((\mathbf{I}+\tilde{\mathbf{F}}^H\mathbf{H}^H\mathbf{H}\tilde{\mathbf{F}})^{-1}\right) \succ^w \rho \\ & \mathrm{Tr}(\tilde{\mathbf{F}}\tilde{\mathbf{F}}^H) \leq P. \end{array} \tag{16.27}$$

Given that $\tilde{\mathbf{F}}^H\mathbf{H}^H\mathbf{H}\tilde{\mathbf{F}}$ is diagonal with diagonal elements in increasing order, we can invoke [9, Lemma 12] or [6, Lemma 3.16] to conclude that the optimal $\tilde{\mathbf{F}}$ can be written as $\tilde{\mathbf{F}} = \mathbf{V}_h\mathrm{diag}(\sqrt{\mathbf{p}})$, implying that $\mathbf{F}^\star = \mathbf{V}_h\mathrm{diag}(\sqrt{\mathbf{p}})\mathbf{\Omega}$. Now, by using the weak supermajorization relation as well as the structure $\tilde{\mathbf{F}} = \mathbf{V}_h\mathrm{diag}(\sqrt{\mathbf{p}})$, (16.27) can be expressed as (16.24). □

If, in addition, f_0 is minimized when the arguments are sorted in decreasing order,[13] then (16.24) can be explicitly written as

$$\begin{array}{ll} \underset{\mathbf{p},\rho}{\text{minimize}} & f_0\left(\rho_1,\ldots,\rho_L\right) \\ \text{subject to} & \sum_{j=i}^{L}\frac{1}{1+p_i\gamma_i} \leq \sum_{j=i}^{L}\rho_j, \quad 1\leq i\leq L \\ & \rho_i \geq \rho_{i+1}, \quad 1\leq i\leq L-1 \\ & \mathbf{p}\geq\mathbf{0},\ \mathbf{1}^T\mathbf{p}\leq P \end{array} \tag{16.28}$$

which is a convex problem if f_0 is a convex function, and thus can be efficiently solved in polynomial time [20]. In fact, the optimal precoder can be further simplified or even obtained in closed form, when the objective f_0 falls into the class of Schur-convex/concave functions.

Corollary 16.2.1. ([9, Theorem 1]) Suppose that the cost function $f_0 : \mathbb{R}^L \mapsto \mathbb{R}$ is increasing in each argument.

(i) If f_0 is Schur-concave, then the optimal solution to (16.23) is given by

$$\mathbf{F}^\star = \mathbf{V}_h\mathrm{diag}(\sqrt{\mathbf{p}})$$

where $\mathbf{p}$ is the solution to the following power allocation problem:

$$\begin{array}{ll} \underset{\mathbf{p}}{\text{minimize}} & f_0\left(\{(1+p_i\gamma_i)^{-1}\}_{i=1}^L\right) \\ \text{subject to} & \mathbf{p}\geq\mathbf{0},\ \mathbf{1}^T\mathbf{p}\leq P. \end{array}$$

(ii) If f_0 is Schur-convex, then the optimal solution to (16.23) is given by

$$\mathbf{F}^\star = \mathbf{V}_h\mathrm{diag}(\sqrt{\mathbf{p}})\mathbf{\Omega}$$

where the power allocation $\mathbf{p}$ is given by

$$p_i = (\mu\gamma_i^{-1/2} - \gamma_i^{-1})^+, \quad 1\leq i\leq L$$

with μ chosen to satisfy $\mathbf{1}^T\mathbf{p} = P$, and $\mathbf{\Omega}$ is a unitary matrix such that $(\mathbf{I}+\mathbf{F}^{\star H}\mathbf{H}^H\mathbf{H}\mathbf{F}^\star)^{-1}$ has equal diagonal elements. $\mathbf{\Omega}$ can be any unitary matrix satisfying $|[\mathbf{\Omega}]_{ik}| = |[\mathbf{\Omega}]_{il}|$, $\forall i,k,l$, such as the unitary DFT matrix or the unitary Hadamard matrix (see Lemma 16.1.3).

[13] In practice, most cost functions are minimized when the arguments are in a specific ordering (if not, one can always use instead the function $\tilde{f}_0(\mathbf{x}) = \min_{\mathbf{P}\in\mathcal{P}} f_0(\mathbf{P}\mathbf{x})$, where $\mathcal{P}$ is the set of all permutation matrices) and, hence, the decreasing order can be taken without loss of generality.

Although Schur-convex/concave functions do not form a partition of all L-dimensional functions, they do cover most of the frequently used global performance measures. An extensive account of Schur-convexity/concavity of common performance measures was provided in [6] and [9] (see also Exercise 16.4.3). For Schur-concave functions, a nice property is that the MIMO channel is fully diagonalized by the optimal transceiver, whereas for Schur-convex functions, the channel is diagonalized subject to a specific rotation $\mathbf{\Omega}$ on the transmit symbols.

16.2.3 Nonlinear MIMO Transceiver Design

In this section, we introduce another paradigm of MIMO transceivers, consisting of a linear precoder and a nonlinear decision feedback equalizer (DFE). The DFE differs from the linear equalizer in that the DFE exploits the finite alphabet property of digital signals and recovers signals successively. Thus, the nonlinear decision feedback (DF) MIMO transceivers usually enjoy superior performance than the linear transceivers. Using majorization theory, the DF MIMO transceiver designs can also be unified, mainly based on the recent results in [11, 21, 22], into a general framework covering diverse design criteria, as was derived independently in [6, 23, 24]. Different from the linear transceiver designs that are based on additive majorization, the DF transceiver designs rely mainly on multiplicative majorization (see Section 16.1.4).

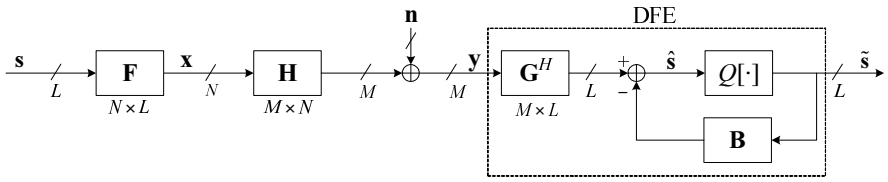

Figure 16.3: Nonlinear MIMO transceiver consisting of a linear precoder and a decision feedback equalizer (DFE).

