Optimal Two-Dimensional Lattices for Precoding of Linear Channels

Dževdan Kapetanović, Hei Victor Cheng, Wai Ho Mow, and Fredrik Rusek

***Abstract*—Consider the communication system model $\mathbf{y} = \mathbf{HFx}+\mathbf{n}$, where $\mathbf{H}$ and $\mathbf{F}$ are the channel and precoder matrices, $\mathbf{x}$ is a vector of data symbols drawn from some lattice-type constellation, such as M-QAM, $\mathbf{n}$ is an additive white Gaussian noise vector and $\mathbf{y}$ is the received vector. It is assumed that both the transmitter and the receiver have perfect knowledge of the channel matrix $\mathbf{H}$ and that the transmitted signal $\mathbf{Fx}$ is subject to an average energy constraint. The columns of the matrix $\mathbf{HF}$ can be viewed as the basis vectors that span a lattice, and we are interested in the precoder $\mathbf{F}$ that maximizes the minimum distance of this lattice. This particular problem remains open within the theory of lattices and the communication theory. This paper provides the complete solution for any non-singular $M \times 2$ channel matrix $\mathbf{H}$. For real-valued matrices and vectors, the solution is that $\mathbf{HF}$ spans the hexagonal lattice. For complex-valued matrices and vectors, the solution is that $\mathbf{HF}$, when viewed in four-dimensional real-valued space, spans the Schläfli lattice D_4.**

***Index Terms*—**

I. Introduction

We consider a complex baseband linear transmission system with 2 inputs and ≥ 2 outputs corrupted by additive white Gaussian noise. The mathematical model of the system under investigation is

$$\mathbf{y} = \mathbf{HFx} + \mathbf{n} \tag{1}$$

where $\mathbf{H}$ represents an $M \times 2$ channel gain matrix, $\mathbf{F}$ is a 2×2 precoding matrix adopted at the transmitter side, $\mathbf{x}$ is a 2-dimensional column vector comprising data symbols, and $\mathbf{n}$ is a vector of independent white Gaussian noise variables. Throughout the paper, we assume that both the transmitter and the receiver are provided with perfect channel state information, and we consider the problem of constructing the precoding matrix $\mathbf{F}$ to improve system performance according to a properly chosen metric. As pointed out in [7], the model in (1) also encompasses the case when transmitting two data streams over an $M \times N_t$ MIMO channel with $M, N_t \geq 2$. Telatar showed in [1] that the capacity-achieving approach is to access the eigenmodes of the channel matrix through a singular value decomposition (SVD) and transmit independent complex Gaussian symbols over each eigenmode with appropriate power allocation. The optimal power allocation can be obtained by applying the classical waterfilling technique. Note that power allocation can be viewed as a precoder of the specific form $\mathbf{F} = \mathbf{VP}$, where the SVD of $\mathbf{H}$ is $\mathbf{H} = \mathbf{USV}^*$, with $(\cdot)^*$ denoting the Hermitian transpose, and the diagonal matrix $\mathbf{P}$ has its diagonal consisting of the power allocation factors.

However, Gaussian data symbols are impractical, and usually discrete QAM-like constellations are used. In this case, the power allocation as a result of the classical waterfilling is no longer optimal in the sense of maximizing the mutual information. With discrete constellations, the best power allocation strategy is known under the name Mercury/Waterfilling and is due to Lozano, Tulino, and Verdu [2]. However, unlike the situation with Gaussian inputs, the best power allocation does not result in the optimal precoder $\mathbf{F}$ for an arbitrarily given discrete constellation. It merely gives the optimal precoder under the constraint that the data symbols in $\mathbf{x}$ can be independently detected without performance loss at the receiver side [2].

To find the precoder $\mathbf{F}$ that maximizes the mutual information between the input and output of the channel, i.e., solve

$$\mathbf{F}_{\text{opt}} = \arg\max_{\mathbf{F}} I(\mathbf{y}; \mathbf{x}), \tag{2}$$

where $I(\cdot;\cdot)$ is the mutual information operator, is a challenging problem. Recently, Perez-Cruz, Rodrigues and Verdu [3] derived the necessary conditions for the first-order optimality, and also proposed a fixed-point iterative algorithm to find efficient precoders $\mathbf{F}$. Continuing in this direction, it was shown in [4] that the information rate is concave with respect to a certain precoder-dependent matrix. This allows the development of more effective numerical algorithms for finding the optimal mutual information precoder. However, it still involves heavy computations, such as computing the non-linear MMSE matrix. It is known that the optimal mutual information precoder converges to the Mercury/Waterfilling policy [3] at low SNR, while it converges to the precoder that maximizes the minimum Euclidean distance[1] between all

Manuscript received March 27, 2012; revised September 6, 2012 and January 29, 2013; accepted March 14, 2013. The associate editor coordinating the review of this paper and approving it for publication was A. Chockalingam.

D. Kapetanović is with the Interdisciplinary Center for Security, Reliability and Trust (SnT), University of Luxembourg, Luxembourg (e-mail: dzevdan.kapetanovic@uni.lu).

H. V. Cheng is with the Department of Electrical Engineering (ISY), Linköping University, 581 83, Linköping, Sweden (e-mail: hei.cheng@liu.se).

W. H. Mow is with the Department of Computer and Electronic Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong (e-mail: eewhmow@ust.hk).

F. Rusek is with the Department of Electrical and Information Technology, Lund University, P. O. Box 118, 22100 Lund, Sweden (e-mail: fredrik.rusek@eit.lth.se).

The work of the first and the fourth author were supported by the Swedish Foundation for Strategic Research through its Center for High Speed Wireless Communication at Lund University, Sweden. The work of the second and third authors were supported by the AoE grant E-02/08 from the University Grants Committee of the Hong Kong Special Administration Region, China. This paper was presented in parts at the International Symposium on Information Theory (ISIT), St. Petersburg, 2011.

Digital Object Identifier 10.1109/TWC.2013.13.120452

[1]It will be simply referred to as the minimum distance in the sequel.

possible received constellation points (i.e. noiseless received vectors) $\mathbf{HFx}$'s at high SNR. This observation and further bounds that link the minimum distance to mutual information were also pointed out by Palomar and Payaro in [5]. The problem of maximizing the minimum distance can be formally stated as

$$\mathbf{F}_{\text{opt}} = \arg\max_{\mathbf{F}} \min_{\mathbf{x},\mathbf{x}'\neq\mathbf{x}} \|\mathbf{HF}(\mathbf{x}-\mathbf{x}^*)\|^2 \text{ such that } \|\mathbf{F}\|^2 \leq P_0, \tag{3}$$

where P_0 represents the allowable maximum average power and is an arbitrary positive constant that is necessary to make the optimization problem meaningful, and $\|.\|^2$ denotes the sum of square magnitudes of all elements in the vector/matrix argument. The problem (3) was shown to be NP-hard in [5].

In general, the problem of constructing precoders, not limited to a MIMO context, such that the minimum distance is maximized is a classical problem in communication theory and has received much attention during the past decades, see for example [6] for an overview. In the field of wireless communications, $\mathbf{H}$ typically represents a MIMO or an OFDM channel, and there has been significant progress towards determining the optimal (or near-optimal) minimum distance constructions of $\mathbf{F}$. In [7], the problem is completely solved for $M \times 2$ channels with BPSK or QPSK constellations. An extension to 16-QAM was made in [8]. Suboptimal designs based on Toeplitz matrices for any dimensions of the channel matrices and constellations were given in [9]. The work in [10] proposes real-valued precoders with low ML-decoding complexity. [11] considers real-valued precoders that are approximately optimal among real-valued ones with respect to the minimum distance for $M \times 2$ systems, and are easy to calculate for large QAM constellations. In [12] and [13], it is proposed to design $\mathbf{F}$ based on dense lattice packings. A lattice-based construction implicitly assumes that the signal constellation is a finite but sufficiently "large" set of lattice points, and the idea is that if the received constellation points $\mathbf{HFx}$'s are arranged as a dense lattice packing, the minimum distance is expected to be "good". However, no exact results on optimality have been presented in either of these papers.

To gain some insight into the problem, let us examine some simple special cases. In real-valued precoding, some specific instances of the problem in (3) can be viewed geometrically. Assume that $\text{tr}(\mathbf{FF}^*) = 4$, with $\text{tr}(\cdot)$ being the trace operator, and the elements of the input $\mathbf{x}$ are identically and independently distributed (i.i.d.) random variables. Consider the special case of a diagonal channel matrix $\mathbf{H}$, i.e., $h_{1,2} = h_{2,1} = 0$, and further normalize $\mathbf{H}$ to have $h_{2,2} = 1$ so that we only need to focus on the effect of varying the value of $h_{1,1}$. (This is less restrictive than it appears because a diagonal matrix can be used to represent a general channel matrix in the SVD representation. Refer to the definition of $\mathbf{S}$ in Section II-B for details.) Since there are only four real-valued elements in $\mathbf{F}$, and they are bounded by the energy constraint, it is possible to determine the optimal $\mathbf{F}$ to (3) for some carefully chosen value of $h_{1,1}$, say, by empirical means. When $\mathbf{H} = \mathbf{I}$ (i.e., $h_{1,1} = 1$), one optimal solution to (3) is $\mathbf{F} = \mathbf{I}$, while another one is

$$\mathbf{F} = \begin{pmatrix} 1 & 0.5 \\ 0 & \sqrt{3/4} \end{pmatrix},$$

which spans a hexagonal lattice. However, as soon as $\mathbf{H}$ deviates from $\mathbf{I}$ (even with a very small change, say, $h_{1,1} = 1.01$), the optimal $\mathbf{F}$ is unique (up to sign changes in the columns) and it gives rise to an $\mathbf{HF}$ that spans a hexagonal lattice. Varying $h_{1,1}$ further, the optimal $\mathbf{F}$ changes in a continuous way, while the received lattice $\mathbf{HF}$ remains the same (up to scaling). This behavior continues until $h_{1,1}$ reaches a certain value, for which the optimal $\mathbf{F}$ suddenly changes in a discontinuous way, resulting in a discontinuous change in $\mathbf{HF}$. However, surprisingly, $\mathbf{HF}$ still spans a hexagonal lattice, in spite of its subtle changes!

