Demixing Radio Waves in MIMO Spatial Multiplexing: Geometry-based Receivers

Francisco A. T. B. N. Monteiro

Instituto de Telecomunicações and ISCTE - Instituto Universitário de Lisboa
frmo@lx.it.pt

Abstract— **The superposition of waves caused by multipath propagation was for a long time considered an unavoidable nuisance in radio communication links. The discovery that multipath interference was central to enable much larger data rates was a breakthrough at the turn of the century. Multiple-input multiple-output (MIMO) spatial multiplexing (SM) allows unprecedented efficiency in the use of the radio spectrum, however, this comes at the cost of high complexity at the receiver because the underlying symbol detection problem belongs to the class of problems of highest computational complexity. In general, lattice problems are simple to describe but rather hard to solve optimally; finding algorithms to deal with the problem has been a central topic in the last decade of research in MIMO SM. This paper contributes to a deeper understanding of the most important types of receivers for SM with a unifying lattice perspective. Capitalising on that, two novel receivers are proposed. The geometric relation between the primal and the dual lattice is clarified, leading to the proposal of a pre-processing technique that greatly reduces the number of candidate solutions via geometric considerations. Then, looking at lattices from a group theory perspective, it is shown that it is possible to approximate the typical lattices encountered in MIMO by a lattice having a trellis representation, translating the problem for the first time into one manageable by the Viterbi algorithm, well known to the semiconductor industry.**

Keywords: *MIMO, spatial multiplexing, lattices, geometry, closest vector problem, trellis.*

I. INTRODUCTION

A. *From MIMO theory to industry standards*

The last ten years of research in communication engineering have been characterised by extensive research in MIMO. Reaching what Shannon found to be possible for the white additive Gaussian channel was a research task spanning 45 years (from 1948 to 1993). On the other hand, the intense research on MIMO since the late 1990's achieved theoretical breakthroughs on MIMO rather more quickly. In fact, in the last ten years, the research community and the industry were able to put in practice many of the predictions of the theory, breeding a new generation of wireless communication standards, for both indoors and mobile systems. One very important contribution to the dramatic increase of the data rates in these new systems comes from the fact that in MIMO SM the capacity linearly increases with $\min(N_T, N_R)$, i.e., with the minimum number of antennas at each side of the link, where N_T and N_R are the number of antennas at the transmitter and at the receiver, respectively.

MIMO, along with larger channel bandwidth, underpins the physical layer of the fourth generation (4G) wireless networks [1], such as IEEE 802.16 (dubbed WiMAX [2]) and Long Term Evolution (LTE) [3]. In its first releases, LTE relied on MIMO mostly for the downlink (i.e., from the base station (BS) to user equipment) [4], using 4 layers in the downlink (DL). In the latest release 10 (known as LTE-Advanced), the role of MIMO also became important in the uplink (UL). Indeed, MIMO is utilised in both uplink and downlink of the IEEE 802.16m standard (WiMAX profile 2.0). The LTE-Advanced release considers DL with eight layers and uplinks with up to four layers, i.e., a BS with eight or more antennas and user terminals with each 4 antennas [5]. Improving the detection performance with affordable complexity in terminals with 8 layers (with 8×8 antennas) is still today a very important problem [6] (p.181). With these configurations LTE-Advanced achieves spectral efficiencies of 30 b/s/Hz in the DL and 15 b/s/Hz in the UL [7] (p.86), [6] (p.181).

Presently, MIMO is entering the vast domestic market via 802.11n [8], [9] (ch.7), the latest generation of Wi-Fi, designed for a peak rate of 600 Mbps (using 40 MHz bandwidth and 4×4 antennas), which is the first commercial product to be based in MIMO-OFDM. The same combination of MIMO-OFDM has also enabled other standards in the IEEE 802.11 universe [10], particularly those specifically designed for high throughput. Notably, the first wireless standards for data transmission rates over 1Gbps are the 802.15.3c [11] (operating at millimetre waves and using beamforming with antenna arrays) and the 802.11ac [12] (using eight parallel layers in the 5GHz band), both incorporate multiple antennas. These standards, which are scheduled to be concluded by early 2013 [10], will put in to practice the dream "gigabit wireless" anticipated by Paulraj et al. in 2004 [13].

Undoubtedly, MIMO research started within wireless communication and remains until today a prolific research community. However, it is worth noting that the concept of SM has also started to be studied for wave propagation in (multi-modal) optical fibres [14] and, more recently, the concept of taking advantage of radio interference has also been extended to the traditional bundles of cables in wired telecommunication networks, turning crosstalk from a nuisance into an ally to increase the transmission capacity [15].

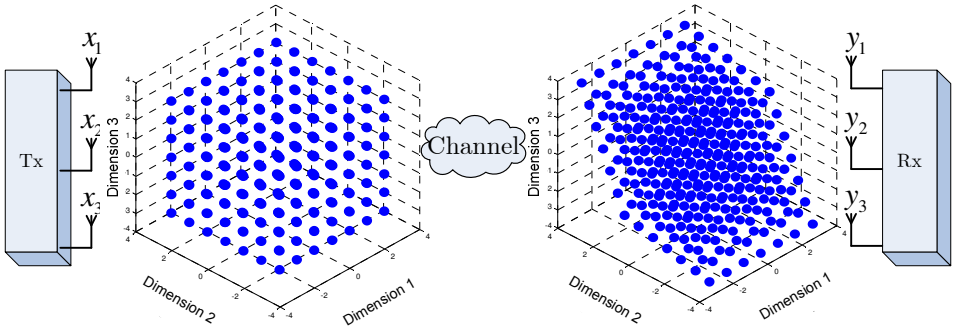


Fig. 1: Spatial multiplexing with real inputs. $\mathbb{Z}^n$ is transmitted and then skewed by the effect of the channel.

How to reduce the complexity that the closest vector problem (CVP) places at the receiver side is a major challenge, and is the main concern of this paper.

B. *Outline of the Paper*

In the next section the MIMO SM setup is established and the detection problem it brings is defined.

Section III introduces the concept of a lattice and clarifies the geometric relationship between a lattice and its dual, which is a central idea to understand some of the stages involved in the two new types of receivers that will be later presented in Section V and Section VI.

Section IV presents the most important detection techniques for MIMO SM: linear filtering, ordered successive interference cancellation (OSIC), lattice-reduction-aided, and the sphere decoding concept. These techniques are presented under a common lattice interpretation where each different detection strategy is shown to consist of some sequence of geometrical manipulations of the lattice structure.

Section V presents a lattice detection strategy for SM which takes advantage of a pre-processing stage based on the geometric relations between the points in the primal lattice and the ones in the dual lattice. This pre-processing finds a set of successive minima in the dual lattice, and is only required at each channel update. The subsequent symbol detection algorithm exclusively involves a linear transformation (the pseudo-inverse) in order to generate a list of candidate solutions for the underlying CVP. The receiver outperforms ordered successive interference cancellation and, in the low signal-to-noise ratio (SNR) regime, also outperforms lattice-reduction-aided receivers at the expense of a "true" sphere-decoder that runs only once per channel update, and not for each received vector.

It is known that some lattices have a trellis representation, however, those lattices require very particular geometries that are not found in lattices randomly generated. Section VI shows that for the typical number of dimensions used in MIMO communication, with high probability, there exists a synthetic lattice that is a member of the family of lattices that have a trellis representation and which is sufficiently close to any given random lattice. For that purpose we present a method to find a trellis-oriented basis for a given random lattice. The basis vectors of the synthetic lattice and the basis vectors of the original lattice are close to each other and, for finite alphabets, the two lattices are roughly the same in the region of interest. Therefore, the optimal decision (Voronoi) regions of both lattices chiefly overlap. A linear transformation then focuses the original lattice onto the synthetic one, known to have a trellis representation. This minimizes the distortion of the Voronoi regions associated with maximum likelihood detection (MLD) and therefore the performance attained in the MIMO-CVP is close to optimal.

II. Spatial Multiplexing: Basic Definitions

A. *System Model for Spatial Multiplexing*

In a MIMO communication channel there is an input vector $\mathbf{x}$ (at a transmitter with N_T elements) and an output $\mathbf{y}$ vector (at the receiver, with N_R elements), which is obtained from $\mathbf{x}$ by means of a linear (matrix) transformation $\mathbf{H}$. When using QAM symbols in each antenna of the transmitter, the set of possible signal can be seen as a set of points carved from a squared multidimensional lattice, the most simple lattice structure one can have. After passing through the channel, this structure is linearly transformed, which geometrically corresponds to warping the squared structure, as suggested in Fig. 1. Furthermore, the detection of $\mathbf{y}$ is perturbed by some noise vector, $\mathbf{n}$, of the same size. In detail, in MIMO SM, with $N_R \geq N_T$, the relationship between the transmitted vector $\mathbf{x}_c = [x_{c,1}, x_{c,2}, \ldots, x_{c,N_T}]^T \in \mathbb{C}^{N_T \times 1}$ and the received vector $\mathbf{y}_c = [y_{c,1}, y_{c,2}, \ldots, y_{c,N_R}]^T \in \mathbb{C}^{N_R \times 1}$ is modelled in the baseband as

$$\mathbf{y}_c = \mathbf{H}_c \mathbf{x}_c + \mathbf{n}_c, \tag{1}$$

where $\mathbf{H}_c \in \mathbb{C}^{N_R \times N_T}$ is the channel matrix with its entries h_{ij} representing the complex coefficient associated with the link between he i^{th} receive antenna and the j^{th} transmit antenna.

In the case of Rayleigh flat fading channel, h_{ij} are taken from a zero-mean circularly symmetric complex Gaussian distribution with unitary variance (i.e., variance 1/2 in both the

real and imaginary components). Furthermore, the noise added to each entry of the received vector is modelled by the column vector $\mathbf{n}_c = [n_{c,1}, n_{c,2}, \ldots, n_{c,N_R}]^T \in \mathbb{C}^{N_R \times 1}$ with independent circularly symmetric complex Gaussian random variables taken with zero mean and variance σ_n^2 (corresponding to a variance $\sigma_n^2 / 2$ in both real and imaginary components). For independent input data, its covariance is $\mathbf{R}_x = E\{\mathbf{x}_c \mathbf{x}_c^H\} = \sigma_x^2 \mathbf{I}_n$. Similarly, the covariance of the independent noise vector is $\mathbf{R}_n = E\{\mathbf{n}_c \mathbf{n}_c^H\} = \sigma_n^2 \mathbf{I}_n$. The energy of the complex transmitted symbols is assumed to be $E\{x_{c,i}^2\} = 1$. The real symbols in each dimension are taken from an alphabet $\mathcal{A}$, with $\sqrt{M}$ symbols, in the case of *M*-QAM modulation in each antenna, which is the modulation type considered in this paper, where

$$\mathcal{A} = \sqrt{\frac{3}{2(M-1)}} \Big\{ -(\sqrt{M}-1), \cdots, -5, -3, -1, +1, +3, +5, \cdots + (\sqrt{M}-1) \Big\}. \quad (2)$$

It is not difficult to prove that by stacking the real and complex parts of the vectors (respectively denoted by $\Re$ and $\Im$), and by appropriate construction of a modified channel matrix, the problem can equivalently be described by means of real variables as

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n} \Leftrightarrow$$
$$\Leftrightarrow \begin{bmatrix} \Re(\mathbf{y}_c) \\ \Im(\mathbf{y}_c) \end{bmatrix} = \begin{bmatrix} \Re(\mathbf{H}_c) & -\Im(\mathbf{H}_c) \\ \Im(\mathbf{H}_c) & \Re(\mathbf{H}_c) \end{bmatrix} \begin{bmatrix} \Re(\mathbf{x}_c) \\ \Im(\mathbf{x}_c) \end{bmatrix} + \begin{bmatrix} \Re(\mathbf{n}_c) \\ \Im(\mathbf{n}_c) \end{bmatrix} \quad (3)$$

with all vectors now real (hence, the "*c*" subscripts in the variables are dropped from now on).

The conversion in (3) makes possible to take any lattice defined in a complex domain (such as the ones arising from signalling with electromagnetic waves), and transform it into full-real lattices with the double of dimensions. Henceforth, this paper will deal with real lattices in $2N_R$ dimensions.

B. Diversity and Multiplexing

From Shannon we know that, in the AWGN channel, a symbol error rate (SER) curve as a function of the signal to noise ratio (SNR), ρ, can be as steep as one wants. In the limit, the SER curve, $P_s(\rho)$, can have an infinite negative slope. For the Rayleigh fading channel it is well known that $P_s(\rho)$ exhibits a -1 slope in the uncoded SISO case, that is, one finds that $P_s(\rho) \propto \rho^{-1}$. One door that MIMO opens is the possibility of increasing (in modulo) that slope, i.e., obtaining a faster reduction of the error rate as SNR increases. One defines the *diversity order*, corresponding to the slope

$$d = \lim_{SNR \to \infty} -\frac{\log(P_s(\rho))}{\log(\rho)}. \quad (4)$$

This *diversity order* measures how many statistically independent copies of the same symbol the receiver is able to receive. In brief, this amounts to the number of independent fading coefficients that the receiver can average in order to produce a reliable estimate of a transmitted symbol. Not surprisingly, the maximum available diversity that can be attained is $d_{\max} = N_T N_R$.

In the MIMO general case, this gain is defined as

$$g = \lim_{\mathrm{SER} \to \infty} \frac{\log(R)}{\log(\rho)} \quad (5)$$

When plotting the symbol error rate (SER) versus the signal to noise ratio (SNR), the existing diversity d in the communication link is simply the slope (in the asymptotic regime) of the SER curve. On the other hand, the interpretation of the multiplexing gain in a typical SER plot is not so straightforward. The metric g indicates how the capacity increases with the SNR, which is a common representation in information theory since Shannon but is less useful in practice. In terms of the SER, the multiplexing gain g measures how fast spectral efficiency can increase with the increase of SNR while keeping the same error rate and corresponds to the maximum number of independent *layers* or *parallel channels* and is limited by

$$g_{\max} = \min(N_T, N_R). \quad (6)$$

In a theoretical breakthrough paper [16], Zheng and Tse showed that there is a trade-off between d and g, i.e., the famous diversity-multiplexing trade-off (DMT): increasing one leads to a decrease in the other.