Considering the MIMO channel in (16.15), we use a linear precoder $\mathbf{F} \in \mathbb{C}^{N\times L}$ at the transmitter to generate the transmitted signal $\mathbf{x} = \mathbf{Fs}$ from a symbol vector $\mathbf{s}$ satisfying $E\left\{\mathbf{ss}^H\right\} = \mathbf{I}$. For simplicity, we assume that $L \le \text{rank}(\mathbf{H})$. The receiver exploits, instead of a linear equalizer, a DFE that detects the symbols successively with the Lth symbol (s_L) detected first and the first symbol (s_1) detected last. As shown in Fig. 16.3, a DFE consists of two components: a feed-forward filter $\mathbf{G}^H \in \mathbb{C}^{L\times M}$ applied to the received signal $\mathbf{y}$, and a feedback filter $\mathbf{B} \in \mathbb{C}^{L\times L}$ that is a strictly upper triangular matrix and feeds back the previously detected symbols. The block $Q[\cdot]$ represents the mapping from the "analog" estimated $\hat{s}_i$ to the closest "digital" point in the signal constellation. Assuming no error propagation,[14] the "analog" estimated $\hat{s}_i$ can be written as

$$\hat{s}_i = \mathbf{g}_i^H\mathbf{y} - \sum_{j=i+1}^{L} b_{ij}x_j, \quad 1 \le i \le L \tag{16.29}$$

where $\mathbf{g}_i$ is the ith column of $\mathbf{G}$ and $b_{ij} = [\mathbf{B}]_{ij}$. Compactly, the estimated symbol vector can be written as

$$\hat{\mathbf{s}} = \mathbf{G}^H\mathbf{y} - \mathbf{Bs} = (\mathbf{G}^H\mathbf{HF} - \mathbf{B})\mathbf{s} + \mathbf{G}^H\mathbf{n}. \tag{16.30}$$

Let $\mathbf{f}_i$ be the ith column of $\mathbf{F}$. The performance of the ith data stream can be measured by the MSE or the SINR as

$$\begin{aligned}
\text{MSE}_i &\triangleq E\left\{\left|\hat{s}_i - s_i\right|^2\right\} \\
&= \left|\mathbf{g}_i^H\mathbf{Hf}_i - 1\right|^2 + \sum_{j=i+1}^{L}\left|\mathbf{g}_i^H\mathbf{Hf}_j - b_{ij}\right|^2 + \sum_{j=1}^{i-1}\left|\mathbf{g}_i^H\mathbf{Hf}_j\right|^2 + \|\mathbf{g}_i\|^2 \\
\text{SINR}_i &\triangleq \frac{\text{desired component}}{\text{undesired component}} \\
&= \frac{\left|\mathbf{g}_i^H\mathbf{Hf}_i\right|^2}{\sum_{j=i+1}^{L}\left|\mathbf{g}_i^H\mathbf{Hf}_j - b_{ij}\right|^2 + \sum_{j=1}^{i-1}\left|\mathbf{g}_i^H\mathbf{Hf}_j\right|^2 + \|\mathbf{g}_i\|^2}.
\end{aligned}$$

[14] Error propagation means that if the detection is erroneous, it may cause more errors in the subsequent detections. By using powerful coding techniques, the influence of error propagation can be made negligible.

Alternatively, the performance can also be measured by $\mathrm{BER}_i \approx \varphi_i(\mathrm{SINR}_i)$ with a decreasing function φ_i. Similar to the linear transceiver case, we consider that the system performance is measured by a global cost function of the MSEs $f_0\left(\{\mathrm{MSE}_i\}_{i=1}^L\right)$ that is increasing in each argument. Then the nonlinear DF MIMO transceiver design is formulated as the following problem:

$$\begin{array}{ll} \underset{\mathbf{F},\mathbf{G},\mathbf{B}}{\text{minimize}} & f_0\left(\{\mathrm{MSE}_i\}\right) \\ \text{subject to} & \mathrm{Tr}(\mathbf{F}\mathbf{F}^H) \le P \end{array} \tag{16.31}$$

where $\mathrm{Tr}(\mathbf{F}\mathbf{F}^H) \le P$ denotes the transmit power constraint.

It is easily seen that to minimize MSE_i, the DF coefficients should be $b_{ij} = \mathbf{g}_i^H \mathbf{H}\mathbf{f}_j$, $1 \le i < j \le L$, or, equivalently,

$$\mathbf{B} = \mathcal{U}(\mathbf{G}^H \mathbf{H}\mathbf{F}) \tag{16.32}$$

where $\mathcal{U}(\cdot)$ stands for keeping the strictly upper triangular entries of the matrix while setting the others zero. To obtain the optimal feed-forward filter, we let $\mathbf{W} \triangleq \mathbf{H}\mathbf{F}$ be the effective channel, and denote by $\mathbf{W}_i \in \mathbb{C}^{M\times i}$ the submatrix consisting of the first i columns of $\mathbf{W}$ and by $\mathbf{w}_i$ the ith column of $\mathbf{W}$. Then, with $b_{ij} = \mathbf{g}_i^H \mathbf{H}\mathbf{f}_j$, the feed-forward filter minimizing MSE_i is given by [6, Sec. 4.3]

$$\mathbf{g}_i = (\mathbf{W}_i \mathbf{W}_i^H + \mathbf{I})^{-1}\mathbf{w}_i, \quad 1 \le i \le L. \tag{16.33}$$

In fact, there is a more computationally efficient expression of the optimal DFE given as follows.

Lemma 16.2.2. ([25]) Let the QR decomposition of the augmented matrix be

$$\mathbf{W}_a \triangleq \left[\begin{array}{c} \mathbf{W} \\ \mathbf{I}_L \end{array}\right]_{(M+L)\times L} = \mathbf{Q}\mathbf{R}$$

and partition $\mathbf{Q}$ into

$$\mathbf{Q} = \left[\begin{array}{c} \bar{\mathbf{Q}} \\ \underline{\mathbf{Q}} \end{array}\right]$$

where $\bar{\mathbf{Q}} \in \mathbb{C}^{M\times L}$ and $\underline{\mathbf{Q}} \in \mathbb{C}^{L\times L}$. The optimal feed-forward and feedback matrices that minimize the MSEs are

$$\mathbf{G}^\star = \bar{\mathbf{Q}}\mathbf{D}_R^{-1} \quad \text{and} \quad \mathbf{B}^\star = \mathbf{D}_R^{-1}\mathbf{R} - \mathbf{I} \tag{16.34}$$

where $\mathbf{D}_R$ is a diagonal matrix with the same diagonal elements as $\mathbf{R}$. The resulting MSE matrix is diagonal:

$$\mathbf{E} \triangleq E\left[(\hat{\mathbf{s}} - \mathbf{s})(\hat{\mathbf{s}} - \mathbf{s})^H\right] = \mathbf{D}_R^{-2}.$$

By using the optimal DFE in (16.34), the MSE and the SINR at the ith data stream are related by

$$\mathrm{SINR}_i = \frac{1}{\mathrm{MSE}_i} - 1$$

which is the same as in the linear equalizer case. Therefore, we can focus w.l.o.g. on the MSE-based performance measures, which, according to Lemma 16.2.2, depend on the diagonal elements of $\mathbf{R}$. The optimal precoder is then given by the solution to the following problem:

$$\begin{array}{ll} \underset{\mathbf{F}}{\text{minimize}} & f_0\left(\{[\mathbf{R}]_{ii}^{-2}\}\right) \\ \text{subject to} & \left[\begin{array}{c} \mathbf{H}\mathbf{F} \\ \mathbf{I}_L \end{array}\right] = \mathbf{Q}\mathbf{R} \\ & \mathrm{Tr}(\mathbf{F}\mathbf{F}^H) \le P. \end{array} \tag{16.35}$$

This complicated optimization can be simplified by using multiplicative majorization.