Figure 1 depicts such a behavior by plotting as vectors the columns of the optimal $\mathbf{F}$ and the corresponding $\mathbf{HF}$ for three different $\mathbf{H}$'s with $h_{1,1} = 1.5, 2.7$ and 2.8, respectively. Assuming the constellation points are integers, the received constellation points $\mathbf{HFx}$ are shown as discrete points. The optimal $\mathbf{F}$ changes continuously as $h_{1,1}$ increases from 1.5 to 2.7, and the columns of $\mathbf{HF}$ are simply being scaled and always span the same hexagonal lattice (up to scaling). When $h_{1,1}$ further increases from 2.7 to 2.8, there is a discontinuous change in the elements of the optimal $\mathbf{F}$. The columns of $\mathbf{HF}$ also change discontinuously, but they still span the hexagonal lattice (up to scaling and rotation). This intriguing behavior of the optimal precoder poses a challenging puzzle, and this paper is a consequence of our effort to offer a satisfactory solution to settle this puzzle.

The main contribution of this paper is to derive the exact solution to the minimum distance precoding problem for the case of an $M \times 2$ channel matrix $\mathbf{H}$ and an infinite signal constellation. Although our results are derived for infinite constellations, the results are applicable to "large" QAM constellations. In our numerical result section, we shall investigate how "large" a QAM constellation is sufficient for the presented results to be fruitfully applied. With the solution at hand, we are able to answer questions, such as the following.

- Is there a general underlying structure of the precoding optimization problem (3)?
- Under what conditions, does the solution to (3) vary with the channel matrix $\mathbf{H}$ in a continuous (respectively, discrete) manner?
- Is it possible to offline construct a codebook of optimal precoders so that there is no need to perform any online optimization?

The answers to these questions are that there is indeed a profound structure in the solution of (3). Remarkably, there is a single precoder structure which is optimal, and it organizes the received constellation points as a hexagonal lattice for real-valued $\mathbf{F}$'s, and as a Schläfli lattice for complex-valued $\mathbf{F}$'s. However, the basis through which the lattice $\mathbf{HF}$ is observed changes (up to scaling) in a discrete fashion when $\mathbf{H}$ changes. This implies that (3) is actually a discrete optimization problem and not a continuous one.

As a remark, we mention that the above precoder construction is optimized under the assumption that a maximum-likelihood detector is used. Much other work on precoders that use less complex receivers exist, see for example [14] for a comprehensive treatment of precoding with MMSE or DFE detectors.

The rest of this paper is organized as follows. Section II

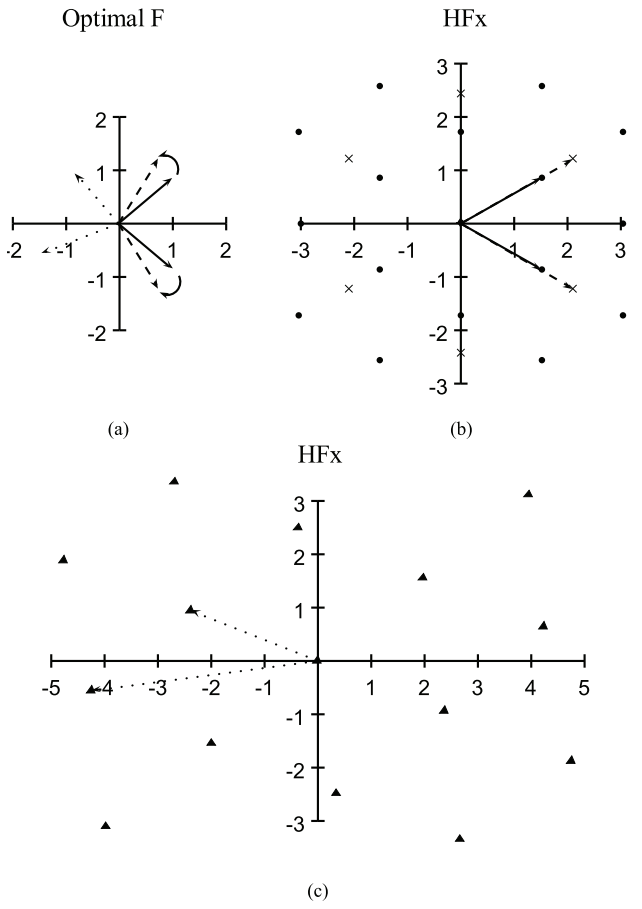


Fig. 1. Visualization of the solution to the precoding optimization problem in (3) in the special case that elements of the input $\mathbf{x}$ are i.i.d., $\mathrm{tr}(\mathbf{F}\mathbf{F}^*) = 4$, and a diagonal channel matrix $\mathbf{H} = \mathrm{diag}([h_{1,1}\ 1])$. (a) Columns of the optimal real-valued precoding matrix $\mathbf{F}$ are plotted. Three different $\mathbf{H}$ are considered: $h_{1,1} = 1.5$ (solid line arrow); $h_{1,1} = 2.7$ (dashed line arrow); $h_{1,1} = 2.8$ (dotted line arrow). Columns of the same matrix are plotted as arrows with the same line style. (b) Columns of the matrices $\mathbf{HF}$ and their corresponding received constellation points $\mathbf{HFx}$'s for $h_{1,1} = 1.5$ (solid line arrows, filled circles) and for $h_{1,1} = 2.7$ (dashed line arrows, crosses) are plotted. (c) Columns of the matrices $\mathbf{HF}$ and their corresponding received constellation points $\mathbf{HFx}$'s for $h_{1,1} = 2.8$ (dotted line arrows, filled triangles) are plotted.

presents the necessary theoretical background and formulates the problem. Section III presents the main results of the work, both for real-valued and complex-valued precoding, along with a description on how to find the optimal minimum distance precoder. All proofs are defered to the Appendices. Section V applies our derived results to MIMO and OFDM communications. Conclusions are presented in Section VI.

II. Problem Statement

This section formally formulates the problem outlined in Section I. We start by briefly introducing lattice theoretic concepts that we shall subsequently make use of. Throughout the paper, let $\mathbb{E}$ denote the expectation operator, $\mathbf{I}$ the 2×2 identity matrix, $\mathbf{0}$ the 2×2 zero matrix and $(\cdot)^{\mathrm{T}}$ matrix transpose.

A. Lattices

Let $\mathbf{L}$ be a 2×2 matrix and $\mathbf{u} = [u_1\, u_2]^{\mathrm{T}}$ a 2×1 vector. A lattice Λ_L is the set of points

$$\Lambda_{\mathbf{L}} = \{\mathbf{L}\mathbf{u} \mid u_1, u_2 \in \mathbb{Z}[i]\}, \tag{4}$$

where $\mathbb{Z}[i]$ is the set of Gaussian integers. $\mathbf{L}$ is called a *generator matrix* for the lattice $\Lambda_{\mathbf{L}}$. The *minimum distance* of $\Lambda_{\mathbf{L}}$ is defined as:

$$d^2_{\min}(\Lambda_{\mathbf{L}}) = \min_{\mathbf{u}\neq\mathbf{v}} \|\mathbf{L}(\mathbf{u}-\mathbf{v})\|^2 = \min_{\mathbf{e}\neq[0\,0]^{\mathrm{T}}} \|\mathbf{L}\mathbf{e}\|^2,$$

where $\mathbf{u}, \mathbf{v}$ and $\mathbf{e} = \mathbf{u} - \mathbf{v}$ are Gaussian integer vectors.

As can be seen from the definition of $\Lambda_{\mathbf{L}}$, the column vectors $\mathbf{l}_1$ and $\mathbf{l}_2$ form a basis for the lattice. There are infinitely many different bases in a lattice, and they all span the same lattice $\Lambda_{\mathbf{L}}$. Assume that $\mathbf{L}'$ is another basis for $\Lambda_{\mathbf{L}}$. It holds that $\mathbf{L}' = \mathbf{L}\mathbf{Z}$, where $\mathbf{Z}$ is a unimodular matrix, i.e., $\mathbf{Z}$ has Gaussian integer entries and $\det(\mathbf{Z}) \in \{\pm 1, \pm i\}$ [15].

From the definition of $d^2_{\min}(\Lambda_{\mathbf{L}})$, it follows that

$$d^2_{\min}(\Lambda_{\mathbf{QLZ}}) = d^2_{\min}(\Lambda_{\mathbf{L}}) \tag{5}$$

where $\mathbf{Q}$ is a unitary matrix.

Note that the above introduced concepts transfer naturally to real-valued matrices and vectors. From the isomorphism between a 2×2 complex-valued matrix $\mathbf{A}$ and its real-valued 4×4 counterpart $\mathbf{A}_r$

$$\mathbf{A}_r = \begin{bmatrix} \mathcal{R}\{\mathbf{A}\} & \mathcal{I}\{\mathbf{A}\} \\ -\mathcal{I}\{\mathbf{A}\} & \mathcal{R}\{\mathbf{A}\} \end{bmatrix}, \tag{6}$$

where $\mathcal{R}, \mathcal{I}$ denote the real and imaginary part of a matrix, respectively, it follows that 2-dimensional complex-valued lattices can be expressed as 4-dimensional real-valued lattices. This transformation will be used later on to convert complex-valued lattices into real-valued ones, since well known lattices are presented in their real-valued forms in the literature.