C. Optimal reception problem

The main research problem in spatial multiplexing in the last ten years has been detecting $\mathbf{x}$ given the noisy observation $\mathbf{y}$ in order to minimise the symbol error rate (SER). For that problem, the maximum a posterior probability (MAP) of $\mathbf{x}$ is

$$\mathbf{x} = \arg\max_{\mathbf{x} \in \mathcal{A}} P\left(\mathbf{x} \middle| \mathbf{y}\right) = \arg\max_{\mathbf{x} \in \mathcal{A}} \frac{P\left(\mathbf{x} \middle| \mathbf{y}\right) P(\mathbf{x})}{P\left(\mathbf{y}\right)} \quad (7)$$

As all vectors $\mathbf{x}$ are equiprobable, $P\left(\mathbf{x} \middle| \mathbf{y}\right)$ is a sufficient statistics for the detection process. Therefore, MAP detection can be reduced to *maximum likelihood* (ML) without any performance loss. For the i.i.d. Rayleigh channel with i.i.d. transmitted symbols with both the $\mathbf{R}_x$ and $\mathbf{R}_n$ given above, one has the *N*-dimensional probability distribution

$$P\left(\mathbf{x} \middle| \mathbf{y}\right) = \frac{1}{\left(2\pi\sigma_n^2\right)^{N/2}} \exp\left(-\frac{\left\| \mathbf{y} - \mathbf{H}\mathbf{x} \right\|^2}{2\sigma_n^2} \right), \quad (8)$$

and therefore the detection problem becomes that of minimizing the exponent in (8):

$$\hat{\mathbf{x}}_{ML} = \arg\max_{\mathbf{x}\in\mathcal{A}} \left\{ \left\| \mathbf{y} - \mathbf{H}\mathbf{x} \right\|^2 \right\}. \quad (9)$$

This problem now has a clear *geometrical* interpretation: the optimal $\mathbf{x}$ (the one that best explains the observation $\mathbf{y}$) is the one that, among all possible input vectors, and after the linear transformation, generates the closest vector $\mathbf{H}\mathbf{x}$ (in the Euclidian sense) to the received vector $\mathbf{y}$. This problem is known in integer optimisation as *integer least squares* and in lattice theory as CVP (as mentioned above): "given a target vector off the lattice, $\mathbf{y}$, which point in the lattice is the closest one?".

The algorithmic complexity of the CVP is proven to be *NP-hard* [17], which, in the current state of understanding of the complexity of algorithms, places it in the worst tier in the hierarchy of complexity classes. Notice that this involves measuring the Euclidian distance of the received vector $\mathbf{y}$ to all the possible $\mathcal{A}_c^n$ transmit points in the (finite) complex lattice (or $\mathcal{A}^{2n}$ in the real equivalent lattice). For example, for a 4×4 configuration with 64 QAM (i.e., 8-PAM per real components in a 8-D lattice), this amounts to a comparison of $64^4 = 16{,}777{,}216$ distance metrics. The same number arises when using the real equivalent model (3), as a result of 8^8 metric comparisons . This is why the 8×8 configuration for LTE receivers is still a major open challenge in 2012 [6].

One should not conclude from this that any hope of finding accurate solutions should be deemed unrealistic. In fact, it will be seen throughout this work that very good approximations to the optimal solution can be found, especially when the number of dimensions is small [18]. As the number of dimensions grows, the complexity of the problem, measured as the number of operations, grows exponentially (this is what is known as "the curse of dimensionality" [19]); however, the complexity of some approximate detection techniques grows only polinomially (zero-forcing, MMSE, SIC, and LLR-based, as presented in Section IV).

III. Primal and Dual Lattice Geometry

A. Definition of a Lattice

Lattices are discrete subgroups in $\mathbb{C}^n$. The most common manner to specify a lattice Λ is based on on a set of vectors which are the columns of a *generator matrix* $\mathbf{H}$:

$$\Lambda = \left\{ \mathbf{y} \in \mathbb{C}^n : \mathbf{y} = \sum_{i=1}^{n} \mathbf{h}_i x_i = \mathbf{H}\cdot\mathbf{x}\,,\ \ x_i \in \mathbb{Z}, \mathbf{h}_i \in \mathbb{C}^n \right\}. \quad \text{c}$$

The coordinates of the lattice points are thus integer combinations of the columns of the complex generator matrix $\mathbf{H}$, as exemplified in Fig. 2.

Given a certain basis of a lattice, the *fundamental region* that is associated to that basis is defined as

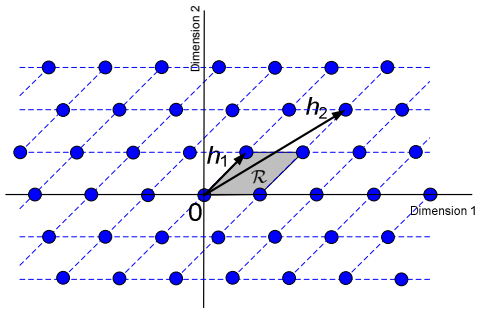


Fig. 2: A lattice in $\mathbb{R}^2$ and the fundamental region associated with a particular basis.

$$\mathcal{R}(\mathbf{H}) = \left\{ \mathbf{H}\mathbf{x} : 0 < x_i < 1 \right\}. \quad (10)$$

The fundamental region cannot contain any lattice point inside it (c.f. Fig. 2). If that happens, then the set of vectors is *not* a basis of the lattice but a basis of one of its sublattices. A sublattice Λ' is also a lattice and the volume is $\mathrm{vol}(\Lambda') > \mathrm{vol}(\Lambda)$ (the technical definition of the volume of a lattice will shortly be given).

Note that different sets of vectors may generate the same lattice. Indeed, the number of admissible bases for a lattice is infinite; it is easy to infer from Fig. 2 that it is always possible to select some point further distant from the origin to replace a generator and still have a fundamental region without including any lattice point in its interior. Moreover, all these different bases are related by unimodular transformations, as it will be described below.

The region of the space where the lattice is embedded that contains all the points in the span of the lattice (i.e., in the continuous Euclidian space where the lattice exists) which are closer to a given lattice point $\mathbf{x}$ than to any other point in the lattice is called the *Voronoi region* and is defined by

$$\mathcal{V}(\Lambda) = \left\{ \mathbf{z} \in \mathrm{span}(\Lambda) : \forall \mathbf{y} \in \Lambda \ \left\| \mathbf{x} - \mathbf{z} \right\| < \left\| \mathbf{y} - \mathbf{z} \right\| \right\}. \quad (11)$$

This (open) region is a characteristic of the lattice and independent of any particular generating matrix, and is the most interesting fundamental region (amid the infinite number of other possible fundamental regions one can define to tile the entire space) as it constitutes the optimal decision region for the closest vector problem in a lattice.

The *Gram matrix* of a lattice is defined by the columns of $\mathbf{H}$ as (in the complex case the Hermitian operator replaces the transposition) as $\mathbf{G} = \mathbf{H}^T\mathbf{H}$.

By construction, the Gram matrix contains all the possible inner products between all the generator vectors: $g_{ij} = \left\langle \mathbf{h}_i, \mathbf{h}_j \right\rangle$; in particular, the diagonal elements are the squared norms $\left\| \mathbf{h}_i \right\|^2$. This fact implies that $\mathbf{G}$ is symmetric and positive definite.

B. The Geometry of the Dual Lattice

Every lattice (said to be the *primal* lattice) has a *dual lattice*, also known as the *polar lattice* or, more commonly, as the *reciprocal lattice*. Note that these names were already in use

in the early 70's [20] (p.24). Since then, the name polar has fallen into disuse, though reciprocal can still be found in some literature. The dual lattice is traditionally defined for real lattices, though the definition has also been extended to complex lattices [21]. Given the intuitive geometrical interpretation that is possible in the real domain, the dual lattice is usually defined for real lattices as

$$\Lambda_D = \left\{ \mathbf{z} \in \mathbb{R} : \left\langle \mathbf{z}, \mathbf{x} \right\rangle \in \mathbb{Z} \quad , \forall \mathbf{x} \in \Lambda \right\}. \tag{12}$$

The dual lattice can also be expressed in terms of the dual basis $\mathbf{H}^{(D)}$ as

$$\Lambda_D = \left\{ \mathbf{z} \in \mathbb{R} : \mathbf{z} = \underbrace{\left(\mathbf{H}^{+}\right)^{T}}_{\mathbf{H}^{(D)}} \mathbf{x}, \quad \mathbf{x} \in \mathbb{Z}^n \right\}. \tag{13}$$

where $\mathbf{H}^{+}$ is the Moore-Penrose pseudo-inverse

$$\mathbf{H}^{+} = (\mathbf{H}^H \mathbf{H})^{-1} \mathbf{H}^H. \tag{14}$$

Hence,

$$\mathbf{H}^{(D)} = \mathbf{H}^T \left(\mathbf{H}^T \mathbf{H}\right)^{-1}. \tag{15}$$

In fact, for $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{Z}^n$,

$$\left\langle \mathbf{z}, \mathbf{y} \right\rangle = \mathbf{z}^T \mathbf{x} = \underbrace{\left(\mathbf{H}^{+}\right)^{T} \mathbf{x}_1}_{\mathbf{z} \in \Lambda^{(D)}} \underbrace{\mathbf{H}\mathbf{x}_2}_{\mathbf{y} \in \Lambda} = \mathbf{x}_1^T \mathbf{H}^{+} \mathbf{H} \mathbf{x}_2 = \mathbf{x}_1^T \mathbf{x}_2 \in \mathbb{Z}.$$

It is also possible to show that each point in the dual lattice can be written as an integer combination of the columns of $\mathbf{H}^{(D)}$. Let us focus on the case of full rank real matrices where $\mathbf{H}^{+} = \mathbf{H}^{-1}$. Denoting the rows of $\mathbf{H}^{-1}$ by $\mathbf{r}_1, \mathbf{r}_2, \cdots, \mathbf{r}_n$, for any point $\mathbf{z} \in \Lambda^{(D)}$ it is possible to write

$$\begin{aligned} \mathbf{z}^T &= \mathbf{z}^T \mathbf{H} \mathbf{H}^{-1} \\ &= (\underbrace{\mathbf{z}^T \mathbf{h}_1}_{\in \mathbb{Z}}) \mathbf{r}_1 + (\underbrace{\mathbf{z}^T \mathbf{h}_2}_{\in \mathbb{Z}}) \mathbf{r}_2 + \cdots (\underbrace{\mathbf{z}^T \mathbf{h}_n}_{\in \mathbb{Z}}) \mathbf{r}_n, \end{aligned} \tag{16}$$

which shows that the point in the dual lattice is defined by a linear combination of the rows of $\mathbf{H}^{-1}$, i.e., a linear combination of the columns of $(\mathbf{H}^{-1})^T$. These arguments can be extended to the cases where the Moore-Penrose inverse is required and also to complex lattices.

One interesting relationship between the two bases is that

$$\left(\mathbf{H}^{(D)}\right)^{T} \mathbf{H} = \mathbf{I}, \tag{17}$$

which is equivalent to saying that $\left\langle \mathbf{h}_i, \mathbf{h}_j^{(D)} \right\rangle = \delta_{i,j}$, using the Kronecker delta.

The volumes of the primal and the dual lattice are related by $\mathrm{vol}\left(\Lambda_D\right) = \left(\mathrm{vol}(\Lambda)\right)^{-1}$ and their Gram matrices are related by $\mathbf{G}^{(D)} = \mathbf{G}^{-1}$.

Obviously, the dual of the dual lattice is the primal lattice itself. The geometry of the dual lattice is closely related to the geometry of the primal lattice. The connection is that each point in the n-dimensional dual lattice defines a family of parallel $(n-1)$ dimensional hyperplanes onto which translates of a $(n-1)$-dimensional sublattice lie. The union of those planes captures all the points of the primal lattice. This means that the shortest vector in the dual lattice will define the most distant $(n-1)$-dimensional hyperplanes, whose union builds up the whole primal lattice. These hyperplanes can be interpreted as parallel layers and (as a consequence of being the ones furthest apart) are the densest ones in the lattice. In MIMO literature, the geometrical interpretation of the dual lattice as a tool for improving detection seems to have been first noticed in [22] (p. 2207) for sphere decoding, and then in [23] and [24], though it is also implied in the detector in [25] (p. 1944). From definition (13), in both Λ and $\Lambda^{(D)}$, the inner product between some given point $\mathbf{z}$ in the dual lattice and any vector in the primal lattice is always an integer, and therefore,

$$\begin{aligned} &\left\langle \mathbf{z}, \mathbf{x} \right\rangle \in \mathbb{Z}, \quad \mathbf{z} \in \Lambda^{(D)}, \mathbf{x} \in \Lambda \Leftrightarrow \\ &\Leftrightarrow \left\| \mathbf{z} \right\| \left\| \mathbf{x} \right\| \cos(\theta) = \left\| \mathbf{z} \right\| \mathrm{Proj}_{\mathbf{e}_z}(\mathbf{x}) \in \mathbb{Z}, \end{aligned} \tag{18}$$

where $\overline{\mathbf{z}} = \mathbf{z} / \left\| \mathbf{z} \right\|$. It is then possible to define a family of parallel hyperplanes $\mathcal{P}(\nu)$, for $\nu \in \mathbb{Z}$, such that $\mathrm{Proj}_{\overline{\mathbf{z}}}(\mathbf{x}) = \left\| \mathbf{z} \right\|^{-1} \nu$. These are planes in dimension $n-1$, with a distance $d = \left\| \mathbf{z} \right\|^{-1}$ between them, as illustrated in Fig. 1. Remark: all vectors $\mathbf{a}_i$ in a given hyperplane have the same inner product with $\mathbf{z}$.