Theorem 16.2.3. *([6, Theorem 4.3]) Suppose that the cost function $f_0 : \mathbb{R}^L \mapsto \mathbb{R}$ is increasing in each argument. Then, the optimal solution to (16.35) is given by*

$$\mathbf{F}^\star = \mathbf{V}_h \mathrm{diag}(\sqrt{\mathbf{p}})\mathbf{\Omega}$$

where

(i) $\mathbf{V}_h \in \mathbb{C}^{N\times L}$ *is a semi-unitary matrix with columns equal to the right singular vectors of matrix* $\mathbf{H}$ *corresponding to the* L *largest singular values in increasing order;*

(ii) $\mathbf{p} \in \mathbb{R}_+^L$ *is the solution to the following power allocation problem:*

$$\begin{array}{ll} \underset{\mathbf{p},\mathbf{r}}{\text{minimize}} & f_0\left(r_1^{-2},\ldots,r_L^{-2}\right) \\ \text{subject to} & \left(r_1^2,\ldots,r_L^2\right) \prec_\times (1+p_1\gamma_1,\ldots,1+p_L\gamma_L) \\ & \mathbf{p}\geq \mathbf{0},\ \mathbf{1}^T\mathbf{p}\leq P \end{array} \tag{16.36}$$

where $\{\gamma_i\}_{i=1}^L$ *are the* L *largest eigenvalues of* $\mathbf{H}^H\mathbf{H}$ *in decreasing order;*

(iii) $\mathbf{\Omega} \in \mathbb{C}^{L\times L}$ *is a unitary matrix such that the matrix* $\mathbf{R}$ *in the QR decomposition*

$$\begin{bmatrix} \mathbf{H}\mathbf{F}^\star \\ \mathbf{I}_L \end{bmatrix} = \mathbf{Q}\mathbf{R}$$

has diagonal elements $\{r_i\}_{i=1}^L$*. To obtain* $\mathbf{\Omega}$*, it suffices to compute the generalized triangular decomposition (GTD) [11]*

$$\begin{bmatrix} \mathbf{H}\mathbf{V}_h\mathrm{diag}(\sqrt{\mathbf{p}}) \\ \mathbf{I}_L \end{bmatrix} = \mathbf{Q}_J\mathbf{R}\mathbf{P}_J^H$$

and then set $\mathbf{\Omega} = \mathbf{P}_J$*.*

Proof: The proof is involved, so we provide only a sketch of it and refer the interested read to [6, Appendix 4.C] for the detailed proof.

Denote the diagonal elements of $\mathbf{R}$ by $\{r_i\}_{i=1}^L$, and the singular values of the effective channel $\mathbf{W}$ and the augmented matrix $\mathbf{W}_a$ by $\{\sigma_{w,1}\}_{i=1}^L$ and $\{\sigma_{w_a,1}\}_{i=1}^L$ in decreasing order, respectively. One can easily see that

$$\sigma_{w_a,i} = \sqrt{1+\sigma_{w,i}^2} \quad 1\leq i\leq L.$$

Consider the SVD $\mathbf{F} = \mathbf{U}_f\mathrm{diag}(\sqrt{\mathbf{p}})\mathbf{\Omega}$. By using Theorem 16.1.6 on the GTD, one can prove that there exists an $\mathbf{\Omega}$ such that $\mathbf{W}_a = \mathbf{Q}\mathbf{R}$ if and only if $\{r_i^2\} \prec_\times \{\sigma_{w_a,i}^2\}$ [6, Lemma 4.9]. Therefore, the constraint

$$\begin{bmatrix} \mathbf{H}\mathbf{F} \\ \mathbf{I}_L \end{bmatrix} = \mathbf{Q}\mathbf{R}$$

can be equivalently replaced by

$$\left(r_1^2,\ldots,r_L^2\right) \prec_\times \left(\sigma_{w_a,1}^2,\ldots,\sigma_{w_a,L}^2\right).$$

Next, by showing that

$$\prod_{i=1}^k \sigma_{w_a,i}^2 = \prod_{i=1}^k (1+\sigma_{w,i}^2) \leq \prod_{i=1}^k (1+\gamma_i p_i), \quad 1\leq k\leq L$$

where the equality holds if and only if $\mathbf{U}_f = \mathbf{V}_h$, one can conclude that the optimal $\mathbf{F}$ occurs when $\mathbf{U}_f = \mathbf{V}_h$. □

Theorem 16.2.3 shows the solution to the general problem with an arbitrary cost function has a nice structure. In fact, when the composite objective function

$$f_0 \circ \exp(\mathbf{x}) \triangleq f_0(e^{x_1},\ldots,e^{x_L}) \tag{16.37}$$

is either Schur-convex or Schur-concave, the nonlinear DF transceiver design problem admits a simpler or even closed-form solution.

Corollary 16.2.2. ([6, Theorem 4.4]) Suppose that the cost function $f_0 : \mathbb{R}^L \mapsto \mathbb{R}$ is increasing in each argument.

(i) If $f_0 \circ \exp$ is Schur-concave, then the optimal solution to (16.35) is given by

$$\mathbf{F}^\star = \mathbf{V}_h\mathrm{diag}(\sqrt{\mathbf{p}})$$

where $\mathbf{p}$ is the solution to the following power allocation problem:

$$\begin{array}{ll} \underset{\mathbf{p}}{\text{minimize}} & f_0\left(\{(1+\gamma_i p_i)^{-1}\}_{i=1}^L\right) \\ \text{subject to} & \mathbf{p}\geq\mathbf{0},\ \mathbf{1}^T\mathbf{p}\leq P. \end{array}$$

(ii) If $f_0 \circ \exp$ is Schur-convex, then the optimal solution to (16.35) is given by

$$\mathbf{F}^\star = \mathbf{V}_h \mathrm{diag}(\sqrt{\mathbf{p}})\mathbf{\Omega}$$

where the power allocation $\mathbf{p}$ is given by

$$p_i = (\mu - \gamma_i^{-1})^+, \quad 1 \le i \le L$$

with μ chosen to satisfy $\mathbf{1}^T\mathbf{p} = P$, and $\mathbf{\Omega}$ is a unitary matrix such that the QR decomposition

$$\begin{bmatrix} \mathbf{H}\mathbf{F}^\star \\ \mathbf{I}_L \end{bmatrix} = \mathbf{Q}\mathbf{R}$$

yields $\mathbf{R}$ with equal diagonal elements.

It is interesting to relate the linear and nonlinear DF transceivers by the Schur-convexity/concavity of the cost function. From Lemma 16.1.5, $f_0 \circ \exp$ being Schur-concave implies that f_0 is Schur-concave, but not vice versa. From Lemma 16.1.4, if f_0 is Schur-convex, then $f_0 \circ \exp$ is also Schur-convex, but not vice versa. The examples of the cost function for which $f_0 \circ \exp$ is either Schur-concave or Schur-convex were provided in [6] (see also Exercise 16.4.4 and a recent survey [26]).