B. Problem Formulation

We consider model (1) under a bounded energy constraint of the transmitted signal $\mathbf{Fx}$. The average energy of $\mathbf{Fx}$ is

$$\mathbb{E}\{\mathbf{x}^*\mathbf{F}^*\mathbf{F}\mathbf{x}\} = \mathbb{E}\{\mathrm{tr}(\mathbf{x}^*\mathbf{F}^*\mathbf{F}\mathbf{x})\} = \mathrm{tr}(\mathbf{F}^*\mathbf{F}\mathbb{E}\{\mathbf{x}\mathbf{x}^*\}).$$

Let $\mathbf{A} = \mathbb{E}\{\mathbf{x}\mathbf{x}^*\}$ be a positive definite covariance matrix of $\mathbf{x}$. The optimization problem studied in this paper is

$$\max_{\mathbf{F}} d^2_{\min}(\Lambda_{\mathbf{HF}}) \text{ subject to } \mathrm{tr}(\mathbf{F}^*\mathbf{F}\mathbf{A}) \leq P_0. \tag{7}$$

It is tempting to connect (7) with the problem of sphere packing, which is perhaps the most classic problem within lattice theory. These two problems are not equivalent, which can be shown in a straightforward, but lengthy, fashion.

Consider next the following optimization

$$\min_{\mathbf{F}} \mathrm{tr}(\mathbf{F}^*\mathbf{F}\mathbf{A}) \text{ subject to } d^2_{\min}(\Lambda_{\mathbf{HF}}) \geq d, \tag{8}$$

for some arbitrary $d > 0$. Since both the constraint function and the objective function are homogeneous of degree two, it holds that the solution $\mathbf{F}_{\mathrm{opt}}$ to (8) is such that $\mathbf{F}_{\mathrm{opt}}\sqrt{P_0/\mathrm{tr}(\mathbf{F}^*_{\mathrm{opt}}\mathbf{F}_{\mathrm{opt}}\mathbf{A})}$ solves (7). That is, the optimal solution to (7) can be obtained from the optimal solution to

(8) with an appropriate scaling constant. Clearly, the optimal solution to (8) satisfies equality in the constraint, and we can therefore without loss of generality assume $d^2_{\min}(\Lambda_{\mathbf{HF}}) = 1$.

We find the problem (8) to be more easily analyzed than the equivalent problem (7), and thus we focus on solving (8). Before presenting the solution, we reformulate the problem into a more analytically tractable form. Let $\mathbf{H} = \mathbf{USV}^*$ denote the SVD of $\mathbf{H}$. The $M \times M$ matrix $\mathbf{U}$ has no impact on the minimum distance and can be removed from the problem formulation. Furthermore, the 2×2 matrix $\mathbf{V}^*$ can be absorbed into the precoder $\mathbf{F}$. After removing these two matrices, it is observed that the case $M > 2$ becomes equivalent to the case $M = 2$, so that we can assume $M = 2$ in the rest of the paper. These observations leave us with the simplified model

$$\mathbf{y} = \mathbf{SFx} + \mathbf{n}. \tag{9}$$

Let $\mathbf{G} = \mathbf{SF}$ be the lattice generator matrix at the receiver. $\mathbf{G}$ can be factorized as $\mathbf{G} = \mathbf{QBZ}$, where $\mathbf{Q}$ is a unitary matrix, $\mathbf{B}$ is a 2×2 matrix and $\mathbf{Z}$ is a unimodular matrix. The lattice structure of $\mathbf{G}$ is determined by the matrix $\mathbf{B}$, while $\mathbf{Z}$ is the basis through which the lattice is represented. The matrix $\mathbf{Q}$ is merely a rotation of the lattice, but plays an important role in the optimization to follow. The reason for introducing this factorization of $\mathbf{G}$ becomes evident in Appendix A, where it is shown that it is possible to obtain an analytical expression for the minimum distance constraint in (8), expressed solely in the elements of the matrix $\mathbf{B}$. With this factorization of $\mathbf{G}$, it follows that $\mathbf{F}$ can be written as

$$\mathbf{F} = \mathbf{S}^{-1}\mathbf{G} = \mathbf{S}^{-1}\mathbf{QBZ}. \tag{10}$$

Since $\mathbf{A}$ in (8) is positive definite, it has a Cholesky decomposition $\mathbf{A} = \mathbf{TT}^*$ where $\mathbf{T}$ is a lower-triangular matrix with non-negative diagonal elements. Inserting (10) and $\mathbf{A} = \mathbf{TT}^*$ into (8) yields

$$\begin{aligned} &\min_{\mathbf{Q},\mathbf{B},\mathbf{Z}} \operatorname{tr}(\mathbf{T}^*\mathbf{Z}^*\mathbf{B}^*\mathbf{Q}^*\mathbf{S}^{-2}\mathbf{QBZT}) \\ &\text{subject to } d^2_{\min}(\Lambda_{\mathbf{QBZ}}) = 1. \end{aligned} \tag{11}$$

For completeness, we shall separately consider two cases: (i) Real-valued precoding, where all quantities in (9)-(11) are real-valued, and (ii) Complex-valued precoding, where all quantities, except $\mathbf{S}$, are complex-valued.

In our proofs, we shall first determine the optimal $\mathbf{Q}$ when $\mathbf{BZT}$ is fixed. Once the optimal matrices $\mathbf{B}$ and $\mathbf{Z}$ are known, the optimal precoder is easily constructed. In the Appendices, the optimization over $\mathbf{Q}$ will be treated directly over the elements of $\mathbf{Q}$. The optimization of $\mathbf{B}$ and $\mathbf{Z}$ is treated separately, and we shall start with $\mathbf{B}$ in Section III, while optimization over $\mathbf{Z}$ is treated in Section IV.

III. Optimal Precoding Lattices

In this section we derive the optimal lattice $\mathbf{B}$ for the real-valued and the complex-valued cases.

For the real-valued case, our main result is:

Theorem 1: For any non-singular channel matrix $\mathbf{S}$, the optimal lattice $\mathbf{B}$ in (11) is the hexagonal lattice, i.e.,

$$\mathbf{B} = \begin{bmatrix} 1 & \frac{1}{2} \\ 0 & \frac{\sqrt{3}}{2} \end{bmatrix}.$$

Proof: See Appendix A. ∎

While the real-valued case is interesting for theoretical purposes, the complex-valued case is more important for practical MIMO or OFDM applications. Nevertheless, the real-valued result has immediate applications to precoding for mitigation of I/Q imbalance in scalar complex-valued channels.

For the complex-valued case, our main result is:

Theorem 2: For any non-singular channel matrix $\mathbf{S}$, the optimal lattice $\mathbf{B}$ in (11) is the complex representation of the Schläfli lattice, i.e.,

$$\mathbf{B} = \begin{bmatrix} 1 & \frac{\pm 1 \pm i}{2} \\ 0 & \pm\frac{1}{\sqrt{2}} \end{bmatrix}.$$

Proof: See Appendix B. ∎

By "complex representation" we mean that if the transformation (6) is performed on $\mathbf{B}$ in Theorem 2, the Schläfli lattice in four real-valued dimensions results.

To summarize, the minimum distance optimal precoder for "large" input constellations is always an instance of the hexagonal or the Schläfli lattice for real-valued and complex-valued precoding, respectively.

IV. Optimal Matrix $\mathbf{Z}$

Since $\mathbf{B}$ is now known, it remains to find the optimal basis matrix $\mathbf{Z}$ in order to solve (11). This section describes the core idea of the algorithms that find the optimal real-valued and complex-valued $\mathbf{Z}$, respectively, for the case when $\mathbf{A} = \mathbf{I}$. Extension to an arbitrary $\mathbf{A}$ is straightforward and is described briefly at the end of this section. A complete Matlab code for the algorithms can be found at *www.eit.lth.se/goto/Zalgorithm.*

Let $c \triangleq (1 + s^2)/2$ and denote by z_{ij} the row-i column-j element of the matrix $\mathbf{Z}$. By optimizing (11) over real-valued $\mathbf{Q}$ and $\mathbf{B}$, we have[2]

$$\mathbf{Z} = \arg\min_{\mathbf{Z}} \mu^r_{\pm}(\mathbf{Z}),$$

where

$$\begin{aligned} \mu^r_{\pm}(\mathbf{Z}) \triangleq\ & c[z_{11}^2+z_{12}^2+z_{21}^2+z_{22}^2 \pm (z_{11}z_{21}+z_{12}z_{22})] \\ & + (c-1)[(z_{11}^2+z_{12}^2)^2+(z_{21}^2+z_{22}^2)^2+4(z_{11}z_{21}+z_{12}z_{22})^2 \\ & -(z_{11}^2+z_{12}^2)(z_{21}^2+z_{22}^2) \\ & \pm 2(z_{11}z_{21}+z_{12}z_{22})(z_{11}^2+z_{12}^2+z_{21}^2+z_{22}^2)]^{1/2}. \end{aligned} \tag{12}$$

Similarly, in the complex-valued case, after optimizing over $\mathbf{Q}$ and $\mathbf{Z}$, we have

$$\mathbf{Z} = \arg\min_{\mathbf{Z}} \mu^c_{\pm}(\mathbf{Z}),$$

where

$$\begin{aligned} \mu^c_{\pm}(\mathbf{Z}) \triangleq\ & c(|z_{11}|^2+|z_{12}|^2+|z_{21}|^2+|z_{22}|^2 \\ & + \mathcal{R}\{(\pm 1 \pm i)(z_{11}z_{21}^*+z_{12}z_{22}^*)\}) \\ & + (c-1)[(|z_{11}|^2+|z_{12}|^2+|z_{21}|^2+|z_{22}|^2 \\ & + \mathcal{R}\{(\pm 1 \pm i)(z_{11}z_{21}^*+z_{12}z_{22}^*)\})^2 - 2]^{1/2}. \end{aligned} \tag{13}$$

The $\pm$ signs in both (12) and (13) can be absorbed into the elements of $\mathbf{Z}$, without changing the unimodularity of $\mathbf{Z}$. Define $\beta_r \triangleq z_{11}^2+z_{12}^2+z_{21}^2+z_{22}^2-(z_{11}z_{21}+z_{12}z_{22})$ and $\beta_c \triangleq$

[2]The optimization of (11) over $\mathbf{Q}$ and $\mathbf{B}$ is treated in the proofs of theorems 1 and 2 provided in the Appendices.