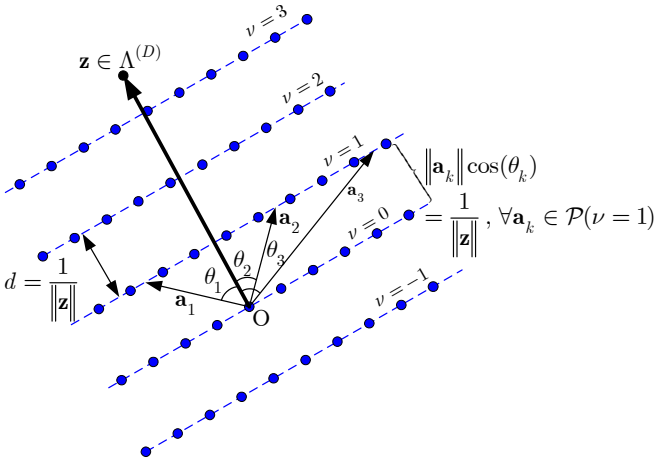


Fig. 1: A primal lattice in n dimensions as the union of translates of a sublattice and these translates lie on $(n-1)$-dimensional hyperplanes.

IV. Geometric Interpretation of MIMO Receivers

This section starts by introducing the most important type of MIMO receivers and the geometric concepts associated with them, which explain their performance loss in respect to the optimum detector. The linear receivers, which are the simplest ones, but also the ones having the worst SER, are the

first ones to be geometrically interpreted. Then, the ordered successive interference cancelation (OSIC) technique is described, followed by the lattice-reduction-aided (LRA) approach, and then the sphere-decoding concept is introduced.

A. Linear receivers: ZF and MMSE

Linear receivers consists of i) a linear transformation $\mathbf{W}$ of the received vector which then followed by ii) a quantisation to the symbol alphabet (also known as *slicing*). The linear transformation is a filter that can be designed with two different criteria, leading to the *zero-forcing* (ZF) detector or to the *minimum mean square error* (MMSE) detector. These receivers constitute the simplest set of (non optimal) receivers to be widely used for MIMO receivers. The detected solution $\mathbf{x}$ given by these techniques is obtained by applying

$$\mathbf{x}_W = \mathbf{W}\mathbf{y}, \tag{19}$$

$$\hat{\mathbf{x}}_W = Q_{\mathbb{Z}}\left[\mathbf{x}_W\right], \tag{20}$$

where $Q_{\mathbb{Z}}[\cdot]$ denotes rounding to the nearest integer and the subscript W in $\hat{\mathbf{x}}_W$ indicates the filter design criterion: ZF or MMSE. Both linear receivers offer solutions to (9) that involve the inversion of $\mathbf{H}$, which implies a number of operations $\mathcal{O}(n^3)$ [26] (p. 170). Note that the number of operations is taken as a standard metric to compare the computational complexity of the different detection techniques.

1) Zero-forcing receiver

It is natural to think first of a solution to (19) involving the linear transformation that *undoes* the linear transformation, which is obviously the inverse matrix, i.e., that takes the structure on the right side of Fig. 1 and reverts it to the original cubic structure on the left side of Fig. 1. In both types of linear receivers the linear transformation $\mathbf{W}$ can be seen as a focusing process of the points in the received lattice back onto $\mathbb{Z}^n$ (or $\mathbb{C}^n$). This "backwards transformation" is of interest because it maps the received lattice back onto $\mathbb{Z}^n$, which lends itself to simple orthogonal *slicing*. Besides this reason, there is no other motivation for this particular design. In Section VI the concept of a having a linear transformation as the first stage of a detection technique will be generalised to the concept of *focusing* a received lattice onto some other given lattice, whose geometric structure is also of interest (although it will no longer be the cubic $\mathbb{Z}^n$).

The Moore-Penrose inverse of $\mathbf{H}$, when $N_R \geq N_T$, always exists and is defined as $\mathbf{H}^+ = (\mathbf{H}^H\mathbf{H})^{-1}\mathbf{H}^H$. If $N_R = N_T$, and for full rank channel matrices, this inversion amounts to a simple matrix inversion $\mathbf{W} = \mathbf{H}^{-1}$, which leans itself to a simple geometrical interpretation.

$$\hat{\mathbf{x}}_{ZF} = Q_{\mathcal{A}}\left[\mathbf{W}_{ZF}(\mathbf{H}\mathbf{x}+\mathbf{n})\right] = Q_{\mathcal{A}}\left[\mathbf{x}+\mathbf{W}_{ZF}\mathbf{n}\right]. \tag{21}$$

The filtered noise is transformed by $W_{ZF} = \mathbf{H}^+$, which constitutes a noise enhancement factor. The receiver structure is shown in Fig. 3.

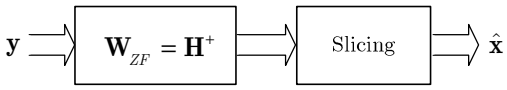


Fig. 3: Zero-forcing receiver.

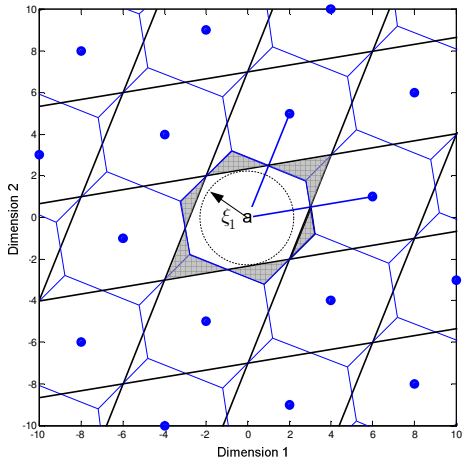
The detected vector $\hat{\mathbf{x}}_{ZF}$, as obtained from (21), is in fact the solution to

$$\hat{\mathbf{x}}_{ZF} = \arg\min_{\mathbf{x}\in\mathbb{R}^{N_T}} \left\{\left\|\mathbf{y}-\mathbf{H}\mathbf{x}\right\|\right\}. \tag{22}$$

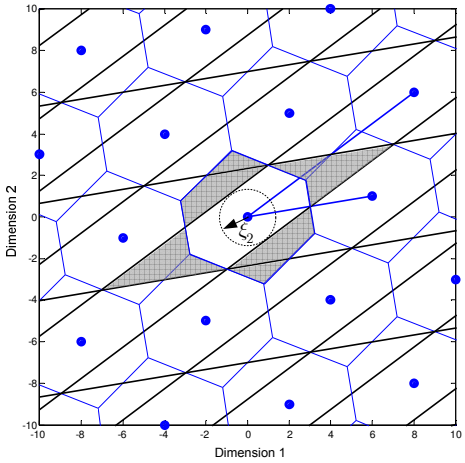
As is mentioned in the previous section, ZF solves the CVP by relaxing it to a search in a continuous neighbourhood instead of computing the distance between the received vector (also called the *target*) and every point in the lattice. The geometrical implication can be better understood thinking of the linear transformation of the hypercubic Voronoi regions of $\mathbb{Z}^n$ by $\mathbf{H}$. The resulting regions are called the *ZF decision regions* and correspond to the space where a lattice point will be interpreted as being close to the lattice point associated with that region.

The decision regions associated with ZF criterion are simple to obtain as they are the fundamental region $\mathcal{R}(\mathbf{H})$, as defined in (10). Because the lattices in MIMO are Gaussian, the basis generated by a channel may have some highly correlated vectors. Geometrically, this corresponds to lattices with very narrow fundamental regions, which are generated by ill-conditioned matrices, i.e., when one or more singular values are close to zero, and consequently the volume of the lattice vanishes. Fig. 4 shows the ZF decision regions associated with two equivalent bases (the relation between the two bases will be shown below in Section IV-A-4).

Let us concentrate in the case where the transmit point was the origin. The shaded areas indicate regions which will lead to wrong decisions when using the ZF technique: either because the point is inside the Voronoi region and outside the ZF decision region or because the closest lattice point would be decided as being the origin while the Voronoi region shows that to be false. It is possible to observe in Fig. 4 that different bases will output different decisions given a target point. For the examples at a given SNR, the SER with $\mathbf{H}_1$ will be always lower, because the *coverage* of the MLD (i.e., Voronoi) regions is larger in the case of basis $\mathbf{H}_2$ (in the sense that the decision region intersects the Voronoi region in a larger volume). The notion of *coverage* is essential to understand MIMO detection [27]. In order to simplify the operational meaning of coverage, Ling [28] introduced the notion of *proximity factors* dependent on the notion of the largest sphere that can be fitted inside the region of coverage. These spheres are also shown in Fig. 4 for the two basis, having *decoding radii* ξ_1 and ξ_2 respectively.



(a) $\mathbf{H}_1 = \begin{bmatrix} 6 & 2 \\ 1 & 5 \end{bmatrix}$.



(b) $\mathbf{H}_2 = \begin{bmatrix} 6 & 8 \\ 1 & 6 \end{bmatrix}$.

Fig. 4: Decision regions associated with the two different bases of the same lattice.

A receiver with a better performance is the one whose decision regions better approximate the shape of the regions associated with MLD. The receiver to be presented in Section VI aims at to maximise this matching between its decision regions and the ones of MLD.

2) Minimum Mean Squared Error Detection

The other (and more sophisticated) linear receiver aims at finding the filter that minimises the mean squared error between the estimated vector and the original vector, i.e., the filter should be

$$\mathbf{W}_{MMSE} = \arg\min_{\mathbf{W}} E\left\{\left\|\mathbf{W}\mathbf{y} - \mathbf{x}\right\|^2\right\}. \qquad (23)$$

This criterion does not aim at cancelling all the interference between layers as ZF does. Instead, the MMSE criterion takes into consideration both the interference *and* the noise in order to minimise the expected error. This minimization implies finding the point where the gradient of the objective function in (23) is zero. There is however a fast track to finding this estimator by applying the *orthogonality principle*, well known in estimation theory and widely used in equalisation problems in the ISI channel [29] (secs. 2.2.3, 2.3.4), [30] (sec. 5.2), [31] (sec. 5.6). The optimum estimator for (23) is the one that produces an error vector $\Delta = \mathbf{W}_{MMSE}\mathbf{y} - \mathbf{x}$ that is orthogonal to received signal, i.e., the two vectors are uncorrelated (as illustrated in Fig. 5).

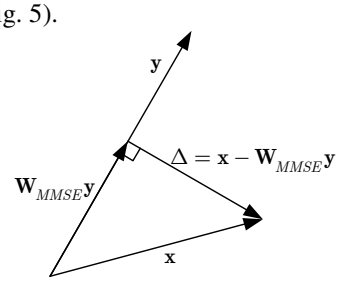


Fig. 5: Orthogonality principle: the expected error is made orthogonal between the receive vector and the space where the best solution is searched.

The MMSE receiver is implemented with the block diagram (and the filter $\mathbf{W}_{MMSE}$) specified in Fig. 6.

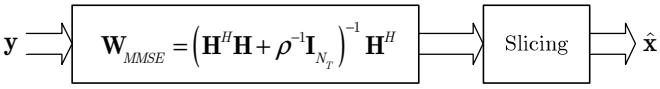


Fig. 6: MMSE receiver.

The minimum norm $\|\Delta\|$ occurs when $\mathbf{W}_{MMSE}\mathbf{y} \perp \Delta$, i.e.,

$$E\left\{(\mathbf{W}_{MMSE}\mathbf{y} - \mathbf{x})\mathbf{y}^H\right\} = 0. \qquad (24)$$

B. Ordered Successive Interference Cancellation

This detection algorithm first proposed in [32], uses the principles of SIC, already known in ISI control and MUD and is also known as the V-BLAST detector. The general principle of SIC is that an initial "best" layer is detected and then, assuming that the symbol was correctly detected, the interference caused by that symbol is replicated and subtracted from all the other layers. The procedure is then applied to the "next best" layer: one symbol more is detected, its interference recreated and then subtracted from the remaining ones.

One important question that arises is the one of determining the order of detection of the N_R antennas. For a MIMO $n \times n$ system one has to find the optimum permutation $\Pi(k)$ of the column indexes $\{1, 2, \cdots, n\}$ that minimises the SER amid all the $n!$ possible permutations. An exhaustive search over all the permutations would rapidly become unbearable as n increases. The optimal solution to this problem was found early on in [33], [32], in the first implementations of the V-BLAST detector. The optimal criterion at each stage is to select the layer that less emphasises the noise power after a ZF

or a MMSE filter. From these facts it is now possible to devise a geometric proof to the optimal ordering of the layers in OSIC and connect it with the geometric ideas of the Babai's nearest plane algorithm [34], [17] (ch.2) in algorithmic number theory, which actually corresponds to SIC in MIMO, as first noticed by[22], [35].

In order to minimise the error probability when deciding layer *j*, the generator vector $\mathbf{h}_j$ to be selected at any given decision step *k*, with $k \in \{1,2,\cdots,n\}$, should be the vector that maximises the projection onto the orthogonal space to the space spanned by the matrix that remains after that same vector is taken out from $\mathbf{H}$.

The initial step is to find the column vector $\mathbf{h}_1$ that, when removed from $\mathbf{H}$, transforms $\mathbf{H}$ into $\mathbf{H}_{\bar{1}}$ (as $\mathbf{H}_{\bar{1}}$ denotes the matrix that is obtained from $\mathbf{H}$ after removing column *j*). $\mathbf{H}_{\bar{1}}$ is the generator of an $(n-1)$ -dimensional lattice Λ_{n-1}. Hence, the original lattice can be written in the form

$$\Lambda = \Lambda_{n-1} + i\mathbf{h}_1, i \in \mathbb{Z}, \quad (25)$$

signifying that Λ can be created from the *union* of *translates* of the Λ_{n-1} sublattice.

The diversity attained by SIC is $N_R - N_T + 1$ and sorting the layers does not contribute to any improvement in this respect, as recently proven in [36]. Sorting can only yield a power gain in SM detection.

The decision regions associated to SIC are hyper-rectangular and it is not difficult to perceive (e.g., from Fig. 7, but particularly in Fig. 8) that these decision regions are unequivocally defined by the Gram-Schmidt vectors of the basis of the lattice (e.g.,[28]).