16.2.4 Impact of Correlation

A measure of correlation

Consider two n-dimensional random vectors $\mathbf{x}$ and $\mathbf{y}$ following the same family/class of distributions with zero means and covariance matrices $\mathbf{R}_x$ and $\mathbf{R}_y$, respectively. One question arising in many practical scenarios is how to compare $\mathbf{x}$ and $\mathbf{y}$ in terms of the degree of correlation. Majorization provides a natural way to measure correlation of a random vector.

Definition 16.2.2. ([7, Sec. 4.1.2]) Let $\lambda(\mathbf{A})$ denote the eigenvalues of a positive semidefinite matrix $\mathbf{A}$. Then, we say $\mathbf{x}$ is more correlated than $\mathbf{y}$, or the covariance matrix $\mathbf{R}_x$ is more correlated than $\mathbf{R}_y$, if $\lambda(\mathbf{R}_x) \succ \lambda(\mathbf{R}_y)$.

Note that comparing $\mathbf{x}$ and $\mathbf{y}$ (or equivalently $\mathbf{R}_x$ and $\mathbf{R}_y$) through the majorization ordering imposes an implicit constraint on $\mathbf{R}_x$ and $\mathbf{R}_y$ that requires $\sum_{i=1}^n \lambda_i(\mathbf{R}_x) = \sum_{i=1}^n \lambda_i(\mathbf{R}_y)$, or equivalently, $\mathrm{Tr}(\mathbf{R}_x) = \mathrm{Tr}(\mathbf{R}_y)$. This requirement is actually quite reasonable. If we consider $E\left\{|x_i|^2\right\}$ as the "power" of the ith element of $\mathbf{x}$, then $\mathrm{Tr}(\mathbf{R}_x) = \sum_{i=1}^n E\left\{|x_i|^2\right\}$ is the sum power of all elements of $\mathbf{x}$. Therefore, the comparison is conducted in a fair sense that the sum power of the two vectors is equal. Nevertheless, Definition 16.2.2 can be generalized to the case where $\mathrm{Tr}(\mathbf{R}_x) \neq \mathrm{Tr}(\mathbf{R}_y)$ by using weak majorization.

From Example 16.1.1, the most uncorrelated covariance matrix has equal eigenvalues, whereas the most correlated covariance matrix has only one non-zero eigenvalue. In the next, we demonstrate through several examples how to use majorization theory along with Definition 16.2.2 to analyze the effect of correlation on communication systems.

Colored Noise in CDMA Systems

Consider the uplink of a single-cell synchronous CDMA system with K users and processing gain N similar to the one that has been considered in Section 16.2.1 but with colored noise. More exactly, with the received signal at the base station given by (16.8), the zero-mean noise $\mathbf{n}$ is now correlated with the covariance matrix $\mathbf{R}_n$. In this case, the sum capacity of the CDMA system is given by [27]

$$C_{\mathrm{sum}} = \frac{1}{2}\log\det\left(\mathbf{I}_N + \mathbf{R}_n^{-1}\mathbf{S}\mathbf{P}\mathbf{S}^T\right) \tag{16.38}$$

where $\mathbf{S}$ is the signature sequence matrix and $\mathbf{P} = \mathrm{diag}\{p_1, \ldots, p_K\}$ contains the received power of each user. The maximum sum capacity is obtained as $C_{\mathrm{opt}} = \max_{\mathbf{S}\in\mathcal{S}} C_{\mathrm{sum}}$, where $\mathcal{S}$ is defined in (16.13).

Denote the EVD of $\mathbf{R}_n$ by $\mathbf{R}_n = \mathbf{U}_n\mathbf{\Lambda}_n\mathbf{U}_n^H$ with eigenvalues $\sigma_1^2, \ldots, \sigma_N^2$. Then, C_{opt} can be characterized as follows.

Lemma 16.2.3. ([27, Lemma 2.2]) The maximum sum capacity of the synchronous CDMA system with colored noise is given by

$$C_{\mathrm{opt}} = \max_{\mathbf{S}\in\mathcal{S}} \frac{1}{2}\sum_{i=1}^N \log\left(1 + \frac{\lambda_i(\mathbf{S}\mathbf{P}\mathbf{S}^T)}{\sigma_i^2}\right). \tag{16.39}$$

Proposition 16.2.1. C_{opt} obtained in (16.39) is Schur-convex in $\sigma^2 \triangleq (\sigma_1^2, \ldots, \sigma_N^2)$.

Proof: Let $\phi(\sigma^2) = \frac{1}{2}\sum_{i=1}^N \log\left(1 + \lambda_i/\sigma_i^2\right)$. Since $g(x_i) = \log\left(1 + \lambda_i/x_i\right)$ is a convex function and $f(\mathbf{x}) = \frac{1}{2}\sum_{i=1}^N x_i$ is increasing and Schur-convex, it follows from Proposition 16.1.2 that $\phi(\sigma^2) = f(g(\sigma_1^2), \ldots, g(\sigma_N^2))$ is Schur-convex. Therefore, given $\sigma_a^2 \prec \sigma_b^2$, we have

$$C_{\text{opt}}(\sigma_a^2) = \max_{\mathbf{S}\in\mathcal{S}} \phi(\sigma_a^2) \le \max_{\mathbf{S}\in\mathcal{S}} \phi(\sigma_b^2) = C_{\text{opt}}(\sigma_b^2).$$

□

Proposition 16.2.1 indicates that the more correlated (according to Definition 16.2.2) the noise is, the higher the sum capacity could be. Intuitively, if one of the noise variances, say σ_N^2, is much larger than the rest, the users can avoid using signals in the direction of $\mathbf{R}_n$ corresponding to σ_N^2 and benefit from a reduced average noise variance (since the sum of all variances keeps unchanged). Apparently, white noise with equal $\sigma_i^2 = \sigma^2$, $i = 1, \ldots, N$, is one of the worst cases that lead to the minimum C_{opt}.

Spatial Correlation in MISO channels

A multiple-input single-output (MISO) channel usually arises in using multiple transmit antennas and a single receive antenna in a wireless link. Consider a block-flat-fading[15] MISO channel with N transmit antennas. The channel model is given by

$$y = \mathbf{x}^H\mathbf{h} + n \tag{16.40}$$

where $\mathbf{x} \in \mathbb{C}^N$ is the transmitted signal, $y \in \mathbb{C}$ is the received signal, the complex Gaussian noise n has zero mean and variance σ^2, and the channel $\mathbf{h} \in \mathbb{C}^N$ is a circular symmetric Gaussian random vector with zero-mean and covariance matrix $\mathbf{R}_h$, i.e., $\mathbf{h} \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}_h)$.