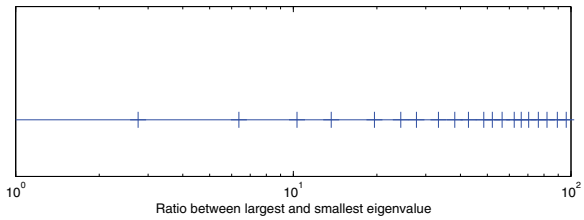


Fig. 2. Change in $\mathbf{Z}$ with respect to the ratio s_1/s_2. The solution to (11) is constant for all $\mathbf{S}$ with a ratio between any two consecutive markers. The scale on the x-axis is logarithmic.

$|z_{11}|^2+|z_{12}|^2+|z_{21}|^2+|z_{22}|^2+\mathcal{R}\{(1+i)(z_{11}z_{21}^*+z_{12}z_{22}^*)\}$, where we do not expliticly denote the dependency of β_r and β_c on $\mathbf{Z}$. Since $|\det(\mathbf{Z})|=1$, (12) and (13) become

$$\mu^r(\beta_r) = c\beta_r + (c-1)\sqrt{\beta_r^2-3} \tag{14}$$

and

$$\mu^c(\beta_c) = c\beta_c + (c-1)\sqrt{\beta_c^2-2}, \tag{15}$$

respectively. If we for the moment drop the constraint that β_r has to be integer-valued, the function $\mu^r(\mathbf{Z})$ in (14) will be minimized over β_r. It can be verified that $\mu^r(\mathbf{Z})$ is a convex function. Differentiating $\mu(\beta_r)$ with respect to β and setting the derivative to 0 gives that $\beta_{r,\text{opt}} = \sqrt{\frac{3c^2}{2c-1}}$ is the optimal point. Since $\mu^r(\mathbf{Z})$ is convex, the minimum of $\mu(\beta_r)$ over unimodular matrices can only occur at two specific matrices. Either it is the $\mathbf{Z}$ that produces the largest β_r smaller than $\beta_{r,\text{opt}}$, or it is the $\mathbf{Z}$ that produces the smallest β_r larger than $\beta_{r,\text{opt}}$. A similar analysis can be applied to the complex-valued (15), and it follows that the largest β_c smaller, or smallest β_c larger, than $\beta_{c,\text{opt}} = \sqrt{2}c/\sqrt{2c-1}$ is optimal. Hence, in the real-valued case, an algorithm can be developed that traverses unimodular $\mathbf{Z}$'s and stops when two matrices $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are found, such that $\mathbf{Z}_1$ gives the β_r that equals the largest integer smaller than $\beta_{r,\text{opt}}$, and $\mathbf{Z}_2$ gives the β_r that equals the smallest integer larger than $\beta_{r,\text{opt}}$. An algorithm for the complex-valued case works in the same way. Due to lack of space and the fact that our algorithms are ad-hoc, we omit the implementation details and refer to the above mentioned homepage where the Matlab code for both algorithms can be found. In the case when $\mathbf{T} \neq \mathbf{I}$, same conclusions as above are reached for the matrix $\mathbf{W} = \mathbf{ZT}$, where β_r and β_c are dependent on $\mathbf{W}$ instead. An algorithm that traverses unimodular $\mathbf{Z}$'s can then be formulated until the new β_r and β_c satisfy the same conditions as above. It is straightforward to include this in the code found at the above homepage.

Since we now know that solving (11) is a discrete optimization problem, it is of interest to see how often the solution changes with varying $\mathbf{S}$. Figure 2 shows the ratio s_1/s_2 on the x-axis, and the markers show the ratios where $\mathbf{Z}$ changes. As seen, the same solution can be used for a wide interval.

V. Applications

In this section we consider a number of practical applications of the optimal minimum distance lattice based precoder and make comparisons to other schemes. As discussed in Section I, minimum distance based precoders are asymptotically optimal in the high SNR regime, but minimum distance plays little role at low SNR so that we can not expect any performance gains there.

We consider first the 2×2 channel studied in [3],

$$\mathbf{S} = \begin{bmatrix} \sqrt{3} & 0 \\ 0 & 1 \end{bmatrix}. \tag{16}$$

In [3], this channel was studied at asymptotically high SNR for BPSK alphabets with real-valued precoding. The objective was to find the real-valued precoder $\mathbf{F}$ that maximizes the mutual information $I(\mathbf{SFx}+\mathbf{n};\mathbf{x})$. For high SNR, it is known that the optimal mutual information precoder converges to the optimal minimum distance precoder, and the numerical optimization framework in [3] thus produced the optimal minimum distance precoder. The precoder is of the following simple form

$$\mathbf{F} = \begin{bmatrix} \sqrt{2} & \sqrt{2} \\ -\sqrt{2} & \sqrt{2} \end{bmatrix}. \tag{17}$$

It can be verified by standard techniques that the combined channel-precoder matrix $\mathbf{SF}$ is an instance of the hexagonal lattice - which is precisly the result if an infinite lattice constellation was used. For such a lattice constellation, the strength of our analysis is that no numerical optimization of the precoder is necessary since it is known a-priori that the hexagonal lattice *must* be the solution, and it only remains to find the optimal basis matrix $\mathbf{Z}$ according to the algorithm mentioned in Section IV. By doing so, we find that the optimal $\mathbf{Z}$ for asymptotically large constellations coincides with the basis matrix that is built into (17). Altogether, for the particular channel (16) studied in [3], a "large" constellation means BPSK and it is known beforehand what structure the solution must have.

In Figure 3 we continue to study the channel (16) but we now evaluate the mutual information achieved by 4QAM inputs when the complex-valued minimum distance optimal precoder for large constellations is used. As comparisons, we also plot the achieved mutual information by 1) no precoding at all, i.e., $\mathbf{F}=\mathbf{I}$, 2) Mercury/Waterfilling from [2], and 3) capacity achieved by Gaussian inputs and waterfilling. The performance of the optimal mutual information precoder coincides with that of Mercury/Waterfilling in the low SNR regime, while it coincides with that of the minimum distance precoder in the high SNR regime. As can be seen, there is a 2 dB gain offered by the minimum distance precoder over uncoded systems and Mercury/Waterfilling at high SNR. At low SNR, the Mercury/Waterfilling policy is optimal and outperforms the minimum distance precoder.

For the channel (16), we observed that the large constellation assumption made in this paper was not very critical as it produced the same result as a BPSK input constellation does. This is, however, not true in general, and we next investigate the impact of the cardinality of the input constellation. We consider diagonal channel matrices $\mathbf{H}$ where each diagonal element is a zero-mean, unit-variance, circulary symmetric complex Gaussian random variable ($\mathcal{CN}(0,1)$). We consider 4QAM and 16QAM input constellations and plot the resulting average mutual information against SNR for 1) the minimum distance optimal precoder for large constellations, 2) minimum

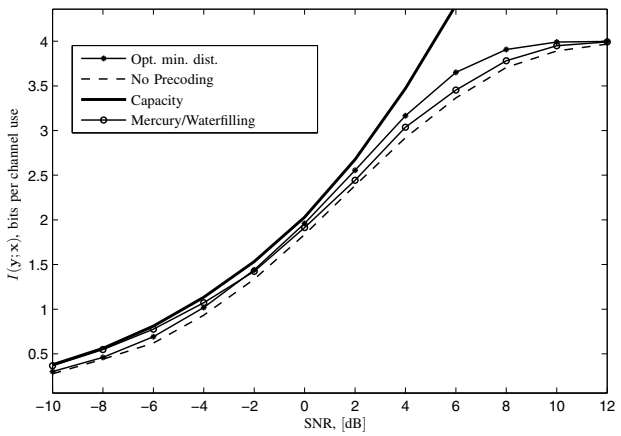


Fig. 3. Mutual information for the channel (16) studied in [3] with 4QAM inputs under different settings. The solid heavy line shows the capacity with waterfilling, the curve marked with asterixes shows the ensuing mutual information from the precoder proposed in this paper and the curve marked with circles show the Mercury/Waterfilling mutual information. The bottom line is the no precoding case.