Once a decision is produced for one layer, the subsequent step is to repeat the process, now in the sublattice with basis $\mathbf{H}_{\bar{j}}$, i.e., by removing generator $\mathbf{h}_j$ from the set. The process repeats itself until a decision is made in a one-dimensional lattice, corresponding to the decision of the last layer to be detected.

Fig. 7 depicts SIC applied to a lattice partitioned as in (25). In a first stage the nearest hyperplane is found and a decision for the layer associated to $\mathbf{h}_j$ is produced. In a second stage, depicted at the bottom of Fig. 7, the same procedure is applied but now conducted in the sublattice Λ_{n-1}.

$$\mathbf{H} = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}.$$

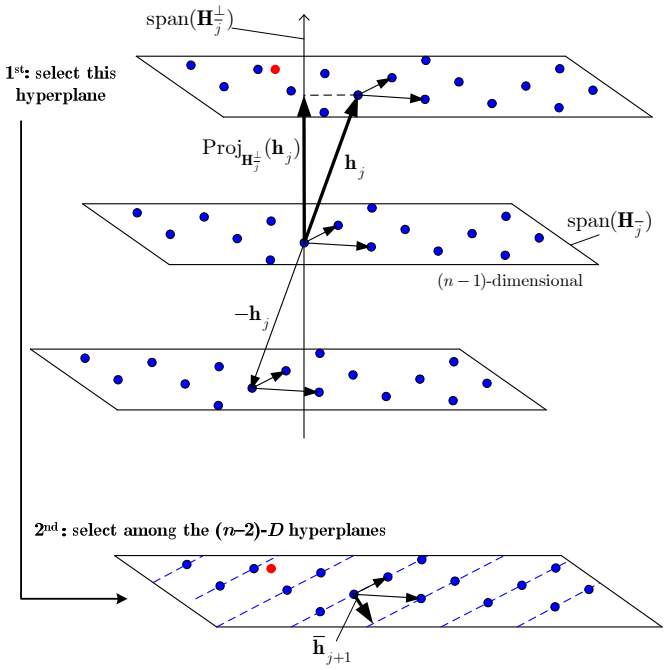


Fig. 7: The nearest plane algorithm with sorting. Choosing the j^{th} generator vector that maximises the distance between parallel hyperplanes. The lattice is the union of such translates.

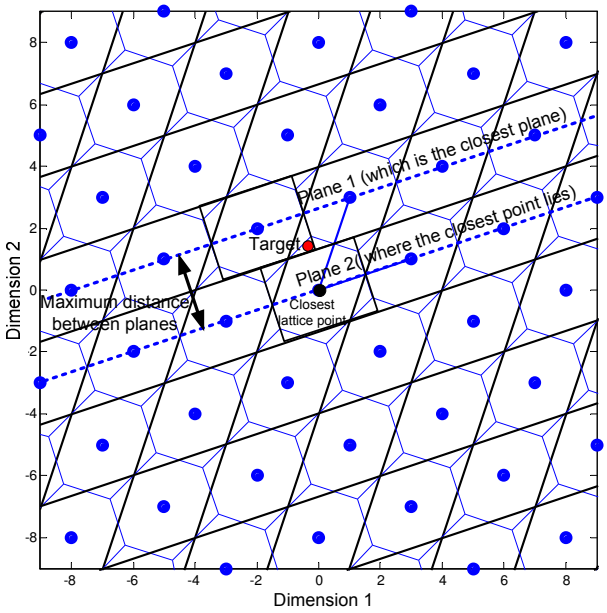


Fig. 8: Errors events in SIC. Plane 1 is selected because it is the closest plane, however, the closest lattice point lies in plane 2. The SIC decision region for the origin is shown.

Fig. 8 shows the SIC decision region for the origin of the lattice with basis. The example shows a target point located in a region where SIC outputs an erroneous decision. The first SIC step in the example in Fig. 8 selects the plane which is the nearest one to the target point. In the example, SIC would decide for plane 1 while the Voronoi region indicates that the correct point lies in plane 2.

The fastest implementation of the original OSIC idea was provided in [37] and was made cubic in n, i.e., with complexity $\mathcal{O}(n^3)$, nevertheless other $\mathcal{O}(n^3)$ algorithms were known for OSIC much before (c.f. [38] (p.39) and references therein). Ling et al. also proposed an OSIC algorithm with $\mathcal{O}(n^3)$ complexity based on the geometric insights offered by the dual lattice [21]. In doing that, the same optimal ordering known for OSIC [32] is proven and the same performance is attained without needing a matrix inversion for each layer to be detected. This approach makes use of the shortest vector in the dual basis at each detection step.

One can now formalise SIC in a very concise manner; the k^{th} index for the permutation is then selected from

$$\Pi(k) = \arg\max_{j \in \mathcal{A}_k} \left\{ \left\langle \mathbf{h}_j^T, \text{Proj}_{\mathbf{H}_{\bar{j}}^{\perp}}(\mathbf{h}_j) \right\rangle \right\}. \tag{26}$$

where $\mathcal{A}_k$ is the set of columns that have not been chosen yet.

Consider now a linear space spanned by the columns of $\mathbf{H}$, i.e., $\text{span}(\mathbf{H})$. The projection of a vector $\mathbf{a}$ onto that space is denoted as $\text{Proj}_{\mathbf{H}}(\mathbf{a})$, the projection onto the space orthogonal to $\text{span}(\mathbf{H})$ is denoted as $\text{Proj}_{\mathbf{H}^{\perp}}(\mathbf{a})$, and they are given by [26] (sec. 8.3)

$$\text{Proj}_{\mathbf{H}}(\mathbf{a}) = \mathbf{P}_{\mathbf{H}}\mathbf{a} = \mathbf{H}\mathbf{H}^{+}\mathbf{a}, \tag{27}$$

$$\text{Proj}_{\mathbf{H}^{\perp}}(\mathbf{a}) = \mathbf{P}_{\mathbf{H}^{\perp}}\mathbf{a} = (\mathbf{I} - \mathbf{H}\mathbf{H}^{+})\mathbf{a}. \tag{28}$$

From (27) and (28), the projection onto the $\text{span}(\mathbf{H}_{\bar{j}})$ is

$$\mathbf{P}_{\mathbf{H}_{\bar{j}}} = \mathbf{H}_{\bar{j}}\mathbf{H}_{\bar{j}}^{+} \tag{29}$$

and the projection onto its orthogonal complement is

$$\begin{aligned} \mathbf{P}_{\mathbf{H}_{\bar{j}}^{\perp}} &= \mathbf{I} - \mathbf{H}_{\bar{j}}\mathbf{H}_{\bar{j}}^{+} \\ &= \mathbf{I} - \mathbf{H}_{\bar{j}}\left(\mathbf{H}_{\bar{j}}^{H}\mathbf{H}_{\bar{j}}\right)^{-1}\mathbf{H}_{\bar{j}}^{H} \end{aligned} \tag{30}$$

where the Moore-Penrose pseudo-inverse was used on the last line.

Using these assisting matrices (29) and (30), one can write the layer ordering problem in the following maner:

$$\Pi(k) = \arg\max_{j \in \mathcal{A}_k} \left\{ h_j^T \left(\mathbf{I} - \mathbf{H}_{k,j}^T (\mathbf{H}_{k,j}^T \mathbf{H}_{k,j})^{-1} \mathbf{H}_{k,j}^T \right) h_j \right\}, \tag{31}$$

which is a very concise expression that summarises the entire OSIC with optimal ordering [39]. Starting with $A_1 = \{1, 2, \cdots, n\}$ (i.e., with all the columns of $\mathbf{H}$), the set A_k is reduced by one element each time a column is selected, and continues until only one is left. Although concise, this formulation for finding the permutation $\Pi(k)$ does not lead to a practical implementation.

There is however a very elegant way of finding $\Pi(k)$ remembering that the distance between hyperplanes in the primal lattice is established by the lattice points in the dual (as proved in Section III-B). Selecting the smallest basis vector in the dual basis ensures that the decision for that layer will be made from selecting between the most distant hyperplanes associated to that basis. Nonetheless, it is important to highlight that these are *not* necessarily the most distant hyperplanes in the lattice, because the search depends on the basis that was given in the first place. This observation confirms why there is room for improving a receiver based on the OSIC principle and why lattice-reduction-based receivers can improve this performance. It is thus natural to look for short vectors in the dual lattice other than the generators constituting the basis. Shorter vectors in the dual lattice would maximise the distance between the parallel hyperplanes and thus minimise erroneous decisions. Finding shorter vectors in the dual lattice is accomplished by means of lattice-reduction-aided (LRA) techniques, which will be presented next.

It is noteworthy that the geometric interpretation presented in this section also sheds light onto the finding by Taherzadeh et al. that reducing the dual matrix is preferable to reducing the primal basis [40].

C. Lattice-Reduction-Aided Detection

As mentioned in Section III-A, two different bases can generate the same lattice although their fundamental regions have different coverage. In order to maximise the coverage of the MLD region, one is interested in bases with vectors that are both short and close to orthogonal, which is called a *reduced basis*.Lattice reduction provides an equivalent basis with shorter (and more orthogonal) generator vectors. Fig. 9 depicts a lattice with a rather "skewed" basis and a reduced basis.

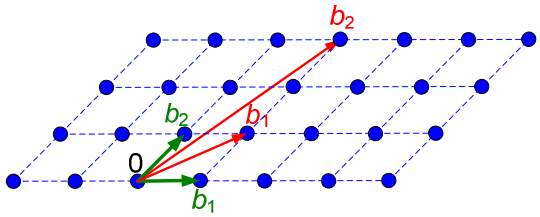


Fig. 9: A reduced basis (shorter) and a skewed basis (larger) for the same lattice.

Two bases of a lattice are related by a unimodular transformation $\mathbf{M}$ (a matrix only with integer elements and with $\det(\mathbf{M}) = 1$). In particular, the two basis in Fig. 4 of Section IV-A-1) are related by

$$\mathbf{H}_1 = \mathbf{H}_2\mathbf{M} \Leftrightarrow \begin{bmatrix} 6 & 2 \\ 1 & 5 \end{bmatrix} = \begin{bmatrix} 6 & 8 \\ 1 & 6 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \tag{32}$$

And, in this case, it is easy to see that $\det(\mathbf{M}) = 1$.

When a ZF filter is used, the coverage of the Voronoi region can maximized if the basis is more orthogonal and with short vectors. For example, the basis in Fig. 4-a) was preferable to the one in Fig. 4-b). To make this possible, in lattice-reduction-aided (LRA) receivers a pre-processing stage is introduced before the detection algorithm, as shown in Fig. 10.

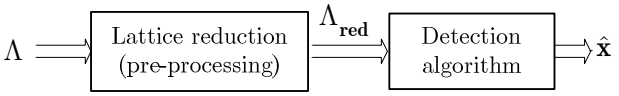


Fig. 10: MIMO detection with lattice-reduction pre-processing.

The application of lattice-reduction-aided (LRA) techniques to MIMO detection was pioneered by Yao and Wornell in 2002 [41], and since then the research in LRA applications to different MIMO contexts (such as the broadcast channel, important in LTE) has boomed. These authors applied the Lenstra Lenstra Lovász reduction (LLL, also sometimes denoted as L^3) [42] to reduce the channel matrix. An overview of the applications of lattice reduction techniques in MIMO (including SM and BC) exists in [43]. LRA detection achieves the maximum diversity available in SM, as proved by Taherzadeh et al. [40] for the case of LLL.

The idea is that the system model can be re-written as

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n} \Leftrightarrow \underbrace{\mathbf{H}\mathbf{M}}_{\mathbf{H}_{\text{red}}} \underbrace{\mathbf{M}^{-1}\mathbf{x}}_{\mathbf{z}} + \mathbf{n} \Leftrightarrow \mathbf{H}_{\text{red}}\mathbf{z} + \mathbf{n} \qquad (33)$$

In this model, $\mathbf{z}$ is a modified data vector that can be detected with a lower SER than would $\mathbf{x}$ without LR. This is true regardless the type of receiver that follows the LR pre-processing (usually ZF, MMSE or OSIC). The original data vector can then be recovered from $\mathbf{z}$ noting that

$$\mathbf{z} = \mathbf{M}^{-1}\mathbf{x} \quad \Rightarrow \quad \mathbf{x} = \mathbf{M}\mathbf{z} \qquad (34)$$

Because *M*-QAM constellations and their PAM equivalent alphabets are defined without the origin and have non unitary distance between the symbols (c.f. (2)), in order to apply the lattice tools as in (33)-(34), it is necessary to make a translation of the constellation, creating the modified received vector

$$\mathbf{y}_{\text{red}} = \frac{1}{2}\left(\mathbf{y} + \mathbf{H}\cdot\mathbf{1}\right) = \frac{1}{2}\left(\mathbf{H}\mathbf{x} + \mathbf{n} + \mathbf{H}\cdot\mathbf{1}\right), \qquad (35)$$

where $\mathbf{1}$ is the column vector of n elements all equal to 1.

Now, in the case of a ZF criterion,

$$\mathbf{z} = \mathbf{H}_{\text{red}}^{+}\mathbf{y}_{\text{red}}, \qquad (36)$$

and in performing

$$\hat{\mathbf{x}} = 2\mathbf{M}\underbrace{Q_{\mathbb{Z}}(\mathbf{z})}_{\hat{\mathbf{z}}} - \mathbf{1}, \qquad (37)$$

the symbol $\hat{\mathbf{z}}$ is detected and put back in the alphabet $\mathcal{A}$.