In MISO (as well as MIMO) channels, the transmit strategy is determined by the transmit covariance matrix $\mathbf{Q} = E\left\{\mathbf{x}\mathbf{x}^H\right\}$. Denote the EVD of $\mathbf{Q}$ by $\mathbf{Q} = \mathbf{U}_q\mathbf{\Lambda}_q\mathbf{U}_q^H$ with the diagonal matrix $\mathbf{\Lambda}_q = \text{diag}\{p_1, \ldots, p_N\}$. Then, the eigenvectors of $\mathbf{Q}$, i.e., the columns of $\mathbf{U}_q$, can be regarded as the transmit directions, and the eigenvalue p_i represents the power allocated to the ith data stream or eigenmode. Assuming that the receiver knows the channel perfectly and the transmitter uses a Gaussian codebook with zero mean and covariance matrix $\mathbf{Q}$, the maximum average mutual information, also termed the ergodic capacity, of the MISO channel is given by [16]

$$C = \max_{\mathbf{Q}\in\mathcal{Q}} E\left[\log(1 + \gamma\mathbf{h}^H\mathbf{Q}\mathbf{h})\right] \tag{16.41}$$

where γ is the signal-to-noise ratio, $\mathcal{Q} \triangleq \{\mathbf{Q} : \mathbf{Q} \succeq \mathbf{0},\ \text{Tr}(\mathbf{Q}) = 1\}$ represents the normalized transmit power constraint, and the expectation is taken over $\mathbf{h}$.

The ergodic capacity depends on what kind of channel state information at the transmitter (CSIT) is available. In the following, we consider three types of CSIT:

- No CSIT. Neither $\mathbf{h}$ nor its statistics are known by the transmitter, and it is usually assumed that $\mathbf{h} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$, i.e., $\mathbf{R}_h = \mathbf{I}$ (see Section 16.2.5).
- Perfect CSIT. That is, $\mathbf{h}$ is perfectly known by the transmitter.
- Imperfect CSIT with covariance feedback. In this case, it is assumed that $\mathbf{h} \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}_h)$ with $\mathbf{R}_h$ known by the transmitter.

Denote the EVD of $\mathbf{R}_h$ by $\mathbf{R}_h = \mathbf{U}_h\mathbf{\Lambda}_h\mathbf{U}_h^H$ with eigenvalues $\mu_1 \ge \cdots \ge \mu_N$ sorted w.l.o.g. in decreasing order, and let $w_1, \ldots, w_N$ be standard exponentially iid random variables. In the case of no CSIT, the optimal transmit covariance matrix is given by $\mathbf{Q} = \frac{1}{N}\mathbf{I}$ [16], which results in

$$C_{\text{noCSIT}}(\mu) = E\left[\log\left(1 + \frac{\gamma}{N}\sum_{i=1}^N \mu_i w_i\right)\right]. \tag{16.42}$$

[15] Block flat-fading means that the channel keeps unchanged for a block of T symbols, and then the channel changes to an uncorrelated channel realization.

With perfect CSIT, the optimal $\mathbf{Q}$ is given by $\mathbf{Q} = \mathbf{h}\mathbf{h}^H / \|\mathbf{h}\|^2$ [16], leading to

$$C_{\text{pCSIT}}(\mu) = E\left[\log\left(1 + \gamma \sum_{i=1}^{N} \mu_i w_i\right)\right]. \tag{16.43}$$

For imperfect CSIT with covariance feedback (i.e., $\mathbf{R}_h$ known), the optimal $\mathbf{Q}$ is given in the form $\mathbf{Q} = \mathbf{U}_h \mathbf{\Lambda}_q \mathbf{U}_h^H$ [28], so the ergodic capacity is obtained by

$$C_{\text{cfCSIT}}(\mu) = \max_{\mathbf{p}\in\mathcal{P}} E\left[\log\left(1 + \gamma \sum_{i=1}^{N} p_i \mu_i w_i\right)\right] \tag{16.44}$$

where $\mathcal{P} \triangleq \{\mathbf{p} : \mathbf{p} \geq \mathbf{0},\ \mathbf{1}^T\mathbf{p} = 1\}$ is the power constraint set. The channel capacities of the three types of CSIT all depends on the eigenvalues of $\mathbf{R}_h$, i.e., μ, which is exactly characterized by the following result.

Theorem 16.2.4. *([29]) While $C_{\text{noCSIT}}(\mu)$ and $C_{\text{pCSIT}}(\mu)$ are both Schur-concave in μ, $C_{\text{cfCSIT}}(\mu)$ is Schur-convex in μ.*

Proof: The Schur-concavity of $C_{\text{noCSIT}}(\mu)$ and $C_{\text{pCSIT}}(\mu)$ follows readily from Corollary 16.1.6, since $f(\mathbf{x}) = \log(1 + a\sum_{i=1}^{N} x_i)$ is a symmetric and concave function for $a > 0$ and $x_i \geq 0$. The proof of the Schur-convexity of $C_{\text{cfCSIT}}(\mu)$ is based on Theorem 16.1.2 but quite involved. We refer the interested reader to [29] for more details. □

Theorem 16.2.4 completely characterizes the impact of correlation on the ergodic capacity of a MISO channel. To see this, assume that $\text{Tr}(\mathbf{R}_h) = N$ (for a fair comparison under Definition 16.2.2) and the correlation vector μ^2 majorizes μ^1, i.e., $\mu^1 \prec \mu^2$. We define the fully correlated vector $\psi = (N, 0, \ldots, 0)$ that majorizes all other vectors, and the least correlated vector $\chi = (1, 1, \ldots, 1)$ that is majorized by all other vectors. Then, according to Theorem 16.2.4, the impact of different types of CSIT and different levels of correlation on the MISO capacity is provided in the following inequality chain [29]:

$$\begin{array}{ccccccccc} & C_{\text{noCSIT}}(\psi) & \leq & C_{\text{noCSIT}}(\mu^2) & \leq & C_{\text{noCSIT}}(\mu^1) & \leq & C_{\text{noCSIT}}(\chi) \\ = & C_{\text{cfCSIT}}(\chi) & \leq & C_{\text{cfCSIT}}(\mu^1) & \leq & C_{\text{cfCSIT}}(\mu^2) & \leq & C_{\text{cfCSIT}}(\psi) \\ = & C_{\text{pCSIT}}(\psi) & \leq & C_{\text{pCSIT}}(\mu^2) & \leq & C_{\text{pCSIT}}(\mu^1) & \leq & C_{\text{pCSIT}}(\chi). \end{array} \tag{16.45}$$

Simply speaking, correlation helps in the covariance feedback case, but degrades the channel capacity when there is either perfect or no CSIT. Nevertheless, the more amount of CSIT is available, the better the performance could be.

16.2.5 Robust Design

The performance of MIMO communication systems depends, to a substantial extent, on the channel state information (CSI) available at both ends of the communication link. While CSI at the receiver (CSIR) is usually assumed to be perfect, CSI at the transmitter (CSIT) is often imperfect due to many practical issues. Therefore, when devising MIMO transmit strategies, the imperfectness of CSIT has to be considered, leading to the so-called robust designs. A common philosophy of robust designs is to achieve worst-case robustness, i.e., to guarantee the system performance in the worst channel [30]. In this section, we use majorization theory to prove that the uniform power allocation is the worst-case robust solution for two kinds of imperfect CSIT.