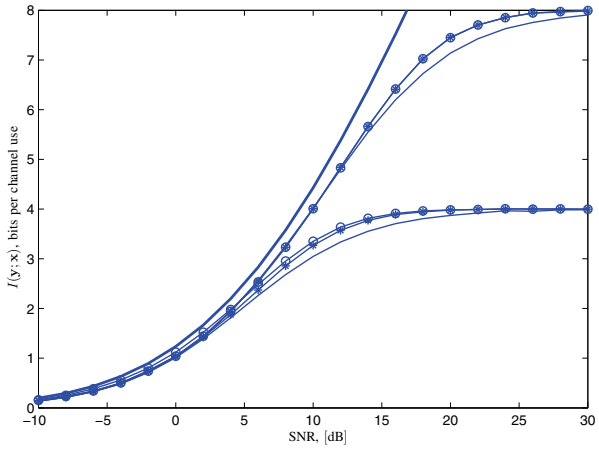


Fig. 4. Average mutual information for random diagonal channels with 4QAM (bottom set) and 16QAM (upper set). The heavy solid line is the capacity with waterfilling. Within each set, the line marked with circles shows the performance of a precoder constructed expliticly for the input constellation used, and the curve marked with asterixes shows the performance of the precoder constructed from an infinite lattice constellation assumption. These two curves are virtually identical for 16QAM. The bottom line within each set corresponds to the no precoding case.

distance optimal precoders for the particular constellations used, and 3) no precoder. The average is evaluated over 10^6 channel realizations by straightforward Monte Carlo simulation. For 4QAM and 16QAM, the minimum distance optimal precoders have been reported in [7], [8], while the optimal precoder for 64QAM has so far not been reported in the literature which is the reason why we do not go beyond 16QAM. The results are shown in Figure 4. The uppermost heavy solid line corresponds to the average capacity of the channel achieved by Gaussian inputs with waterfilling. The lower set of curves corresponds to 4QAM while the upper corresponds to 16QAM. Within each set of curves, the lower curve (without markers) shows the no precoder case, the middle curve (marked with asterixes) is the performance of the precoder constructed from a large constellation assumptions, and the upper curve (marked with circles) is the performance of the precoder explicitly constructed for the input constellation used. For 4QAM inputs, a small loss of the large constellation construction can be seen, while for 16QAM the ensuing mutual information from a large constellation assumption is virtually indistinguishable from that of a construction expliticitly made for 16QAM. Hence, we can conclude from this example that an 16QAM input constellation can be replaced by an infinite lattice constellation without appreciatly affecting the results. It can also be numerically verified, in accordance with the results in [7], [8], [11], that when the constellation size increases, the value of s_1/s_2 for which the precoder changes gets higher and higher. Thus, above a certain cardinality of the alphabet, the same precoder is optimal for all channels $\mathbf{S}$ for which s_1/s_2 is below some large threshold. Therefore our precoder approaches the optimal one for M-QAM as the constellation size increases beyond a certain alphabet, and simulations show that 16-QAM is large enough. This greatly simplifies the precoder optimization problem since lattice theoretic tools can be applied.

In Figure 5 we turn our attention towards the error probability of 2×2 MIMO systems with 1) the minimum distance optimal precoder for large constellations, 2) minimum distance optimal precoders for the particular constellations used, and 3) no precoding. We consider 4QAM, 16QAM, and 64QAM input constellations, and a maximum likelihood detector. The lines marked with circles correspond to the minimum distance optimal precoder for large constellations, the lines marked with asterixes correspond to the optimal precoder designed for the particular input constellations used, and the unmarked lines correspond to the no-precoder case. As can be seen, there is a large gain of explicitly taking the input constellation into account for 4QAM. However, for 16QAM inputs, this gain reduces significantly, so that the precoder designed for large constellations performs close to optimal. For 64QAM, the gap to the optimal precoder designed expliticly for 64QAM can not be determined. However, given the large reduction of the gap between the 4QAM and 16QAM cases, we expect that the gap for 64QAM is minor, so that the precoder designed for large constellations is virtually optimal.

As a final example we consider an OFDM system with N sub-carriers having their channel gains $\{h_k\}_{k=1}^N$. For simplicity, all sub-carrier channel gains are assumed to be independent zero-mean, unit-variance, circulary symmetric complex Gaussian random variables ($\mathcal{CN}(0,1)$). In practice, adjacent carriers are strongly correlated but for the transceiver system to be considered, N is large and such correlations are immaterial. We follow the approach taken in [16] and use the 2×2 minimum distance optimal precoder constructed from the large constellation assumption as a building block to construct much larger precoder structures. The N sub-carriers are first grouped into $N/2$ pairs. The particular pairing used in [16] is to combine the strongest sub-carrier with the weakest sub-carrier, the second strongest with the second weakest etc. Let $\{\tilde{h}_k\}_{k=1}^N$ denote the set of sub-carrier channel gains $\{h_k\}_{k=1}^N$, but sorted

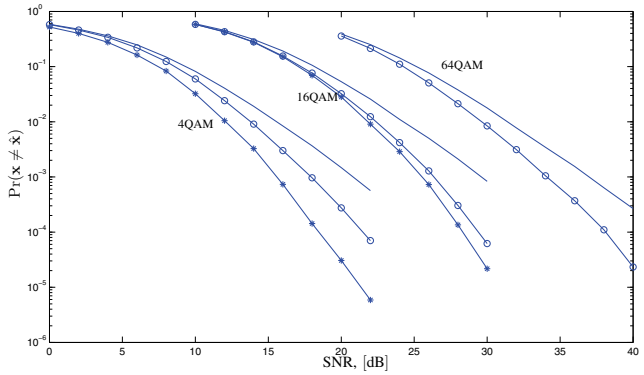


Fig. 5. Maximum likelihood receiver tests of various precoders with 4QAM, 16QAM, and 64QAM. Within each set, the rightmost curve is the no precoding case, the middle curve is the precoder constructed from an infinite lattice constellation assumption, and the leftmost curve is the performance of a precoder constructed expliticly for the input constellation used (not present for 64QAM).

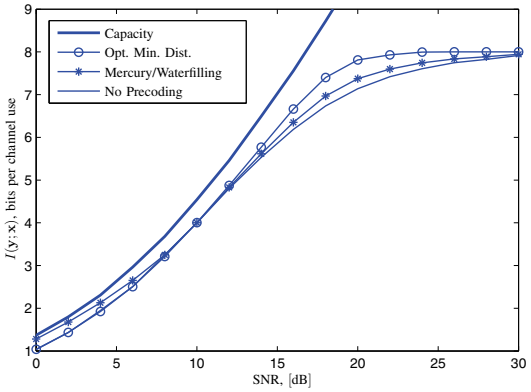


Fig. 6. Average mutual information per sub-carrier pair with 16QAM inputs under different settings. The solid heavy line shows the capacity with waterfilling, the curve marked with circles shows the ensuing mutual information from the precoder proposed in this paper and the curve marked with asterixes shows the mercury/waterfilling mutual information. The bottom line is the no precoding case.

according to their strengths so that $|\tilde{h}_1| \geq |\tilde{h}_1| \geq \ldots \geq |\tilde{h}_N|$. We have $N/2$ independent transmissions

$$\mathbf{y}_k = \begin{bmatrix} \tilde{h}_k & 0 \\ 0 & \tilde{h}_{N-k+1} \end{bmatrix} \mathbf{F}_k \mathbf{x}_k + \mathbf{n}_k = \tilde{\mathbf{H}}_k \mathbf{F}_k \mathbf{x}_k + \mathbf{n}_k, \ 1 \leq k \leq N/2$$

and we need to construct $N/2$ precoders $\{\mathbf{F}_k\}_{k=1}^{N/2}$. A total energy of $N\,P/2$ is assumed, and we allocate a fraction γ_k to $\mathbf{F}_k$ under the constraint that $\sum \gamma_k = N\,P/2$. Our power allocation policy is that all channel-precoder pairs $\tilde{\mathbf{H}}_k \mathbf{F}_k$ should have equal minimum distances. We can find the precoders according to this policy as follows:

- Design $\{\mathbf{F}_k\}_{k=1}^{N/2}$ according to the constraint $\mathrm{Tr}(\mathbf{F}_k^\dagger \mathbf{F}_k) = 1$.
- From lattice theory, it is guaranteed that the minimum distance for each channel-precoder pair equals the length of the shortest vector of the lattice spanned by $\tilde{\mathbf{H}}_k \mathbf{F}_k$. Let D_k^2 denote the minimum distance.
- The power allocation that equalizes all minimum distance is proportional to

$$\gamma_k \propto \frac{1}{D_k^2}$$

and the overall power constraint $\sum \gamma_k = N\,P/2$ finally yields the set of precoders.

We shall compare the ensuing average mutual information of this strategy with the no-precoder case, Mercury/Waterfilling, and the capacity of the channel. The input constellation is 16QAM in all cases (except for the capacity case). The results are shown in Figure 6. Note that we have plotted the average mutual information per channel-precoder pair. The top heavy solid curve is the average capacity of the channel, the curve marked by circles is the system based on the minimum distance optimal precoder described above, the curve marked with asterixes is the Mercury/Waterfilling system, and the bottom curve shows the performance of the no-precoder case. As in the previous examples, there are no gains at low-moderate SNR by the minimum distance optimal precoder, while the gains are significant at high SNR. Note that the Mercury/Waterfilling is close to optimal at low SNR while it suffers from large penalties at high SNR.

VI. Conclusions

This work has provided the optimal minimum distance precoder for $M \times 2$, $M \geq 2$, real-valued and complex-valued channel matrices under the assumption that the input constellation is an infinite lattice. The importance of the minimum distance optimal precoder is that the mutual information precoder converges to the optimal minimum distance precoder as the SNR grows. We have found a profound structure for the minimum distance optimal precoders, namely that for real-valued precoding, the optimal precoder corresponds to the hexagonal lattice at the receiver for *every* non-singular channel matrix. For complex-valued precoding, the precoder corresponds to a complex lattice that is equivalent to the Schläfli lattice in four-dimensional real-valued space. Efficient algorithms to construct the optimal basis for the lattices are given. By numerical studies, we have found that the infinite lattice input constellations can be approximated by conventional 16QAM constellations.