D. Sphere decoding

Sphere decoding (SD) is an exact detection method (i.e., it achieves the same performance as MLD) with a complexity that, on average, is much lower than MLD. The idea is that a rigid rotation $\mathbf{Q}$ can be applied to the ensemble $\{\Lambda, \mathbf{y}\}$, for which the CVP needs to be solved, so that the lattice can be described by an equivalent lattice in u.t form. The u.t. property allows describing the norm of any lattice point to be detected as a sum that can be computed incrementally, taking in consideration the cumulative effect of each vector components. Consider now that an upper bound (UB) on the norms is established. The u.t. property of the basis allows all the possible values in the last component of the data vector, $x(n)$, to be detected. As the norm can be computed as a sum of "ordered" contributions, if some of the tested values in $x(n)$ generates a total vector norm that is larger than the UB, then those values of $x(n)$ need not to be considered further as possible values in the solution. This procedure can extended to the next layer $n-1$, where only the possible values of $x(n)$ are considered. In conclusion, finding vectors with norm smaller then the UB is a problem that can be solved by expanding and pruning a tree that represents the lattice points. All these ideas can be converted to the CVP, if the lattice is shifted to the target $\mathbf{y}$.

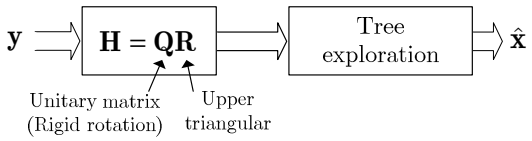


Fig. 11: Sphere decoder receiver structure.

A sphere decoder has the structure shown in Fig. 11. After traversing the tree with a particular symbol enumeration, the MLD solution is always found if the initial radius that is chosen is large enough to contain a lattice point inside the hypersphere. Fig. 12 gives an example tree associated with 3×3 antenna and 4-PAM, showing the branches that have been expanded at each tree level.

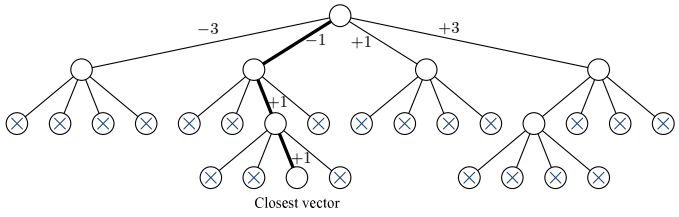


Fig. 12: Tree exploration of a tree with 3 layers, considering a 4-PAM alphabet.

As mentioned above, one can define an UB for the radius (or, equivalently, for the squared radius) of the sphere around the received point, i.e., $\left\|\mathbf{y} - \mathbf{H}\mathbf{x}\right\|^2 \leq \xi^2$. Finally, remembering

that $\mathbf{R}$ is upper triangular, for full rank $\mathbf{H}$ the problem can be written as the sum

$$\sum_{i=1}^{m}\left(y_i - \sum_{j=i}^{m} r_{ij}x(j)\right)^2 \leq \xi^2, \tag{38}$$

which is in the form of incremental contributions, thus allowing to be solved by a tree exploration as exemplified in Fig. 12.

The average complexity of SD is exponential, given by $\mathcal{O}(M^{\alpha N_T})$ with $0 \leq \alpha \leq 1$ [44], however, for low dimensional lattices, that number is affordable.

V. A Geometry-based Receiver

This section capitalises on the geometric relation between primal and dual lattices and proposes a dual-lattice-aided (DLA), for slow fading channels, that samples some points nearby the received vector and generates a list of candidate solutions to the CVP. This sampling makes use of projections onto distinct families of parallel hyperplanes where the density of lattice points is maximised.

A. *Successive Minima in the Dual Lattice*

The dual lattice, as is the case of any lattice, has a shortest vector (which comes at least paired with its symmetrical vector) and a set of other successive minima. Because some of these vectors may be linearly dependent, the interesting definition of successive minima imposes independence. Hence, λ_i is the i^{th} *successive minimum* of a lattice if λ_i is the smallest real number that is the smallest radius of a sphere that contains i independent vectors, all with norms smaller or equal to λ_i. The shortest vector obviously has norm λ_1.

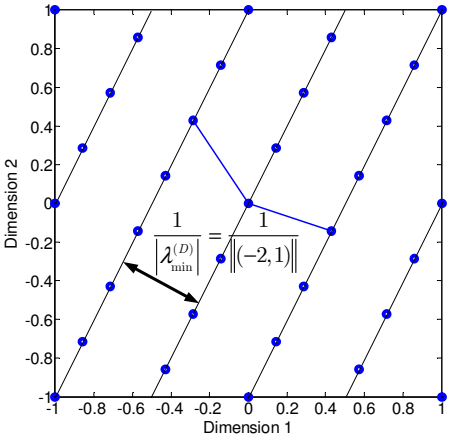
From Section III-B it is possible to conclude that the hyperplanes which are furthest apart from each other (and thus having the highest density of lattice points on them) are defined by the shortest vector in the dual lattice. This observation is essential to explain which layer must be detected first in a OSIC receiver. The selection of the next layer is determined by the same observation, applied now to the sublattice spanned by the matrix obtained after striking out from $\mathbf{H}$ the column generator preciously detected.

In what follows, this paper proposes that not only the family of hyperplanes that are furthest apart are used, but also that other families of hyperplanes associated with some of the successive minima of the dual lattice should also be brought to use. Fig. 13 (a) and (b) shows an example of two of those different partitions of a lattice defined by

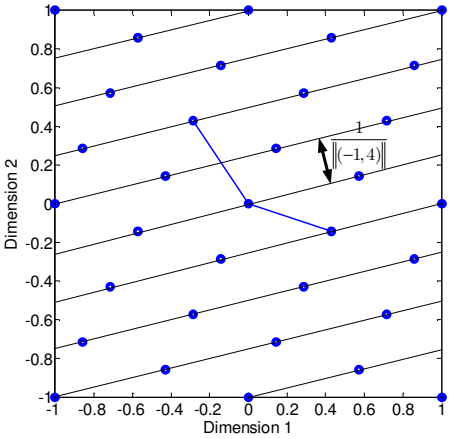
$$\mathbf{H} = \begin{bmatrix} 3/7 & -2/7 \\ -1/7 & 3/7 \end{bmatrix} \quad \text{and with} \quad \mathbf{H}^{(D)} = \begin{bmatrix} 3 & 1 \\ 2 & 3 \end{bmatrix},$$

associated with two different choices of vectors in $\Lambda^{(D)}$.

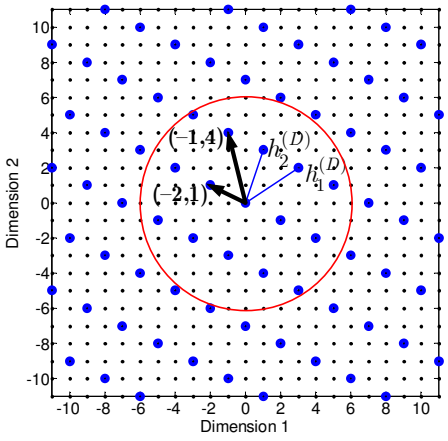
Consider the hyperplanes selected by the first L successive minima in $\Lambda^{(D)}$, i.e., $\lambda_i, \cdots \lambda_L$. Finding the shortest vector in a lattice is itself a NP-hard problem, which implies the same complexity for obtaining the L shortest ones. Nevertheless, if this is only required at a pre-processing stage, and not needed for each received vector, then using a sphere decoder is acceptable.



(a) Hyperplanes in the primal lattice associated with (–2,1) in the dual lattice.



(b) Hyperplanes in the primal lattice associated with (–1,4) in the dual lattice.



(c) Vectors selected in the dual lattice (black arrows).

Fig. 13: Identification of the hyperplanes in the primal lattice associated with a given vector in the dual lattice.

While its complexity is exponential in the dimension of the lattice [45], this cost is only necessary whenever the channel changes, which is tolerable for slow fading channels.

Using the SD principle, applied to the origin of the lattice, it is possible to list all the points of a lattice inside a hypersphere of radius ρ centred in the origin, which for this reason only requires a very simple implementation of the SD such as the one given in [45]. We are interested in the set of lattice points spanned by $\mathbf{H} \in \mathbb{R}^{m \times n}$ which verify

$$\left\{ \mathbf{y} \in \mathbb{R} : \mathbf{y} = \left\| \mathbf{H} \cdot \mathbf{x} \right\|^2 \le \xi^2, \quad \mathbf{x} \in \mathbb{Z} \right\}. \tag{39}$$

which can be written in the form (38), permitting the characteristic tree exploration of SD. Because one is interested in planes with different distances, not all the lattice points with $\left\| \mathbf{y} \right\| < \xi$ are a successive minima and they need to be expunged from the list.

When centred at the origin, an implementation of SD such as that in [45], outputs a list of (column) vectors arranged as

$$[\underbrace{\mathbf{s}_1, \mathbf{s}_2, \cdots \mathbf{s}_{N/2-1}}_{N/2-1}, \mathbf{0}, \underbrace{\mathbf{s}_{1N/2+1}, \cdots \mathbf{s}_N}_{N/2-1}]$$

where $\mathbf{0}$ is the origin, which is always captured in the set for any $\rho > 0$ and N is the number of lattice points inside the sphere of radius ξ.

The two sides of the output around 0 have the same vectors up to their sign and therefore the selection of the first $N/2-1$ suffices. In addition to that selection, one will just take one vector for each distinctive norm, even if there are several linearly independent ones. This widens the range of different distances between hyperplanes. The resulting set of vectors in the dual will be dubbed *unique successive minima* (USM). This concept is depicted in Fig. 13 (c), where L=7 USM are found inside the sphere in $\Lambda^{(D)}$.

B. Projections onto Hyperplanes

The L USM in the dual lattice are denoted by $\mathbf{v}_1^{(D)}, \mathbf{v}_2^{(D)}, \cdots, \mathbf{v}_L^{(D)}$. Naturally, the unit vectors which are orthogonal to the families of hyperplanes are

$$\overline{\mathbf{v}}_i^{(D)} = \mathbf{v}_i^{(D)} \Big/ \left\| \mathbf{v}_i^{(D)} \right\| \tag{40}$$

and we further define the vectors $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_L$, each one respectively collinear with $\mathbf{v}_i^{(D)}$, but forced to have norm $d = \left\| \mathbf{v}_i^{(D)} \right\|^{-1}$, as has been suggested in Fig. 1. Hence, from (40), these vectors should be $\mathbf{v}_i = \mathbf{v}_i^{(D)} \Big/ \left\| \mathbf{v}_i^{(D)} \right\|^2$.

The projections of the received vector (i.e., the target in the CVP) onto a family of $\mathcal{P}_{\mathbf{v}_i}(\nu)$ hyperplanes generate the set of projection points

$$\begin{aligned} \mathbf{y}_p(\mathbf{v}_i, \nu) &= \mathbf{y} + \underbrace{\left[Q_{\mathbb{Z}} \left(\frac{\langle \mathbf{y}, \mathbf{v}_i \rangle}{\left\| \mathbf{v}_i \right\|^2} \right) - \frac{\langle \mathbf{y}, \mathbf{v}_i \rangle}{\left\| \mathbf{v}_i \right\|^2} \right]}_{\Omega} \mathbf{v}_i + \nu \, \mathbf{v}_i \\ &= \mathbf{y} + \left(\Omega(\mathbf{y}, \mathbf{v}_i) + \nu \right) \mathbf{v}_i \end{aligned} \tag{41}$$

where $\nu \in \mathbb{Z}$ and $Q_{\mathbb{Z}}(\cdot)$ denotes rounding to $\mathbb{Z}$. Notice that, in the case of a zero noise vector, the index Ω will be *always* an integer, indicating in which hyperplane the lattice lies, for each family $\mathcal{P}_{\mathbf{v}_i}$.

C. List of Candidate Solutions

Fixing L as the number of USM, and setting $\nu_{\max}$ as the maximum value of ν that will be explored, it is possible to obtain a set C consisting of the candidate vectors obtained from

$$\mathbf{y}_i^{(C)} = \mathbf{H} Q_{\mathcal{A}} \left(\mathbf{H}^{+} \mathbf{y}_p(\overline{\mathbf{v}}_i, \nu) \right) \quad , i = 1, 2, \cdots, 2L\nu_{\max}, \tag{42}$$

The total number of candidates considered in (44) is given by the number of families of hyperplanes considered (i.e., the number of USM inside a sphere) multiplied by the number parallel of hyperplanes considered (the closest one and the adjacent ones with non-zero index ν):

$$| C | = L \cdot (2 \cdot \nu_{\max} + 1). \tag{43}$$

This amounts to performing zero-forcing detection not only to $\mathbf{y}$ but also to the set of all projections onto $\mathcal{P}_{\mathbf{v}_i}(\nu)$.

This concept is depicted in Fig. 14, which shows the projections of the target point onto the three densest families of hyperplanes with lattice points. The projections are made through the dark black segments and, for this example with $\nu = -1, 0, +1$, there are three projections along each black segment onto the corresponding three nearest hyperplanes.

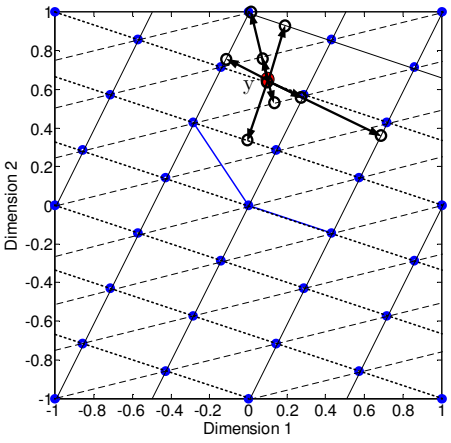


Fig. 14: Dual-lattice-aided generation of candidate solutions considering $\nu_{\max}$=1 and considering L=3 families of hyperplanes: the two families in Fig. 13 and also the family associated with the dual vector $h_2^{(D)} = (1, 3)$.

The circle lines in Fig. 14 correspond to the nine projections that are generated, i.e., three projections in each family of hyperplanes.

The solution to the CVP is then obtained by applying the maximum likelihood principle to all the vectors in the set of candidates *C*

$$\hat{\mathbf{y}} = \arg\min_{\mathbf{y}_i^{(C)} \in C} \left\{ \left\| \mathbf{y} - \mathbf{y}_i^{(C)} \right\|^2 \right\}. \tag{44}$$

Table 1 shows the number of candidates, $|C|$, generated when the number of USM is set to $L = 4n$ and the maximum index of the hyperplanes is $\nu_{\max} = 2$ and $\nu_{\max} = 4$.