Deterministic imperfect CSIT

Consider the MIMO channel model in (16.15), where the transmit strategy is given by the transmit covariance matrix $\mathbf{Q}$. Indeed, assuming the transmit signal $\mathbf{x}$ is a Gaussian random vector with zero mean and covariance matrix $\mathbf{Q}$, i.e., $\mathbf{x} \sim \mathcal{CN}(\mathbf{0}, \mathbf{Q})$, the mutual information is given by [16]

$$\Psi(\mathbf{Q}, \mathbf{H}) = \log\det\left(\mathbf{I} + \mathbf{H}\mathbf{Q}\mathbf{H}^H\right) = \log\det\left(\mathbf{I} + \mathbf{Q}\mathbf{H}^H\mathbf{H}\right). \tag{16.46}$$

If $\mathbf{H}$ is perfectly known by the transmitter, i.e., perfect CSIT, the channel capacity can be achieved by maximizing $\Psi(\mathbf{Q}, \mathbf{H})$ under the power constraint $\mathbf{Q} \in \mathcal{Q} \triangleq \{\mathbf{Q} : \mathbf{Q} \succeq \mathbf{0},\ \text{Tr}(\mathbf{Q}) = P\}$.

In practice, however, the accurate channel value is usually not available, but belongs to a known set of possible values, often called an uncertainty region. Since $\Psi(\mathbf{Q}, \mathbf{H})$ depends on $\mathbf{H}$ through $\mathbf{R}_H = \mathbf{H}^H\mathbf{H}$, we can conveniently define an uncertainty region $\mathcal{H}$ as

$$\mathcal{H} \triangleq \{\mathbf{H} : \mathbf{R}_H \in \mathcal{R}_H\} \tag{16.47}$$

where the set $\mathcal{R}_H$ could, for example, contain any kind of spectral (eigenvalue) constraints as

$$\mathcal{R}_H \triangleq \{\mathbf{R}_H : \{\lambda_i(\mathbf{R}_H)\} \in \mathcal{L}_{R_H}\} \tag{16.48}$$

where $\mathcal{L}_{R_H}$ denotes arbitrary eigenvalue constraints. Note that $\mathcal{H}$ defined in (16.47) and (16.48) is an isotropic set in the sense that for each $\mathbf{H} \in \mathcal{H}$ we have $\mathbf{HU} \in \mathcal{H}$ for any unitary matrix $\mathbf{U}$.

Following the philosophy of worst-case robustness, the robust transmit strategy is obtained by optimizing $\Psi(\mathbf{Q}, \mathbf{H})$ in the worst channel within the uncertainty region $\mathcal{H}$, thus resulting in a maximin problem

$$\max_{\mathbf{Q}\in\mathcal{Q}} \min_{\mathbf{H}\in\mathcal{H}} \Psi(\mathbf{Q}, \mathbf{H}). \tag{16.49}$$

The optimal value of this maximin problem is referred to as the compound capacity [31]. In the following, we show that the compound capacity is achieved by the uniform power allocation.

Theorem 16.2.5. *([32, Theorem 1]) The optimal solution to (16.49) is* $\mathbf{Q}^\star = \frac{P}{N}\mathbf{I}$ *and the optimal value is*

$$C(\mathcal{H}) = \min_{\mathbf{H}\in\mathcal{H}} \log\det\left(\mathbf{I} + \frac{P}{N}\mathbf{H}^H\mathbf{H}\right).$$

Proof: Denote the eigenvalues of $\mathbf{Q}$ by $p_1 \geq \cdots \geq p_N$ w.l.o.g. in decreasing order. From [32, Lemma 1], the optimal $\mathbf{Q}$ depends only on $\{p_i\}$ and thus the inner minimization of (16.49) is equivalent to

$$\underset{\{\lambda_i(\mathbf{R}_H)\}\in\mathcal{L}_{R_H}}{\text{minimize}} \sum_{i=1}^{N} \log\left(1 + p_i\lambda_i(\mathbf{R}_H)\right)$$

with $\lambda_1(\mathbf{R}_H) \leq \cdots \leq \lambda_N(\mathbf{R}_H)$ in increasing order. Consider the function $f(\mathbf{x}) = \sum_{i=1}^N g_i(x_i) = \sum_{i=1}^N \log(1 + a_i x_i)$ with $\{a_i\}$ in increasing order. It is easy to verify that $g_i'(x) \leq g_{i+1}'(y)$ whenever $x \geq y$. Thus, from Theorem 16.1.1, $f(\mathbf{x})$ is a Schur-concave function, whose maximum is achieved by a uniform vector $\mathbf{x}$. Under the power constraint $\sum_{i=1}^N p_i = P$, it follows that

$$\min_{\{\lambda_i(\mathbf{R}_H)\}\in\mathcal{L}_{R_H}} \sum_{i=1}^{N} \log\left(1 + p_i\lambda_i(\mathbf{R}_H)\right) \leq \min_{\{\lambda_i(\mathbf{R}_H)\}\in\mathcal{L}_{R_H}} \sum_{i=1}^{N} \log\left(1 + \frac{P}{N}\lambda_i(\mathbf{R}_H)\right)$$

where the equality holds for the uniform power allocation. □

The optimality of the uniform power allocation is actually not very surprising. Due to the symmetry of the problem, if the transmitter does not uniformly distribute power over the eigenvalues of $\mathbf{Q}$, then the worst channel will align its highest singular value (or eigenvalue of $\mathbf{R}_H$) to the lowest eigenvalue of $\mathbf{Q}$. Therefore, to avoid such a situation and achieve the best performance in the worst channel, an appropriate way is to use equal power on all eigenvalues of $\mathbf{Q}$, which is formally proved in Theorem 16.2.5.

Stochastic imperfect CSIT

Tracking the instantaneous channel value may be difficult when the channel varies rapidly. The stochastic imperfect CSIT model assumes that the channel is a random quantity with its statistics such as mean or/and covariance known by the transmitter. Sometimes, even the channel statistics may not be perfectly known. The interests of this model would be on optimizing the average system performance using the channel statistics.

For simplicity, we consider the MISO channel in (16.40), where the channel $\mathbf{h}$ is a circular symmetric Gaussian random vector with zero-mean and covariance matrix $\mathbf{R}_h$, i.e., $\mathbf{h} \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}_h)$. Mathematically, the channel can be expressed as

$$\mathbf{h} = \mathbf{R}_h^{1/2}\mathbf{z} \tag{16.50}$$

where $\mathbf{z} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$. Different from the covariance feedback case where $\mathbf{R}_h$ is assumed to be known by the transmitter (see Section 16.2.4), here we consider an extreme case where the transmitter does not even know exactly $\mathbf{R}_h$. Instead, we assume that $\mathbf{R}_h \in \mathcal{R}_h$ with

$$\mathcal{R}_h \triangleq \{\mathbf{R}_h : \{\lambda_i(\mathbf{R}_h)\} \in \mathcal{L}_{R_h}\} \tag{16.51}$$

where $\mathcal{L}_{R_h}$ denotes arbitrary constraints on the eigenvalues of $\mathbf{R}_h$. In the case of no information on $\mathbf{R}_h$, we have $\mathcal{L}_{R_h} = \mathbb{R}_+^N$. To combat with the possible bad channels, the robust transmit strategy should take into account the worst channel covariance, thus leading to the following maximin problem

$$\max_{\mathbf{Q}\in\mathcal{Q}} \min_{\mathbf{R}_h\in\mathcal{R}_h} E\left[\log(1+\mathbf{h}^H\mathbf{Q}\mathbf{h})\right] = E\left[\log(1+\mathbf{z}^H\mathbf{R}_h^{1/2}\mathbf{Q}\mathbf{R}_h^{1/2}\mathbf{z})\right] \tag{16.52}$$

where $\mathcal{Q} \triangleq \{\mathbf{Q} : \mathbf{Q} \succeq \mathbf{0},\ \mathrm{Tr}(\mathbf{Q}) = P\}$ and $\mathbf{z} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$. The following result indicates that the uniform power allocation is again the robust solution.