Appendix A: Proof of Theorem 1

First, the constraint in (11) is made more manageable. It follows from (5) that $d^2_{\min}(\Lambda_{\mathbf{QBZ}}) = d^2_{\min}(\Lambda_{\mathbf{B}})$. Let $\mathbf{b}_1, \mathbf{b}_2 \in \mathbb{R}^2$ be the columns of $\mathbf{B}$ and assume that $\|\mathbf{b}_1\| \leq \|\mathbf{b}_2\|$. In 1801, C.F. Gauss noted [17] that if $\mathbf{b}_1$ and $\mathbf{b}_2$ fulfill $|\mathbf{b}_2 \cdot \mathbf{b}_1| \leq \|\mathbf{b}_1\|^2/2$, where "$\cdot$" is the scalar product between vectors, then $d^2_{\min}(\Lambda_{\mathbf{B}}) = \|\mathbf{b}_1\|^2$. This basis is said to be *Gaussian reduced*. Given $\mathbf{b}_1$, the set of all $\mathbf{b}_2$ satisfying the inequality is the *minimum distance region* of $\mathbf{b}_1$. Figure 7 depicts this region geometrically. $\mathbf{b}_1$ and $\mathbf{b}_2$ are actually the *shortest basis* for the lattice, since $\|\mathbf{b}_1\|$ is the length of the shortest vector in the lattice, and it can be shown that $\|\mathbf{b}_2\|$ is the length of the next shortest vector in the lattice. Every lattice has a Gaussian reduced basis, and Gaussian reduction [17] is an algorithm that finds a Gaussian reduced basis from a starting basis of the lattice. Hence, by putting $\|\mathbf{b}_1\| = 1$ and letting $\mathbf{b}_2$ be any

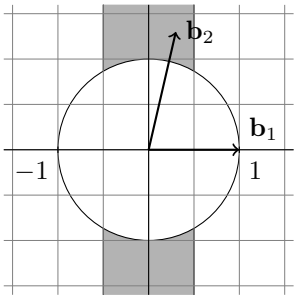


Fig. 7. The minimum distance region of $\mathbf{b}_1$ is shaded. All $\mathbf{b}_2$ inside the shaded region generate a lattice, spanned by $\mathbf{b}_1$ and $\mathbf{b}_2$, with a minimum distance equal to the length of $\mathbf{b}_1$.

vector in the minimum distance region of $\mathbf{b}_1$, the matrix $\mathbf{B}$ will be a generator matrix for any lattice in the plane with unit minimum distance.

Let $r = \|\mathbf{b}_2\|$. The constraint $d_{\min}(\Lambda_{\mathbf{QBZ}}) = 1$ can be written as $r \geq 1$ and $|\cos(\phi)| \leq 1/2r$ where ϕ is the angle between $\mathbf{b}_1$ and $\mathbf{b}_2$. Hence, $\mathbf{QB}$ can be written as

$$\mathbf{QB} = \underbrace{\begin{pmatrix} \sin(\alpha) & -\cos(\alpha) \\ \cos(\alpha) & \sin(\alpha) \end{pmatrix}}_{\mathbf{Q}} \underbrace{\begin{pmatrix} 1 & r\cos(\phi) \\ 0 & r\sin(\phi) \end{pmatrix}}_{\mathbf{B}} = \begin{pmatrix} \sin(\alpha) & r\sin(\alpha-\phi) \\ \cos(\alpha) & r\cos(\alpha-\phi) \end{pmatrix}. \quad (18)$$

Let $\mathbf{W} = \mathbf{ZT}$ in (11). The optimization (11) can now be formulated over α, ϕ and r:

$$\min_{\alpha,\phi,r} \operatorname{tr}(\mathbf{W}^*\mathbf{B}^*\mathbf{Q}^*\mathbf{S}^{-2}\mathbf{QBW}) \quad \text{subject to } r \geq 1,\ |\cos(\phi)| \leq 1/2r. \quad (19)$$

It follows that the intervals for α and ϕ are $0 \leq \alpha \leq 2\pi$, $\cos^{-1}(1/2r) \leq \phi \leq \cos^{-1}(-1/2r)$, $-\cos^{-1}(-1/2r) \leq \phi \leq -\cos^{-1}(1/2r)$.

Let s_1, s_2 be the diagonal elements of $\mathbf{S}$ and assume $s_1 \geq s_2$. Define $s \triangleq s_2/s_1$ and

$$a = \frac{w_{11}^2 + w_{12}^2}{\|\mathbf{W}\|^2} \quad b = \frac{w_{11}w_{21} + w_{12}w_{22}}{\|\mathbf{W}\|^2} \quad c = \frac{1+s^2}{2}. \quad (20)$$

In order to obtain easier expressions, we scale the objective function (19) with $1/s_2\|\mathbf{W}\|^2$ which has no impact on the solution, and by doing so we get the following objective function

$$\begin{aligned} f(\alpha,\phi,r) &\triangleq \operatorname{tr}(\mathbf{W}^*\mathbf{B}^*\mathbf{Q}^*\mathbf{S}^{-2}\mathbf{QBW})/s_2\|\mathbf{W}\|^2 \\ &= c(a + r^2(1-a) + 2br\cos(\phi)) \\ &+ (1-c)(a\cos(2\alpha) + (1-a)r^2\cos(2\alpha+2\phi) \\ &+ 2br\cos(2\alpha+\phi)). \end{aligned} \quad (21)$$

Since $0 \leq s \leq 1$, it follows that $1/2 \leq c \leq 1$.

First, we minimize $f(\alpha,\phi,r)$ over α by making use of the following lemma

Lemma 1: Let $g(x) = \sum_{j=1}^n a_j\cos(x+\theta_j)$ for some real-valued constants $\{a_j\}$ and $\{\theta_j\}$. It holds that

$$\min_x g(x) = -\sqrt{\sum_{j=1,k=1}^{n} a_j a_k \cos(\theta_j - \theta_k)}. \quad (22)$$

Proof: Rewrite $g(x)$ as $g(x) = \mathcal{R}\{\sum_{j=1}^n a_j e^{i(x+\theta_j)}\} = \mathcal{R}\{e^{ix}(\sum_{j=1}^n a_j e^{i\theta_j})\} = \mathcal{R}\{e^{ix}z\}$, where $z \triangleq \sum_j a_j e^{i\theta_j} = |z|e^{i\beta}$. The minimum occurs when z is rotated to the negative part of the real axis, i.e., $x = \pi - \beta$, and the minimum value is then equal to $-|z|$. This gives expression (22). ■

Applying Lemma 1 to (21) in order to minimize over α, we get

$$\begin{aligned} h(\phi,r) &\triangleq \min_\alpha f(\alpha,\phi,r) = c(a + r^2(1-a) + 2rb\cos(\phi)) \\ &+(c-1)[a^2+r^4(1-a)^2+4r^2b^2 \\ &+ 2r^2a(1-a)\cos(2\phi)+4rb(a+r^2(1-a))\cos(\phi)]^{1/2}. \end{aligned}$$

Using the identity $\cos(2\phi) = 2\cos^2(\phi) - 1$ and defining $t \triangleq \cos(\phi)$, we get

$$\begin{aligned} q(t,r) &\triangleq h(\cos^{-1}(t),r) = c(a+r^2(1-a)+2rbt) \\ &+ (c-1)[a^2+r^4(1-a)^2+4r^2b^2 \\ &-2r^2a(1-a)+4r^2a(1-a)t^2+4rb(a+r^2(1-a))t]^{1/2}. \end{aligned} \quad (23)$$

From the definition of t, it follows that $-1/2r \leq t \leq 1/2r$. It can be verified that $q(t,r)$ is a concave function in t. This implies that the minimum of $h(t,r)$ over t is attained at one of the two end points $t = \pm 1/2r$. For these values, and with the variable substitution $\rho = r^2$, we get

$$\begin{aligned} l_\pm(\rho) &\triangleq q(\pm 1/2r, r) \\ &= c(a+\rho(1-a)\pm b)+(c-1)[a^2+\rho^2(1-a)^2+4b^2\rho \\ &\quad -2\rho a(1-a)+a(1-a)\pm 2b(a+\rho(1-a))]^{1/2}, \end{aligned} \quad (24)$$

where $\rho \geq 1$. $l_+(\rho)$ has "+" instead of $\pm$ and $l_-(\rho)$ has "−". The functions $l_\pm(\rho)$ are both concave in ρ, since they geometrically correspond to a hyperbola opening downward. Now, since $l_\pm(\rho)$ is the objective function of (19), which is always positive, it follows that $l_\pm(\rho)$ is positive. The minimum of a positive, concave one-dimensional function is always at the leftmost point of the interval of definition. Hence, the minimum of $l_\pm(\rho)$ must be at $\rho = 1$, which implies that $r = 1$ in (23). This implies that the minimum over t in (23) occurs at $t = \pm 1/2$, which corresponds to $\phi \in \{\pm\pi/3, \pm 2\pi/3\}$ in (21). This shows that the minimum of $f(\alpha,\phi,r)$ in (21) occurs at $r = 1$ and $\phi \in \{\pm\pi/3, \pm 2\pi/3\}$. Inserting these values in the generator matrix $\mathbf{B}$, one obtains the generator matrix for the hexagonal lattice as stated in the theorem. This completes the proof.