Table 1: Number of candidates in DLA.

	$n=4$	$n=6$	$n=8$	$n=12$
$L=4n$, $\nu_{\max}=2$	80	120	160	240
$L=4n$, $\nu_{\max}=4$	144	216	288	432

D. Simulation Results

The performance of the proposed dual-lattice-aided (DLA) receiver is assessed in terms of the (complex) symbol error rate (SER) with *L*=4*n*, for 3×3 and 4×4 antennas (i.e., lattices with *n*=6 and *n*=8 dimensions, i.e., with *L*=24 and *L*=32 respectively) using 64-QAM. The results are shown in Fig. 15 and Fig. 16. These figures also include the following traditional receivers: ZF, MMSE, OSIC, LRA using LLL pre-processing with ZF and also with OSIC-ZF, besides ML (using SD). In both cases OSIC is outperformed.

One can observe that DLA detection outperforms OSIC using a reasonable number of candidates, $|C|$, obtained from the projections (41) and listed in Table 1. At $\mathrm{SER} = 10^{-3}$, the gain in respect to OSIC amounts to 5 dB with 3×3 antennas, and is 4dB with 4×4 antennas.

The DLA receiver exhibits better performance than LRA in the low SNR regime, but because LRA achieves the full diversity of the channel [46] the SER of LRA eventually drops below the one of the proposed algorithm. It has also been found that, as expected, when either *L* or $\nu_{\max}$ decreases, the performance degrades.

The complexity involved in DLA is concentrated in the pre-processing stage that is only required each time the CSI is updated at the receiver. This involves solving a SVP via SD, which has complexity $\mathcal{O}(M^{\alpha N_T})$, as mentioned in Section IV-D. However, once the USM are determined, DLA detection amounts to first computing the projections according to (41), and perform afterwards the matrix multiplications in (42), involving the Moore-Penrose inverse of the channel. Therefore, given that matrix multiplications have a number of operations $\mathcal{O}(n^3)$ (which is the same as the complexity to compute the pseudo-inverse), the total number of operations is then bound by $2L\nu_{\max}\mathcal{O}(n^3)$, with $n = 2N_T$, because there are $2L\nu_{\max}$ candidates in (42).

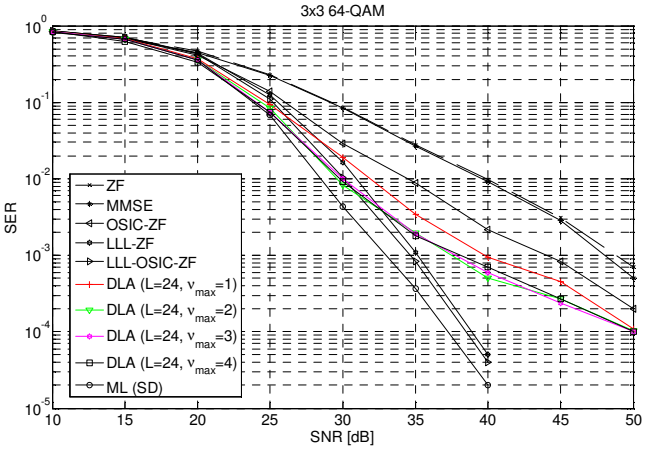


Fig. 15: SER vs SNR for 3×3 antennas and 64-QAM.

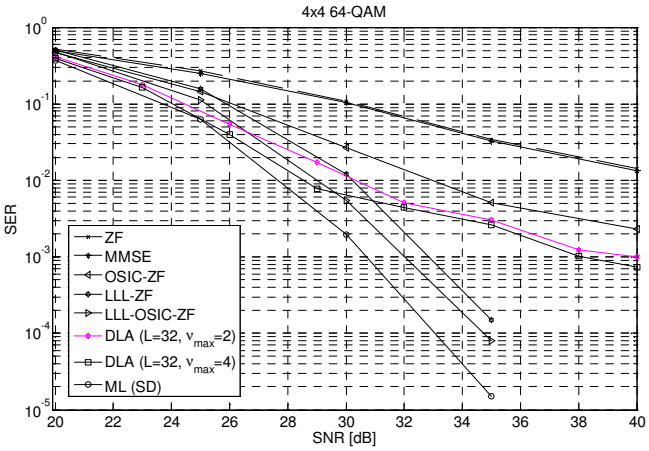


Fig. 16: SER vs SNR for 4×4 antennas and 64-QAM.

VI. TRELLIS-BASED RECEIVERS

This section presents a novel concept for SM receivers which by converting the problem into the search of a minimum cost path in a trellis, which can be solved by the Viterbi algorithm (a well known technique to radio communication engineers).

A. Trellis representation of lattices

The regularity of a lattice lends itself for the representation of problems where signals are interpreted as a point in a multidimensional space defined in some basis. Forney's pioneering work [47] showed that some lattices can be described by a trellis, where each segment of the trellis is associated with the coordinates of the lattice points in each dimension of the space. The lattices for which a trellis exists, can be said to constitute a *family* of lattices, denoted by $\mathcal{L}_R$. These properties have been used in coding theory for detecting

lattice codes [48]. However, this approach requires a rather restricted type of lattice allowing a trellis representation. Some well known lattices belong to $\mathcal{L}_R$ (such as A_2, D_4, E_8, or the Leech lattice in $\mathbb{R}^{24}$) [49] but others are constructed imposing a specific geometrical structure during the design of the code. MLD can in those cases be attained through trellis detection, and therefore the CVP is in those cases solved with the Viterbi algorithm. This circumvents the exponential complexity of MLD.

Clearly, the trellis detection approach cannot be extended to any random lattice. However, one should ask the question, for any given lattice, can one find a lattice that is sufficiently "similar" or "close" to it, and yet is simultaneously a member of the family of lattices with a trellis representation, $\mathcal{L}_R$. This paper deals with that question. As lattices are defined by generator matrices, the problem can be seen as a *matrix nearness problem* [50]; as in many other matrix nearness problems, the one we formulate also does not seem to have an analytical solution and therefore we take an algorithmic approach. To the best of our knowledge the approximation of a random lattice by a lattice in $\mathcal{L}_R$ is a new approach to MIMO detection. In [51] the authors use a trellis detector but their approach is clearly sub-optimal, as it is based on a transformation of a tree data structure (associated with a sphere decoder) into a trellis data structure, and ends up losing many of the branches.

In this section one derives the property that makes a lattice a member of $\mathcal{L}_R$ and presents an algorithm to find such a lattice which is "nearby" a given random lattice. Then the paper presents results for typical detection in MIMO SM, comparing the results with the most important sub-optimal receivers and MLD.

B. Focusing Onto the $\mathcal{L}_R$ Family

We call $\mathcal{M}$ the set of all possible lattices in $\mathbb{R}^n$. Hence, $\mathbb{Z}^n$ is just one particular lattice in $\mathcal{M}$ (see Fig. 18). Moreover, all lattices with a trellis representation are also members of that $\mathcal{M}$ and we say that they constitute the $\mathcal{L}_R$ family of lattices.

It is well known that the simplest way of solving the CVP amounts to the least-squares solution given by the Moore-Penrose pseudo-inverse of the generator matrix. In the MIMO context this is known as the zero-forcing (ZF) solution. Geometrically, this type of linear receiver applies a linear transformation that takes the received lattice Λ and transforms it back into the original $\mathbb{Z}^n$. We will call this procedure a *focusing* of the received lattice Λ onto $\mathbb{Z}^n$, and we propose to generalize this concept of focusing by means of a linear transformation $\mathcal{F}$. The ZF focusing approach presents the lowest complexity among all sub-optimal receivers but also results in the poorest performance (in terms of erroneous decisions). The poor performance is a direct consequence of the potentially huge mismatch between the optimal decision regions in MLD and the decision regions associated with focusing onto $\mathbb{Z}^n$. These decision regions are nothing but linear transformations of *n*-dimensional hypercubes. Note that the convenience of the ZF receiver comes from the fact that the destination lattice is $\mathbb{Z}^n$, which allows detection by means of a simple slicer.

One argues that it is possible to perform a linear transformation from any received lattice Λ onto other lattices in $\mathcal{M}$ which also lend themselves to another convenient detection method, namely, the Viterbi algorithm. This concept gives rise to a new kind of MIMO receiver with the steps illustrated in Fig. 17.

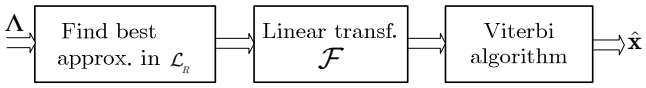


Fig. 17: Detection on an approximated trellis representation.

Fig. 18 depicts the set of all lattices, including the particular $\mathcal{L}_R$ family. Any given lattice may be closer to one lattice in $\mathcal{L}_R$ than to $\mathbb{Z}^n$, as those are infinitely many more. Again, notice that ZF would focus any received lattice always onto $\mathbb{Z}^n$, regardless the distance to it.

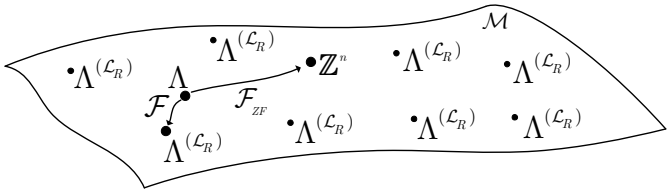


Fig. 18: The set of lattices and the focusing operator. A received lattice Λ can be focused onto the nearest member of $\mathcal{L}_R$ or onto $\mathbb{Z}^n$.

When the distance between lattices is reduced, then the matching (or coverage) between their decision regions is maximized, which minimizes the distortion created by linearly transforming one lattice onto another one. If there is a member of $\mathcal{L}_R$ nearby Λ (i.e. very "similar" to Λ), then *i)* its MLD regions will mach closely the ones of the original lattice and *ii)* the distortion involved in the focusing operation will be small.

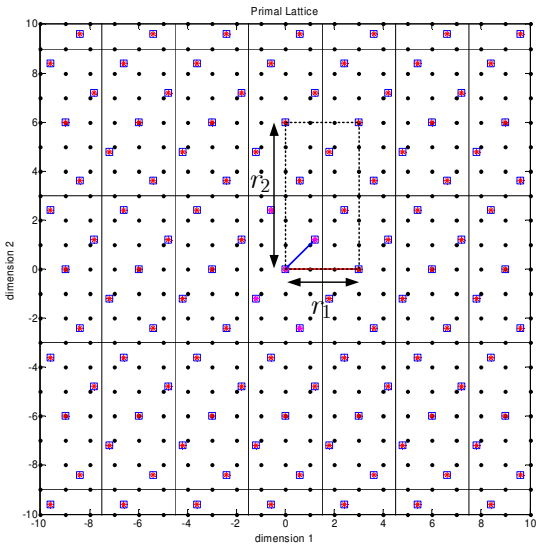
C. The $\mathcal{L}_R$ Family of Lattices

A lattice has a trellis if it can be written as the union of a *rectangular sublattice* Λ_R and translated versions of it. As noticed by Forney [47], such a lattice is given by $\Lambda = \Lambda_R + [\Lambda/\Lambda_R]$, where $[\Lambda/\Lambda_R]$ is a "system of coset representatives" for the cosets of Λ_R in Λ or, equivalently, for the elements of the quotient group Λ/Λ_R. As Λ_R is a rectangular lattice, by definition it can be expressed by a Cartesian product, i.e., $\Lambda_R = r_1\mathbb{Z} \times \cdots \times r_n\mathbb{Z}$.

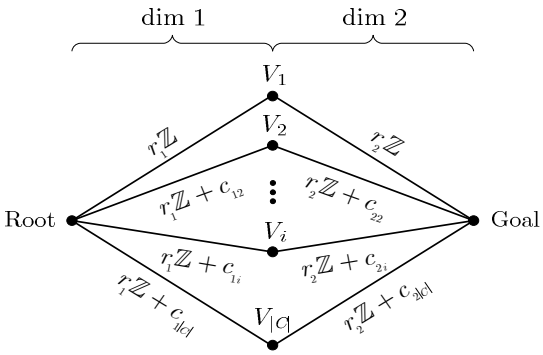
Fig. 19 shows an example of a lattice in $\mathbb{Z}^2$ and its representation by a trellis. It is possible to observe the rectangular quotient group and its translated versions.The

lattice is then the union of the cosets of Λ_R in Λ . For the case in Fig. 19 , the index of Λ_R in Λ is $|C|=|\Lambda/\Lambda_R|=5$. In general,

$$\left|C\right| = \frac{\det(\Lambda_R)}{\det(\Lambda)}. \tag{45}$$



(a) Rectangular sub-lattice in a lattice that has a trellis representation.



(b) Trellis of the 2D lattice.

Fig. 19: A rectangular sub-lattice in a random lattice and the trellis representation of the lattice.

Using the origin as a representative of Λ_R, the set constituted by the origin together with all the other points with coordinates (c_{1i}, c_{2i}), i=1,2....|C|, that are inside the central rectangular region are the *coset representatives* of the quotient group. The whole lattice can now be seen as a tiling of the space using that fundamental region. The coefficients $r_1, r_2, \cdots r_n$ have now a simple geometrical interpretation as they define the lengths of the fundamental hyper-rectangle.

There is a strong connection between the way the trellis of binary block codes and group codes are obtained from a trellis-oriented generator matrix [48], [52] , and how the trellis of a lattice is obtained from the basis of a lattice in $\mathcal{L}_R$, [47], [53]. The n-dimensional orthogonal sublattice has its basis vectors along one-dimensional subspaces W_i, i=1,...,n. From these we can define the sequence of spaces $\{\mathbf{0}\} \subset V_0 \subset V_1 \subset \cdots \subset V_n = \mathbb{R}^n$ and each W_i, is the 1-D orthogonal complement of V_{i-1} to V_i. We denote the projections onto V_i and W_i respectively by P_i and the P_{W_i} and define the intersection lattices $\Lambda_i = \Lambda \cap V_i$ and the one-dimensional lattices $\Lambda_{W_i} = \Lambda \cap W_i$.