Theorem 16.2.6. *The optimal solution to (16.52) is* $\mathbf{Q}^\star = \frac{P}{N}\mathbf{I}$ *and the optimal value is*

$$C(\mathcal{R}_h) = \min_{\mathbf{R}_h\in\mathcal{R}_h} E\left[\log\left(1+\frac{P}{N}\sum_{i=1}^N \lambda_i(\mathbf{R}_h)w_i\right)\right]$$

where $w_1, \ldots, w_N$ *are standard exponentially iid random variables.*

Proof: Denote the eigenvalues of $\mathbf{Q}$ by $p_1 \geq \cdots \geq p_N$ w.l.o.g. in decreasing order. Considering that $\mathcal{Q}$ and $\mathcal{R}_h$ impose no constraint on the eigenvectors of $\mathbf{Q}$ and $\mathbf{R}_h$, respectively, and that $\mathbf{Uz}$ has the same distribution as $\mathbf{z}$ for any unitary matrix $\mathbf{U}$, the optimal $\mathbf{Q}$ should be a diagonal matrix depending on the eigenvalues $\{p_i\}$ (see e.g. [28]) and thus (16.52) is equivalent to

$$\max_{\mathbf{Q}\in\mathcal{Q}} \min_{\mathbf{R}_h\in\mathcal{R}_h} E\left[\log\left(1+\sum_{i=1}^N p_i\lambda_i(\mathbf{R}_h)w_i\right)\right] \tag{16.53}$$

with $w_i = |z_i|^2$, where z_i is the ith element of $\mathbf{z}$.

Given $\{p_i\}$ in decreasing order, the minimum of the inner minimization of (16.53) must be achieved with $\{\lambda_i(\mathbf{R}_h)\}$ in increasing order, otherwise a smaller objective value can be obtained by changing the order of $\{\lambda_i(\mathbf{R}_h)\}$. Then, following the similar steps in the proof of Theorem 16.2.5, one can show that $E\left\{\log(1+\sum_{i=1}^N p_i\lambda_i(\mathbf{R}_h)w_i)\right\}$ is a Schur-concave function with respect to $(p_i)_{i=1}^N$. Hence, the maximum of (16.53) is achieved by a uniform power vector that is majorized by all other power vectors under the constraint $\sum_{i=1}^N p_i = P$. □

Another interesting problem is to investigate the worst channel correlation for all possible transmit strategies, which is given by the solution to the following minimax problem:

$$\min_{\mathbf{R}_h\in\mathcal{R}_h} \max_{\mathbf{Q}\in\mathcal{Q}} E\left[\log(1+\mathbf{h}^H\mathbf{Q}\mathbf{h})\right]. \tag{16.54}$$

Through the similar steps in the proof of Theorem 16.2.6, one can find that the solution to (16.54) is proportional to an identity matrix, i.e., $\mathbf{R}_h = \alpha\mathbf{I}$ with $\alpha \geq 0$. This provides a robust explanation for the assumption that $\mathbf{h} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$ in the case of no CSIT (see Section 16.2.4): $\mathbf{R}_h = \mathbf{I}$ is the worst correlation among $\mathcal{R}_h = \{\mathbf{R}_h : \mathrm{Tr}(\mathbf{R}_h) = N\}$ [29].

16.3 Conclusions and Further Readings

This chapter introduced majorization as a partial order relationship for real-valued vectors and described its main properties. This chapter also presented applications of majorization theory in proving inequalities and solving various optimization problems in the fields of signal processing and wireless communications. For a more comprehensive treatment of majorization theory and its applications, the readers are directed to Marshall and Olkins book [1]. Applications of majorization theory to signal processing and wireless communications are also described in the tutorials [6] and [7] .

16.4 Exercises

Exercise 16.4.1. Schur-convexity of sums of functions.

a. Let $\phi(\mathbf{x}) = \sum_{i=1}^n g_i(x_i)$, where each g_i is differentiable. Show that ϕ is Schur-convex on $\mathcal{D}_n$ if and only if

$$g_i'(a) \geq g_{i+1}'(b) \quad \text{whenever} \quad a \geq b, \quad i = 1, \ldots, n-1.$$

b. Let $\phi(\mathbf{x}) = \sum_{i=1}^{n} a_i g(x_i)$, where $g(x)$ is decreasing and convex, and $0 \leq a_1 \leq \cdots \leq a_n$. Show that ϕ is Schur-convex on $\mathcal{D}_n$.

Exercise 16.4.2. Schur-convexity of products of functions.

a. Let $g : \mathcal{I} \to \mathbb{R}_+$ be continuous on the interval $\mathcal{I} \subseteq \mathbb{R}$. Show that $\phi(\mathbf{x}) = \prod_{i=1}^{n} g(x_i)$ is (strictly) Schur-convex on $\mathcal{I}^n$ if and only if $\log g$ is (strictly) convex on $\mathcal{I}$.

b. Show that $\phi(\mathbf{x}) = \prod_{i=1}^{n} \Gamma(x_i)$, where $\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du$ denotes the Gamma function, is strictly Schur-convex on $\mathbb{R}_{++}^n$.

Exercise 16.4.3. Linear MIMO Transceiver.

a. Prove Corollary 16.2.1, which shows that when the cost function f_0 is either Schur-concave or Schur-convex, the optimal linear MIMO transceiver admits an analytical structure.

b. Show that the following problem formulations can be rewritten as minimizing a Schur-concave cost function of MSEs:

- Minimizing $f(\{\mathrm{MSE}_i\}) = \sum_{i=1}^{L} \alpha_i \mathrm{MSE}_i$.[16]
- Minimizing $f(\{\mathrm{MSE}_i\}) = \prod_{i=1}^{L} \mathrm{MSE}_i^{\alpha_i}$.
- Maximizing $f(\{\mathrm{SINR}_i\}) = \sum_{i=1}^{L} \alpha_i \mathrm{SINR}_i$.
- Maximizing $f(\{\mathrm{SINR}_i\}) = \prod_{i=1}^{L} \mathrm{SINR}_i^{\alpha_i}$.
- Minimizing $f(\{\mathrm{BER}_i\}) = \prod_{i=1}^{L} \mathrm{BER}_i$.

c. Show that the following problem formulations can be rewritten as minimizing a Schur-convex cost function of MSEs:

- Minimizing $f(\{\mathrm{MSE}_i\}) = \max_i\{\mathrm{MSE}_i\}$.
- Maximizing $f(\{\mathrm{SINR}_i\}) = \left(\prod_{i=1}^{L} \mathrm{SINR}_i^{-1}\right)^{-1}$.
- Maximizing $f(\{\mathrm{SINR}_i\}) = \min_i\{\mathrm{SINR}_i\}$.
- Minimizing $f(\{\mathrm{BER}_i\}) = \sum_{i=1}^{L} \mathrm{BER}_i$.
- Minimizing $f(\{\mathrm{BER}_i\}) = \max_i\{\mathrm{BER}_i\}$.