Appendix B: Proof of Theorem 2

As in the Proof of Theorem 1, we define $\mathbf{W} \triangleq \mathbf{ZT}$. It turns out that there is a similar minimum distance preserving condition for complex-valued $\mathbf{B}$ as for real-valued ones. In [18], the authors prove that if $\|\mathbf{b}_1\| \leq \|\mathbf{b}_2\|$ and

$$|\mathcal{R}\{\mathbf{b}_1^*\mathbf{b}_2\}| \leq \frac{1}{2}\|\mathbf{b}_1\|^2 \text{ and } |\mathcal{I}\{\mathbf{b}_1^*\mathbf{b}_2\}| \leq \frac{1}{2}\|\mathbf{b}_1\|^2, \quad (25)$$

then $d^2_{\min}(\Lambda_{\mathbf{B}}) = \|\mathbf{b}_1\|^2$. The matrix $\mathbf{QB}$ now has the general form

$$\begin{aligned}\mathbf{QB} &= \underbrace{\begin{pmatrix} e^{i(\phi_1-\gamma_1)} & 0 \\ 0 & e^{i(\phi_3-\gamma_1)}\end{pmatrix} \times \begin{pmatrix} \sin(\alpha)e^{-i\phi_1} & \cos(\alpha)e^{-i\phi_2} \\ \cos(\alpha)e^{-i\phi_3} & -\sin(\alpha)e^{-i\phi_4}\end{pmatrix}}_{Q} \\ &\quad \times \underbrace{\begin{pmatrix} re^{i\gamma_1} & \sin(\omega)e^{i\gamma_2} \\ 0 & \cos(\omega)e^{i\gamma_3}\end{pmatrix}}_{\mathbf{B}} \\ &= \begin{pmatrix} r\sin(\alpha)\sin(\alpha)\sin(\omega)e^{i\theta_1} + \cos(\alpha)\cos(\omega)e^{i\theta_2} \\ r\cos(\alpha)\cos(\alpha)\sin(\omega)e^{i\theta_1} - \sin(\alpha)\cos(\omega)e^{i\theta_2}\end{pmatrix} \quad (26)\end{aligned}$$

where $\phi_1 - \phi_2 \equiv \phi_3 - \phi_4 \pmod{2\pi}$, $\theta_1 = \gamma_2 - \gamma_1$ and $\theta_2 = \gamma_3 - \gamma_1 + \phi_1 - \phi_2$. Note that $\mathbf{QB}$ can be factorized into the product of a rotation matrix ($\mathbf{Q}$) and a basis matrix ($\mathbf{B}$) with one basis vector having a coordinate of 0. Furthermore, the angles γ_1 and γ_3 can be set to 0, since ϕ_1 and ϕ_2 can be used to change θ_2. The conditions (25) become

$$|\mathcal{R}\{\sin(\omega)e^{-i\theta_1}\}| \le \frac{1}{2r} \text{ and } |\mathcal{I}\{\sin(\omega)e^{-i\theta_1}\}| \le \frac{1}{2r}, \quad (27)$$

where $r \ge 1$. Define $f(\alpha,\omega,\theta_1,\theta_2,r) \triangleq \mathrm{tr}(\mathbf{W}^*\mathbf{B}^*\mathbf{Q}^*\mathbf{S}^{-2}\mathbf{QBW})/s_2$. We have

$$\begin{aligned}f(\alpha,\omega,\theta_1,\theta_2,r) &= c\left[r^2(|w_{11}|^2+|w_{12}|^2)\right. \\ &+|w_{21}|^2+|w_{22}|^2+2\mathcal{R}\{(rw_{11}w_{21}^*+rw_{12}w_{22}^*)\sin(\omega)e^{-i\theta_1}\}\big] \\ &+(1-c)\left[r^2(|w_{11}|^2+|w_{12}|^2)-(|w_{21}|^2+|w_{22}|^2)\cos(2\omega)\right. \\ &+2\mathcal{R}\{(rw_{11}w_{21}^*+rw_{12}w_{22}^*)\sin(\omega)e^{-i\theta_1}\}\big]\cos(2\alpha) \\ &-(1-c)\left[(|w_{21}|^2+|w_{22}|^2)\sin(2\omega)\cos(\theta_1-\theta_2)\right. \\ &+2\mathcal{R}((rw_{11}w_{21}^*+rw_{12}w_{22}^*)\cos(\omega)e^{-i\theta_2})\big]\sin(2\alpha), \quad (28)\end{aligned}$$

where $c = (1+(s_2/s_1)^2)/2$. First, we minimize over α. It is seen that f depends on α as

$$\begin{aligned}f(\alpha,\omega,\theta_1,\theta_2,r) &= a_1 + a_2\cos(2\alpha) + a_3\sin(2\alpha) \\ &= a_1 + \sqrt{a_2^2+a_3^2}\left(\frac{a_2}{\sqrt{a_2^2+a_3^2}}\cos(2\alpha)\right. \\ &\quad \left. + \frac{a_3}{\sqrt{a_2^3+a_3^2}}\sin(2\alpha)\right) \\ &= a_1 + \sqrt{a_2^2+a_3^2}(\sin(\psi)\cos(2\alpha) \\ &\quad + \cos(\psi)\sin(2\alpha)) \\ &= a_1 + \sqrt{a_2^2+a_3^2}\sin(2\alpha+\psi), \quad (29)\end{aligned}$$

where the constants a_1, a_2 and a_3 are easily read of from (28) and ψ is such that $\sin(\psi) = a_2/\sqrt{a_2^2+a_3^2}$. The minimum of (29) over α occurs at $\alpha = -\pi/4 - \psi/2$, which gives $f(-\pi/4-\psi/2,\omega,\theta_1,\theta_2,r) = a_1 - \sqrt{a_2^2+a_3^2}$. Since only a_3 depends on θ_2, minimizing f over θ_2 implies maximizing a_3^2 over θ_2. We have

$$\begin{aligned}a_3 &= -(1-c)\left[(|w_{21}|^2+|w_{22}|^2)\sin(2\omega)\cos(\theta_1-\theta_2)\right. \\ &\quad \left. +2\mathcal{R}((rw_{11}w_{21}^*+rw_{12}w_{22}^*)\cos(\omega)e^{-i\theta_2})\right] \\ &= -(1-c)\mathcal{R}\{e^{-i\theta_2}((|w_{21}|^2+|w_{22}|^2)\sin(2\omega)e^{i\theta_1} \\ &\quad +2\cos(\omega)(rw_{11}w_{21}^*+rw_{12}w_{22}^*))\}.\end{aligned}$$

It follows that the maximizing θ_2 is such that $e^{i\theta_2}$ rotates the expression it multiplies to the real axis. We get

$$\begin{aligned}&\min_{\theta_2} f(-\pi/4-\psi/2,\theta_1,\theta_2,\omega,r) = l(\theta_1,\omega,r) \\ &= c\left[r^2(|w_{11}|^2+|w_{12}|^2)+|w_{21}|^2+|w_{22}|^2\right. \\ &+ 2\mathcal{R}\{\sin(\omega)e^{-i\theta_1}(rw_{11}w_{21}^*+rw_{12}w_{22}^*)\}\big] \\ &+ (c-1)\big[\big(r^2(|w_{11}|^2+|w_{12}|^2)+|w_{21}|^2+|w_{22}|^2 \\ &+ 2\mathcal{R}\{\sin(\omega)e^{-i\theta_1}(rw_{11}w_{21}^*+rw_{12}w_{22}^*)\}\big)^2 \\ &- 4\cos^2(\omega)(\det(\mathbf{W}))^2\big]^{1/2}. \quad (30)\end{aligned}$$

As in the real-valued case, it can easily be shown that the expression in (30) is concave in $\sin(\omega)$, since it is a hyperbola opening downward. Thus, the minimum is attained at the endpoints of $\sin(\omega)$. The constraints in (27) can be written as $|\sin(\omega)\cos(\theta_1)| \le 1/2r$ and $|\sin(\omega)\sin(\theta_1)| \le 1/2r$. Assume $|\sin(\theta_1)| \le |\cos(\theta_1)|$. It follows that the interval for $\sin(\omega)$ is $-1/(2r\cos(\theta_1)) \le \sin(\omega) \le 1/(2r\cos(\theta_1))$, while the interval for θ_1 is $-\pi/4 \le \theta_1 \le \pi/4$. Inserting either one of these endpoints for $\sin(\omega)$ in (30) and using the trigonometric identity $1/\cos^2(x) = 1+\tan^2(x)$, we get that l takes on the following form

$$\begin{aligned}l(\theta_1,r) &= c(b_1+b_2\tan(\theta_1)) \\ &\quad + (c-1)\left[(b_1+b_2\tan(\theta_1))^2 + \frac{|\det(\mathbf{W})|^2}{r^2}\tan^2(\theta_1)\right. \\ &\quad \left. + |\det(\mathbf{W})|^2(4-1/r^2)\right]^{1/2}, \quad (31)\end{aligned}$$

where b_1 and b_2 are constants with respect to θ_1. Again, it is clear that (31) is concave in $\tan(\theta_1)$, and thus the minimum is attained at one of the endpoints of θ_1, which are $-\pi/4$ and $\pi/4$. If we instead assumed that $|\sin(\theta_1)| \ge |\cos(\theta_1)|$, the only difference is that $\tan(\theta_1)$ becomes $\cot(\theta_1)$ and $\pi/4 \le \theta_1 \le 3\pi/4$. This gives rise to the same behavior of $l(\theta_1,r)$ and thus same results are obtained.