Using these definitions, the state space of a trellis of a lattice in the coordinate system $\{W_i\}_{i=1}^n$ is $P_i(\Lambda)/\Lambda_1$ and the label group for the trellis branches is $P_{W_i}(\Lambda)/\Lambda_{W_i}$ [54], [55] [56].

D. Orthogonal Sublattices

We are interested in finding what properties a generator matrix must have so that it generates a lattice $\Lambda \in \mathcal{L}_R$.

1) Properties of the generator matrix

A lattice can only be written as in the form $\Lambda = \Lambda_R + [\Lambda/\Lambda_R]$ if and only if it contains a rectangular sublattice. Given a lattice, to find if a rectangular sublattice in it is believed to be itself an NP-hard problem. Micciancio calls it the *quasi orthogonal set problem* [17] that we may appropriately call it the *quasi orthogonal sublattice problem* (QOSP). This problem deserved virtually no attention in the literature, apparently due to lack of applications.

In addition to the problem of discovering a rectangular sublattice we add an additional constraint: we want to find the rectangular sublattice that minimizes the index number of the quotient group in order to minimize the number of trellis paths. The problem does not seem to have an analytical solution; consequently, we revert to an algorithmic approach.

Let us consider a random rational lattice defined by a rational $\mathbf{H}$ with entries $h_{ij}=n_{ij}/d_{ij}$ and whose inverse is the rational matrix $\mathbf{W}=\mathbf{H}^{-1}$ with entries p_{ij}/q_{ij}. For lattice points, because $\mathbf{x} = \mathbf{H}^{-1}\mathbf{y}$ (no noise), one should force

$$\begin{bmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \frac{p_{11}}{q_{11}} & \frac{p_{12}}{q_{12}} & \cdots & \frac{p_{1n}}{q_{1n}} \\ \frac{p_{21}}{q_{21}} & \frac{p_{22}}{q_{22}} & \cdots & \frac{p_{2n}}{q_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{p_{n1}}{q_{n1}} & \frac{p_{n2}}{q_{n2}} & \cdots & \frac{p_{nn}}{q_{nn}} \end{bmatrix} \cdot \begin{bmatrix} kr_1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{p_{11}}{q_{11}} kr_1 \\ \frac{p_{21}}{q_{21}} kr_1 \\ \vdots \\ \frac{p_{n1}}{q_{n1}} kr_1 \end{bmatrix} \tag{46}$$

for $k \in \mathbb{Z}$. As $\mathbf{x} \in \mathbb{Z}$, then $\frac{p_{i1}}{q_{i1}} kr_1 \in \mathbb{Z} \Rightarrow \frac{r_1}{q_{i1}} \in \mathbb{Z}$ and thus $q_{i1} \left| r_1 \right.$, where $q_{i1} \left| r \right.$ denotes that q_{i1} divides r . Hence,

$$r_1 = \operatorname{lcm}\left(q_{11}, q_{21}, \cdots, q_{n1}\right), \tag{47}$$

where lcm stands for *lowest common multiple*. Applying the same reasoning to each dimension one gets the rule $r_i = \mathrm{lcm}\left(q_{1i}, q_{2i}, \cdots, q_{ni}\right)$. Finally, we can interpret the property in terms of the columns of $\mathbf{H}^{(D)}$, the generator matrix of the dual lattice (henceforth called the *dual matrix*). In conclusion, the sublattice Λ_R of $\Lambda \in \mathcal{L}_R$ in the original system of coordinates is completely defined by the denominators q_{ij} of the dual lattice, so that

$$r_i = \mathrm{lcm}\left(q_{i1}, q_{i2}, \cdots, q_{in}\right),\ i = 1, 2, \ldots n\,. \tag{48}$$

2) Geometrical interpretation: distortion vs number of cosets

The number of cosets in a quotient group is

$$\Phi = \left|\frac{\tilde{\Lambda}}{\Lambda_R}\right| = \frac{\mathrm{Vol}\left(\Lambda_R\right)}{\mathrm{Vol}\left(\tilde{\Lambda}\right)} = \frac{\prod_{i=1}^{n} r_i}{\det\left(\tilde{\mathbf{H}}\right)} = \prod_{i=1}^{n} r_i \cdot \det\left(\tilde{\mathbf{H}}_D\right). \tag{49}$$

In order to calculate $\det(\tilde{\mathbf{H}}_D)$ one should observe that (55) is uniquely defined by two matrices: one is $\mathbf{P}$, comprising the denominators of $\tilde{\mathbf{H}}_D$, and the other we call $\mathbf{R}$, with the numerators of $\tilde{\mathbf{H}}_D$, and both matrices are u.t.:

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ 0 & p_{22} & \cdots & p_{21} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p_{nn} \end{bmatrix} \text{ and } \mathbf{R} = \begin{bmatrix} r_1 & r_1 & \cdots & r_1 \\ 0 & r_2 & \cdots & r_2 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & q_n \end{bmatrix}. \tag{50}$$

Note that the non-zero elements of $\mathbf{R}$ in each row are forced to be equal. The determinant of $\tilde{\mathbf{H}}_D$ is then

$$\det\left(\tilde{\mathbf{H}}_D\right) = \prod_{i=1}^{n} \frac{p_{ii}}{r_i} = \underbrace{\prod_{i=1}^{n} p_{ii}}_{\substack{\text{product diagonal} \\ \text{numerators}}} \cdot \underbrace{\prod_{i=1}^{n} \frac{1}{r_i}}_{\substack{\text{volume of} \\ \text{quantization grid}}} \tag{51}$$

and (49) becomes

$$\Phi = \prod_{i=1}^{n} r_i \cdot \prod_{i=1}^{n} p_i \cdot \prod_{i=1}^{n} \frac{1}{r_i} = \prod_{i=1}^{n} p_i\,. \tag{52}$$

The number of cosets is thus solely determined by $\mathrm{diag}(\mathbf{P})$. Geometrical insight into the problem can now be given from:

$$\left|\frac{\Lambda}{\Lambda_R}\right| = \frac{\mathrm{Vol}\left(\Lambda_R\right)}{\mathrm{Vol}\left(\tilde{\Lambda}\right)} = \mathrm{Vol}\left(\Lambda_R\right)\mathrm{Vol}\left(\tilde{\Lambda}_D\right)$$

$$= \prod_{i=1}^{n} r_i \cdot \mathrm{Vol}\left(\tilde{\Lambda}_D\right) = \mathrm{Vol}\left(\tilde{\Lambda}_D\right) \Big/ \frac{1}{\prod_{i=1}^{n} r_i} = \frac{\mathrm{Vol}\left(\tilde{\Lambda}_D\right)}{\mathrm{Vol}(\varepsilon_q)}. \tag{53}$$

The denominator $(\prod_{i=1}^{n} r_i)^{-1}$ corresponds to the volume of the elementary quantization grid (see Fig. 20).

In order to reduce the number of paths in the trellis of a lattice in $\mathcal{L}_R$, one wants to keep low value entries in $\mathrm{diag}(\mathbf{P})$, while at the same time, a good approximation that minimizes $\left\|\mathbf{H}^{(D)} - \tilde{\mathbf{H}}^{(D)}\right\|_F$, implies having larger r_i values (as these ratios are fixed, this constitutes another constraint into the problem). For that purposes one presents Algorithm 1, which outputs $\tilde{\Lambda}(\tilde{\mathbf{H}})$. By construction, the shape of the Voronoi regions of this latter lattice are similar to the ones of the original Λ. Using the concept introduced in Section VI-B, the focusing linear transformation is

$$\mathcal{F}(\mathbf{H}, \tilde{\mathbf{H}}) = \tilde{\mathbf{H}} \cdot \mathbf{H}^{-1}, \tag{54}$$

with $\mathcal{F}$ close to the identity, i.e., $\left\|\mathcal{F} - \mathbf{I}\right\|_F < \varepsilon$. By allowing an increasing number of cosets, ε can be reduced towards zero.

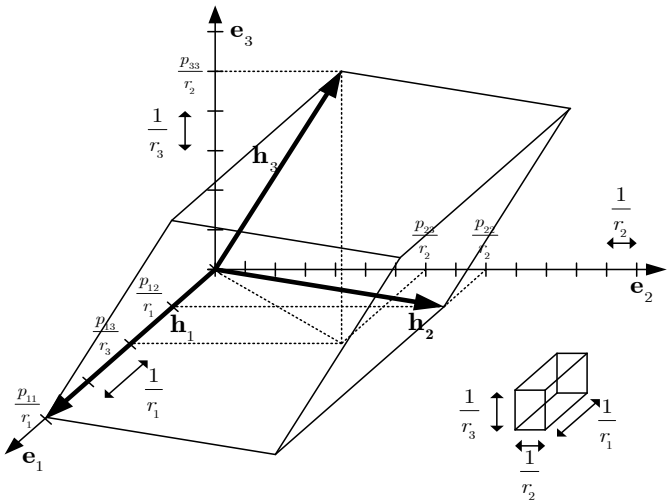


Fig. 20: Approximation versus number of cosets: the dilemma of the approximation in the dual lattice (example in a 3D space).

3) Algorithm

Given the condition (48) for the dual matrix, one reduces the problem to finding a close dual generator matrix (in the Frobenius sense). One starts by applying a QR decomposition to the dual matrix, reducing it to an upper triangular (u.t.) form via a unitary rigid rotation of the lattice, $\mathbf{Q}$. To make the elements in this matrix shorter, one has to: *i)* LLL-reduce this rotated dual lattice, and then *ii)* find rational approximations for the matrix elements via a greedy algorithm. The algorithm finds an approximated (or *synthetic*) dual lattice defined by

$$\tilde{\mathbf{H}}^{(D)} = \left(\mathbf{h}_1^{(D)}\ \mathbf{h}_2^{(D)} \cdots \mathbf{h}_n^{(D)}\right) = \begin{pmatrix} \frac{p_{11}}{r_1} & \frac{p_{12}}{r_1} & \cdots & \frac{p_{1n}}{r_1} \\ 0 & \frac{p_{22}}{r_2} & \cdots & \frac{p_{2n}}{r_2} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \frac{p_{nn}}{r_n} \end{pmatrix}. \tag{55}$$

Note that the Diophantine approximation problem is itself a NP-hard problem, solvable by another CVP [57]).

The core of Algorithm 1 is carried out on the dual matrix $\mathbf{H}^{(D)}$. The algorithm starts by shortening the generator vectors of $\mathbf{H}^{(D)}$ and then sorts them in ascendant order of their norm from the leftmost column to the rightmost one. This procedure minimises p_{11} and therefore constitutes a first step towards minimising Φ, from (49) The algorithm then enters a search mode where in each step the largest numerator p_{ii} is found and its accuracy relaxed so that p_{ii} may diminish. Then, all the remaining off-diagonal elements in that row are written with the denominator that has just been found, so that $r_i = q_{ii}$, as seen in (55). The rational approximation of the diagonal elements is relaxed by means of a continuous fraction algorithm.

ALGORITHM 1: SYNTHESIS OF A LATTICE IN $\mathcal{L}_R$

Input: Generator $\mathbf{H}$, Admissible numb. of paths Γ.
Output: Approximation $\tilde{\mathbf{H}} \in \mathcal{L}_R$; number of cosets |C|.

1: $\mathbf{H}_{\text{red}}^{(D)}, \mathbf{M} \leftarrow \text{LLL}\{(\mathbf{H}^{-1})^T\}$, $\mathbf{M}$ unimodular
2: Sort columns by increasing norm
3: $\mathbf{Q}, \widehat{\mathbf{H}}_{\text{red}}^{(D)}, \mathbf{J} \leftarrow QR\left(\text{sort}(\mathbf{H}_{\text{red}}^{(D)})\right)$; $\mathbf{J}$ permutation, $\widehat{\mathbf{H}}_{\text{red}}^{(D)}$ u.t.
4: Do until $|C| < \Gamma$
5: Obtain $\tilde{\mathbf{H}}_{\text{red}}^{(D)}$: for each row i, obtain rational approximation of each q_{ij} using a common denominator r_i and with maximum error δ
6: $\mathbf{P}, \mathbf{R} \leftarrow \tilde{\mathbf{H}}_{\text{red}}^{(D)}$; as in expression (50)
7: $|C| = \prod_{i=1}^{n} p_i$
8: increment δ
9: end loop
10: $\tilde{\mathbf{H}} = \mathbf{Q}^T \left((\tilde{\mathbf{H}}_{\text{red}}^{(D)} \cdot \mathbf{J}^{-1})^{-1} \right)^T \mathbf{M}^{-1}$

The complexity of Algorithm 1 is dominated by the LLL reduction, $\mathcal{O}(n^4)$. In addition, the QR decomposition is $\mathcal{O}(n^3)$, sorting is $\mathcal{O}(n\log n)$ [58] (sec. 6.5), and the iterations for rational approximation is dominated by a continued fractions algorithm, having $\mathcal{O}(n^3)$ [59] . Sphere decoding is well known to have a random number of branch expansions during the exploration of the tree (unless fixed complexity sphere decoding is used [60]). That number varies each time a received vector is decoded, and is highly dependent upon the noise power. We note that, while in the proposed detector the number of cosets is also a random variable, it only affects the pre-processing stage. Then, the complexity remains constant over the coherence time of that lattice instance.

Fig. 21 shows an example of how the number of cosets evolves as the error tolerance δ increases. At the same time, it is also possible to see the corresponding increase of the Frobenius distance between the synthetic dual lattice and the original dual lattice, $\left\| \mathbf{H}^{(D)} - \tilde{\mathbf{H}}^{(D)} \right\|_F$ (shown as a percentage). Notice that this is the distance between the dual lattices and *not* the distance between the primal lattices (as matrix inversion changes the Frobenius norm). This example in Fig. 21 is for a *n*=8 dimensional (real) lattice, with a limit of $\Gamma = 500$ cosets. The algorithm terminates outputting a synthetic lattice with Φ=324 cosets after 144 iterations (many of them do not lead to a change of variables, which explains why there are fewer than 144 points plotted).