Exercise 16.4.4. Nonlinear MIMO Transceiver.

a. Prove Corollary 16.2.2, which shows that the optimal nonlinear DF MIMO transceiver can also be analytically characterized if the composite cost function $f_0 \circ \exp$ is either Schur-concave or Schur-convex.

b. Show that the following problem formulations can be rewritten as minimizing a Schur-concave $f_0 \circ \exp$ of MSEs:

- Minimizing $f(\{\mathrm{MSE}_i\}) = \prod_{i=1}^{L} \mathrm{MSE}_i^{\alpha_i}$.
- Maximizing $f(\{\mathrm{SINR}_i\}) = \sum_{i=1}^{L} \alpha_i \mathrm{SINR}_i$.

c. Show that, in addition to all problem formulations in Exercise 16.4.3.c, the following ones can also be rewritten as minimizing a Schur-convex $f_0 \circ \exp$ of MSEs:

- Minimizing $f(\{\mathrm{MSE}_i\}) = \sum_{i=1}^{L} \mathrm{MSE}_i$.
- Minimizing $f(\{\mathrm{MSE}_i\}) = \prod_{i=1}^{L} \mathrm{MSE}_i$.
- Maximizing $f(\{\mathrm{SINR}_i\}) = \prod_{i=1}^{L} \mathrm{SINR}_i$.

[16] Assume w.l.o.g. that $0 \leq \alpha_1 \leq \cdots \leq \alpha_L$.

Bibliography

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

[2] G. H. Hardy, J. E. Littlewood, and G. Pólya, *Inequalities*, 2nd ed. London and New York: Cambridge University Press, 1952.

[3] R. Bhatia, *Matrix Analysis*. New York: Springer-Verlag, 1997.

[4] R. A. Horn and C. R. Johnson, *Matrix Analysis*. New York: Cambridge University Press, 1985.

[5] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, 3rd ed. Wiley, 2003.

[6] D. P. Palomar and Y. Jiang, "MIMO transceiver design via majorization theory," *Foundations and Trends in Communications and Information Theory*, vol. 3, no. 4-5, pp. 331–551, 2006.

[7] E. A. Jorswieck and H. Boche, "Majorization and matrix-monotone funcstions in wireless communications," *Foundations and Trends in Communications and Information Theory*, vol. 3, no. 6, pp. 553–701, Jul. 2006.

[8] P. Viswanath and V. Anantharam, "Optimal sequences and sum capacity of synchronous CDMA systems," *IEEE Trans. Inform. Theory*, vol. 45, no. 6, pp. 1984–1993, Sep. 1999.

[9] 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.

[10] C. T. Mullis and R. A. Roberts, "Synthesis of minimum roundoff noise fixed point digital filters," *IEEE Trans. on Circuits and Systems*, vol. CAS-23, no. 9, pp. 551–562, Sept. 1976.

[11] Y. Jiang, W. Hager, and J. Li, "The generalized triangular decomposition," *Mathematics of Computation*, Nov. 2006.

[12] E. A. Jorswieck and H. Boche, "Outage probability in multiple antenna systems," *European Transactions on Telecommunications*, vol. 18, no. 3, pp. 217–233, Apr. 2007.

[13] S. Verdú, *Multiuser Detection*. New York, NY: Cambridge University Press, 1998.

[14] S. Ulukus and R. D. Yates, "Iterative construction of optimum signature sequence sets in synchronous CDMA systems," *IEEE Trans. Inform. Theory*, vol. 47, no. 5, pp. 1989–1998, Jul. 2001.

[15] C. Rose, S. Ulukus, and R. D. Yates, "Wireless systems and interference avoidance," *IEEE Trans. Wireless Commun.*, vol. 1, no. 3, pp. 415–428, Jul. 2002.

[16] I. E. Telatar, "Capacity of multi-antenna Gaussian channels," *European Trans. Telecommun.*, vol. 10, no. 6, pp. 585–595, Nov.-Dec. 1999.

[17] D. P. Palomar, M. A. Lagunas, and J. M. Cioffi, "Optimum linear joint transmit-receive processing for MIMO channels with QoS constraints," *IEEE Trans. Signal Process.*, vol. 52, no. 5, pp. 1179–1197, May 2004.

[18] L. Zheng and D. N. C. Tse, "Diversity and multiplexing: A fundamental tradeoff in multiple-antenna channels," *IEEE Trans. Inform. Theory*, vol. 49, no. 5, pp. 1073–1096, May 2003.

[19] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1993.

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

[21] Y. Jiang, W. Hager, and J. Li, "The geometric mean decomposition," *Linear Algebra and Its Applications*, vol. 396, pp. 373–384, Feb. 2005.

[22] ——, "Tunable channel decomposition for MIMO communications using channel state information," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4405–4418, Nov. 2006.

[23] F. Xu, T. N. Davidson, J. K. 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.

[24] A. A. D'Amico, "Tomlinson-Harashima precoding in MIMO systems: A unified approach to transceiver optimization based on multiplicative schur-convexity," *IEEE Trans. Signal Process.*, vol. 56, no. 8, pp. 3662–3677, Aug. 2008.

[25] B. Hassibi, "A fast square-root implementation for BLAST," in *Proc. of Thirty-Fourth Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, USA, Nov. 2000.

[26] P. P. Vaidyanathan, S.-M. Phoong, and Y.-P. Lin, *Signal Processing and Optimization for Transceiver Systems*. New York, NY: Cambridge University Press, 2010.

[27] P. Viswanath and V. Anantharam, "Optimal sequences for CDMA under colored noise: A Schur-saddle function property," *IEEE Trans. Inform. Theory*, vol. 48, no. 6, pp. 1295–1318, Jun. 2002.

[28] S. A. Jafar and A. Goldsmith, "Transmitter optimization and optimality of beamforming for multiple antenna systems," *IEEE Trans. Wireless Commun.*, vol. 3, no. 4, pp. 1165–1175, Jul. 2004.

[29] E. A. Jorswieck and H. Boche, "Optimal transmission strategies and impact of correlation in multiantenna systems with different types of channel state information," *IEEE Trans. Signal Process.*, vol. 52, no. 12, pp. 3440–3453, Dec. 2004.

[30] J. Wang and D. P. Palomar, "Worst-case robust MIMO transmission with imperfect channel knowledge," *IEEE Trans. Signal Process.*, vol. 57, no. 8, pp. 3086–3100, Aug. 2009.

[31] A. Lapidoth and P. Narayan, "Reliable communication under channel uncertainty," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2148–2177, Oct. 1998.

[32] D. P. Palomar, J. M. Cioffi, and M. A. Lagunas, "Uniform power allocation in MIMO channels: A game-theoretic approach," *IEEE Trans. Inform. Theory*, vol. 49, no. 7, pp. 1707–1727, Jul. 2003.