To recap, we showed that the minimum for $l(\theta_1,\omega,r)$ in (30) over θ_1,ω occurs when $\theta_1 = \pm\pi/4$ and at the endpoints for $\sin(\omega)$, which are then $\sin(\omega) = \pm 1/(2r\cos(\theta_1)) = \pm 1/\sqrt{2}r$. We now continue by inserting this expression for $\sin(\omega)e^{-i\theta_1}$ in (30) and obtain a one-dimensional function in $\rho = r^2$ of the form

$$\begin{aligned}l_1(\rho) &= k_1 + k_2\rho + (c-1)[k_3\rho^2 + k_4\rho + k_5 \\ &\quad + |\det(\mathbf{W})|^2(2/\rho-4)]^{1/2}, \quad (32)\end{aligned}$$

where the k_j are constants with regard to ρ and with k_3 positive. If we instead study the function $l_2(\rho) = k_1 + k_2\rho + (c-1)\sqrt{k_3\rho^2+k_4\rho+k_5-2|\det(\mathbf{W})|^2}$, it follows from the same concavity arguments as before that $l_2(\rho)$ is a concave function and thus the minimum is attained at the endpoints, which are $\rho = 1$ and $\rho = \infty$. From the concavity of $l_2(\rho)$ it follows that if the minimum is attained at ∞, then the minimum value is $-\infty$, which is impossible since our trace function is always positive; thus the minimum of $l_2(\rho)$ must be attained at $\rho = 1$. Now comparing $l_2(\rho)$ with $l_(\rho)$, the only difference is the term $|\det(\mathbf{W})|^2(2/\rho-4)$ in the square root, with maximum value of $2|\det(\mathbf{W})|^2$ attained at $\rho = 1$; hence $l_2(1) = l_1(1)$. Since $c-1$ is always non-positive, it follows that $l_2(\rho) \le l_1(\rho)$ for $\rho \ge 1$, which gives that the minimum

of $l_1(\rho)$ occurs when $\rho = r = 1$ (because the minimum of $l_2(\rho)$ occurs for $\rho = 1$).

We have now showed that the minimum of $l(\theta_1, \omega, r)$ in (30) occurs for $\theta_1 = \pm\pi/4$, $\sin(\omega) = \pm 1/\sqrt{2}r$, $r = 1$. Inserting these values into the lattice generator $\mathbf{MR}$ in (26), we arrive at the following optimal lattice generator $\mathbf{B} = \mathbf{MR}$

$$\mathbf{B} = \begin{pmatrix} 1 & \frac{\pm 1 \pm i}{2} \\ 0 & \pm\frac{1}{\sqrt{2}} \end{pmatrix}. \tag{33}$$

Extending $\mathbf{B}$ to its real-valued representation by means of (6), it holds that for each realization of $\pm$ as $+$ or $-$, that $\mathbf{B}_r$ is a generator matrix for the Schläfli lattice D4.

REFERENCES

[1] I. Telatar, "Capacity of multi-antenna Gaussian channels," *Euro. Trans. Telecommun.*, vol. 10, pp. 585–595, 1999.

[2] A. Lozano, A. M. Tulino, and S. Verdu, "Mercury/waterfilling: optimum power allocation with arbitrary input constellations," *IEEE Trans. Inf. Theory*, vol. 52, no. 7, pp. 3033–3051, July 2006

[3] F. Perez-Cruz, M. R. D. Rodrigues, and S. Verdu, "MIMO Gaussian channels with arbitrary inputs: optimal precoding and power allocation," *IEEE Trans. Inf. Theory*, vol. 56, no. 3, pp. 1070–1084, Mar. 2010.

[4] C. Xiao, Y. R. Zheng, and Z. Ding, "Globally optimal linear precoders for finite alphabet signals over complex vector Gaussian channels," *IEEE Trans. Signal Process.,* vol. 59, no. 7, pp. 3301–3314, July 2011.

[5] M. Payaro and D. P. Palomar, "On optimal precoding in linear vector Gaussian channels with arbitrary input distribution," in *Proc. 2009 IEEE International Symposium on Information Theory*.

[6] J. B. Anderson and A. Svensson, *Coded Modulation Systems*. Kluwer Academic/Plenum Publishers, 2003.

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

[8] Q.-T. Ngo, O. Berder, B. Vrigneau, and O. Sentieys, "Minimum distance based precoder for MIMO-OFDM systems using a 16-QAM modulation," in *Proc. 2009 IEEE Intl. Conf. Commun.*

[9] D. Kapetanovic and F. Rusek, "Design of close to optimal Euclidean distance MIMO-precoders," in *Proc. 2009 IEEE International Symposium on Information Theory*.

[10] S. K. Mohammed, E. Viterbo, Y. Hong, and A. Chockalingam, "MIMO precoding with X- and Y-codes," *IEEE Trans. Inf. Theory*, vol. 57, no. 6, pp. 3542–3566, June 2011.

[11] K. P. Srinath and B. S. Rajan, "A low ML-decoding complexity, full-diversity, full-rate MIMO precoder," *IEEE Trans. Signal Process.*, vol. 59, no. 11, pp. 5485–5498, Nov. 2011.

[12] S. Bergman and B. Ottersten, "Lattice-based linear precoding for MIMO channels with transmitter CSI," *IEEE Trans. Signal Process.*, vol. 56, no. 7, pp. 2902–2914, July 2008.

[13] G. D. Forney Jr. and L.-F. Wei, "Multidimensional constellations—part I: introduction, figures of merit, and generalized cross constellations," *IEEE J. Sel. Areas Commun.*, vol. 7, no. 6, pp. 877–892, Aug. 1989.

[14] D. P. Palomar and Y. Jiang, "MIMO transceiver design via majorization theory," *Foundations and Trends in Commun. and Inf. Theory*, vol. 3, nos. 4-5, 2007.

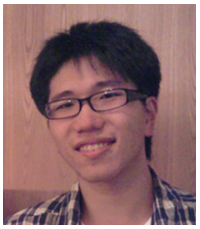
[15] J. H. Conway and N. J. A. Sloane, *Sphere Packings, Lattices and Groups*. Springer-Verlag, 1999.

[16] B. Vrigneau, J. Letessier, P. Rostaing, L. Collin, and G. Burel, "Extension of the MIMO precoder based on the minimum Euclidean distance: a cross-form matrix," *IEEE J. Sel. Topics Signal Process.*, vol. 2, no. 2, pp. 135–146, Apr. 2008.

[17] C. F. Gauss, *Disquisitiones Arithmeticae*, Leipzig 1801. German translation: *Untersuchungen uber die hohere Arithmetik*. Springer, 1889 (reprint: Chelsea, 1981.)

[18] H. Yao and G. W. Wornell, "Lattice-reduction-aided detectors for MIMO communication systems," in *Proc. 2002 IEEE Global Telecomm. Conf.*

Dževdan Kapetanović received his M.Sc. degree in Computer Science in 2007 and the Ph.D. degree in electrical engineering in 2012, both from Lund University, Lund, Sweden. Currently, he is a research associate at the Security and Trust (SnT) center, University of Luxembourg, Luxembourg. His research interests are communication theory and applied information theory.



Hei Victor Cheng received the B.Eng. degree in Electronic Engineering from Tsinghua University, Beijing, China, in 2010, the M.Phil. degree in Electronic and Computer Engineering from Hong Kong University of Science and Technology (HKUST) in 2013. He is now pursuing his Ph.D. studies in the Division for Communication Systems in the Department of Electrical Engineering (ISY) at Linkping University (LiU) in Linkping, Sweden. His current research interests include large-scale MIMO, wireless communications, statistical signal processing and optimization theory.

Wai Ho Mow (S'89-M'93-SM'99) received the M.Phil. and Ph.D. degrees in information engineering from the Chinese University of Hong Kong in 1991 and 1993, respectively. From 1997 to 1999, he was with the Nanyang Technological University, Singapore. He has been with the Hong Kong University of Science and Technology (HKUST) since March 2000. He was the recipient of seven fellowships from ve countries, such as the Humboldt Research Fellowship. His research interests are in the areas of wireless communications, coding, and information theory. He pioneered the lattice approach to signal detection problems (such as sphere decoding and complex lattice reduction-aided detection) and unied all known constructions of perfect (or CAZAC) root-of-unity sequences (widely used as preambles and sounding sequences). He has published one book, and has coauthored over 20 led patent applications and over 150 technical publications, among which he is the sole author of over 40. He coauthored a paper that received the ISITA2002 Paper Award for Young Researchers. Since 2002, he has been the Principal Investigator of 15 funded research projects. In 2005, he chaired the Hong Kong Chapter of the IEEE Information Theory Society. He was the Technical Program Co-Chair of ve conferences, and served the technical program committees of numerous conferences, such as ICC, Globecom, ITW, ISITA, and VTC. He was a Guest (Associate) Editor for three special sections of the *IEICE Transactions on Fundamentals*. He was an industrial consultant for Huawei, ZTE, and Magnotech Ltd. He was a member of the Radio Spectrum Advisory Committee, Ofce of the Telecommunications Authority, Hong Kong S.A.R. Government from 2003 to 2008.

Fredrik Rusek was born in Lund, Sweden in 1978. He received the M.S. and Ph.D. degrees in electrical engineering from Lund University, Sweden, in 2003 and 2007. He currently holds an associate professorship at the Department of Electrical and Information Technology at Lund Institute of Technology. From 2012, he is also part time with Huawei in Lund, Sweden. He is an associate editor for Elsevier Press, and has served in the TPC of a number of conferences. His research interests include modulation theory, equalization, wireless communications and applied information theory.