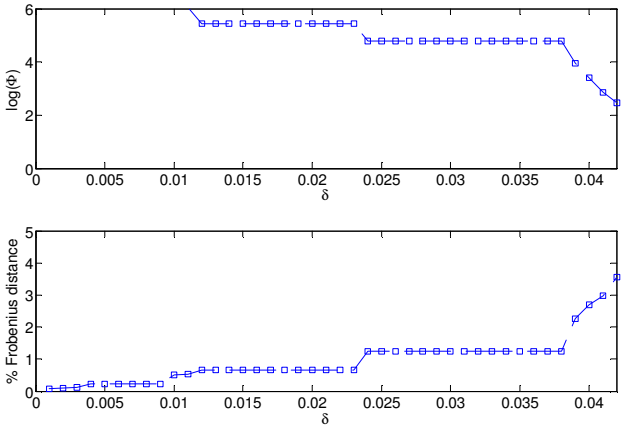


Fig. 21: Examples of the evolution of the number of cosets and Frobenius distance in Algorithm 1 with *n*=8 dimensional random lattices, as the error tolerance δ increases.

E. Performance Results and Discussion

We have assessed the proposed receiver using lattices which arise in MIMO communications under Rayleigh flat fading channel and compared its performance with the one of lattice-reduction-aided receivers (with ZF and with ordered successive interference cancellation (OSIC) schemes), which are well known for capturing the full diversity provided by MLD [61]. The performances of linear ZF, linear minimum mean square error (MMSE) and OSIC without lattice-reduction are also included in the results presented in Fig. 22. The proposed trellis-based detection also attains full diversity while reducing the gap between lattice reduction and MLD and the required number of cosets needed to achieve quasi-optimum detection is surprisingly small.

Algorithm 1 searches for an approximate lattice with a specified maximum number of cosets Γ (i.e., number of paths in a trellis). However, their average number is about half of the specified Γ. Fig. 22 shows the performance for a typical benchmarking MIMO configuration (4×4 antennas with 64 QAM). Limiting the admissible number of cosets to Γ=100, we observe that an average of 38 paths is enough to synthesise good approximated lattices in $\mathcal{L}_R$ to achieve a performance

about 1.2 dB away from ML, coinciding with the performance of LLL-OSIC-ZF. With an average of 506 cosets, the gap shortens to 0.6 dB.

For 2×2 configuration with 64-QAM, an average of 20 cosets have been found to assure the same performance as MLD and for the 3×3 setup, the performance is 0.2 dB away from MLD with 34 cosets on average.

The complexity in Algorithm 1 is dominated by the LLL reduction, $\mathcal{O}(n^4)$, added to $\mathcal{O}(n^3)$ in the QR decomposition, and the complexity of the iterations for rational approximation, dominated by a continued fractions algorithm, $\mathcal{O}(n^3)$ [59]. Sphere decoding is well known to have a random number of branch expansions during the exploration of the tree (unless fixed complexity sphere decoding is used [60]). That number varies each time a received vector is decoded, and is highly dependent on the noise power. We note that, while in the proposed detector the number of cosets is also a random variable, it only affects the pre-processing stage. Then, the complexity remains constant over the coherence time of that lattice instance.

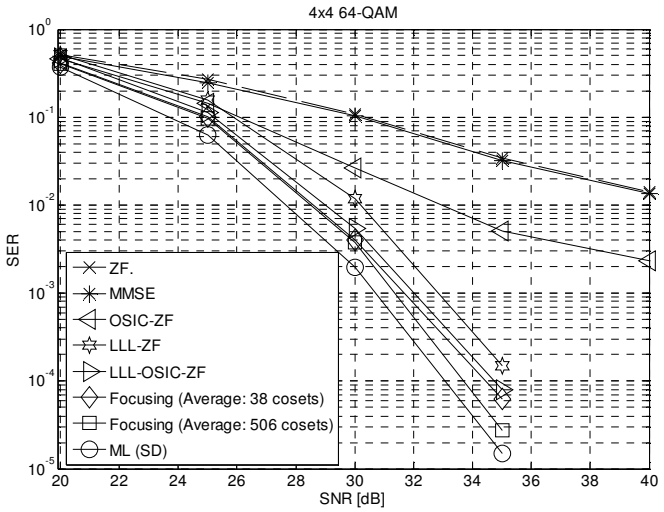


Fig. 22: Symbol error probability when detecting in a lattice with n=8 real dimensions (4×4 MIMO configuration) with 64-QAM.

F. Number of cosets

Algorithm 1 outputs synthetic lattices with a number of cosets Φ which is a random variable. The average number of cosets in those returned latticesshould be $E\{\Phi\} < \Gamma$, however, there are particular instances when a lattice cannot be found with a number of cosets $\Phi < \Gamma$. Those instances are counted in the rightmost bin in the histogram, with "*" above the bin. The visible in the probability distribution function (pdf) was chosen for "*" to represent not more than 1 per cent of the lattices. Moreover, all histograms computed were normalised, so that they can be a good approximation to pdfs. As an example, it is shown a n=8 case in Fig. 23 (which includes a graphical representation of the standard deviation as a bar on top of the pdf and centred at the average). The total number of lattices taken into account to generate the pdfs corresponds to the sum of all the instances considered to obtain all the points in the SER curves.

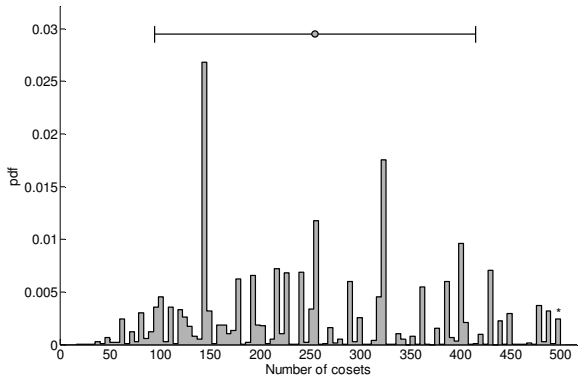


Fig. 23: Pdf of the number of cosets when generating lattices with Algorithm 1 for n=8 real dimensions (4×4 configurations), limiting to Γ=500.

One surprising result is that the pdfs showing the number of cosets present in the various dimension sizes analysed are far from being uniform. Indeed, for all the dimension sizes investigated, and regardless of the limit Γ, the number of cosets tends to cluster around some particular numbers and the resulting probability density functions (pdfs) are "almost discrete" and "almost periodic", though they cannot strictly be considered as such. This fact is difficult to explain and can only be justified by the existence of an underlying number theoretic property governing the possible number of coset groups in n dimensions, which is a mathematical problem beyond the scope of this engineering approach.

From the generation of the distribution of the number of cosets for several MIMO configurations (omitted in this paper due to space limitations), it was observed that the number of cosets needed for near-optimal performance diminishes for smaller alphabets (smaller M). This happens because the distortion between the received lattice and the approximated lattice in $\mathcal{L}_{\mathrm{R}}$ increases as one gets further away from the origin.

As expected, as the dimensionality goes up, the average number of cosets, $E\{\Phi\}$, required for quasi-optimal detection, also grows, however it is found that $E\{\Phi\}$ is still affordable up to 8 dimensions (4×4), which includes the most important scenarios in MIMO.

Finally, note that, i) by construction, the number of trellis paths behaves has an upper bound to the number of trellis states and ii) note that the length of the trellises (number of segments) is determined by the dimensionality of the real lattice (n=2N_R) and therefore, for the typical number of antennas in MIMO, these trellises are rather short.

VII. CONCLUSIONS

The existence of multipath propagation in wireless systems is known today as an effect than must be exploited in order to increase the capacity of a radio link. However, in open loop configurations (channel estimated by the receiver, although not know by the transmitter), this carries a large complexity

problem to the receiver side, which, in essence, is required to compute a closest vector problem in a lattice for each received vector of symbols. As almost all lattice problems, this problem is simple to describe but rather hard to solve optimally. This paper frames several suboptimal solutions for CVP in MIMO communication systems in the context of lattice geometry.

When describing the fundamental properties of lattices, the geometric relation between a lattice and its dual lattice was clarified, which is a geometric relation much ignored in the literature. Capitalising on that relation, a technique was devised in that samples points lying on sets of hyperplanes that have the highest density of lattice points on them. Those samples are then quantised to the lattice via zero forcing and the best candidate is declared. The technique exhibits a considerable gain (up to 7 dB in the 4×4 / 64-QAM case) in comparison to OSIC with ZF.

A major contribution in this paper was the proposal of a new type of receiver that maps the problem onto a sequence detection problem, solvable by the well known Viterbi algorithm. First, a lattice with a trellis is "synthesised" near the original received lattice. Then, the original lattice problem is focused (i.e., linearly transformed) onto the "synthetic" lattice, which processes a trellis description. One can see this as a generalization of the zero-forcing concept, extended to a much larger family of lattices to map (or *focus*) the problem onto.

The derivation of the property that makes a lattice a member of the family of lattices with a trellis structure and an algorithm was given to create one of those lattices "nearby" the typical Gaussian lattices in MIMO. The basis vectors of the synthetic lattice and the basis vectors of the original lattice are close to each other and for finite QAM alphabets the two lattices are roughly the same in the region of interest (i.e., not far from the origin). Given this geometric similarity, the Voronoi regions of both lattices chiefly overlap. Hence, the linear transformation that focuses the original lattice onto the synthetic one, is well-conditioned and close to identity, leading to a very small noise increase. The distortion of the Voronoi regions associated with maximum-likelihood detection is therefore minimized and consequently the performance attained in the MIMO-CVP is close to optimal.

For 2×2, 3×3, 4×4 and 6×6 configurations (not all depicted for space reasons), decoding on the synthetic lattice outperforms all the most used detection techniques with or without pre-processing. Moreover, likewise SD, it also attains the same diversity as MLD.

As expected, the number of cosets necessary for near-optimum performance increases with the dimension of the lattices. Note that while the number of nodes to explore in SD is random and depends on both the lattice and each received vector itself [62], decoding in a trellis has fixed-complexity during the period of time while that particular lattice represents the channel matrix coefficients, and the algorithm run after each channel update is polynomial , $\mathcal{O}(n^4)$, the same as lattice-reduction-aided.

ACKNOWLEDGMENTS

The author is thankful to both Ian Wassell (in Cambridge) and Frank Kschischang (in Toronto) for having supervised this research. The author is also thankful to Oded Regev (Tel Aviv University) for discussions on the *quasi orthogonal sublattice problem.* The author was funded for this research with grants from the Foundation for Science and Technology (FCT), the Gulbenkian Foundation, the Royal Academy of Engineering, the Cambridge Philosophical Society, The Computer Laboratory, and from Fitzwilliam College (University of Cambridge).

REFERENCES

[1] H. Yang, "A road to future broadband wireless access: MIMO-OFDM-Based air interface," *IEEE Communications Magazine*, vol. 43, no. 1, pp. 53-60, January 2005.

[2] Jeffrey G. Andrews, Arunabha Ghosh, and Rias Muhamed, *Fundamentals of WiMAX*. Upper Saddle River, New Jersey, USA: Prentice Hall - Pearson Education, 2007.

[3] F. Khan, *LTE for 4G Mobile Broadband: Air Interface Technologies and Performance*. Cambridge, UK: Cambridge University Press, 2009.

[4] Q. H. Spencer, C. B. Peel, A. L. Swindlehurst, and M. Haardt, "An introduction to the multi-user MIMO downlink," *IEEE Communications Magazine*, vol. 42, no. 10, pp. 60-67, October 2004.

[5] A. Ghosh, R. Ratasuk, B. Mondal, N. Mangalvedhe, and T. Thomas, "LTE-advanced: next-generation wireless broadband technology," *IEEE Wireless Communications*, vol. 17, no. 3, pp. 10-22, June 2010.

[6] D. Bai, H. Nguyen, T. Kim, and I. Kang, "LTE-Advanced modem design: challenges and perspectives," *IEEE Communications Magazine*, vol. 50, no. 2, pp. 178-186, February 2012.

[7] S. Parkvall, A. Furuskär, and E. Dahlman, "Evolution of LTE toward IMT-advanced," *IEEE Communications Magazine*, vol. 49, no. 2, pp. 84-91, February 2011.

[8] E. Perahia and R. Stacey, *Next Generation Wireless LANs - Throughput, Robustness, and Reliability in 802.11n*. Cambridge, UK: Cambridge University Press, 2008.

[9] A. Sibille, C. Oestges, and A. Zanella, Eds., *MIMO: From Theory to Implementation*. Amsterdam, Netherlands: Academic Press, 2011.

[10] G. R. Hiertz et al., "The IEEE 802.11 universe," *IEEE Communications Magazine*, vol. 48, no. 1, pp. 62-70, January 2010.

[11] T. Baykas et al., "IEEE 802.15.3c: the first IEEE wireless standard for data rates over 1 Gb/s," *IEEE Communications Magazine*, vol. 49, no. 7, pp. 114-121, July 2011.

[12] R. V. Nee, "Breaking the gigabit-per-second barrier with 802.11ac," *IEEE Wireless Communications*, vol. 18, no. 2, p. 4, April 2011.

[13] A. J. Paulraj, D. A. Gore, R. U. Nabar, and H. Bölcskei, "An overview of MIMO communications - a key to gigabit wireless," *Proceedings of the IEEE*, vol. 92, no. 2, pp. 198-218, February 2004.

[14] H. R. Stuart, "Dispersive Multiplexing in Multimode Optical Fiber," *Science*, vol. 289, pp. 281-283, July 2000.

[15] P. Ödlig et al., "The fourth generation broadband concept," *IEEE Communications Magazine*, vol. 43, no. 1, pp. 63-69, January 2009.

[16] L. Zheng and D. N.C. Tse, "Diversity and multiplexing:a Fundamental tradeoff in multiple antenna channels," *IEEE Transactions on Information Theory*, vol. 49, no. 5, pp. 1073-1096, May 2003.

[17] D. Micciancio and S. Goldwasser, *Complexity of Lattice Problems - A Cryptographic Perspective*. Norwell, Massachusetts, USA: Kluwer Academic Publishers, 2002.

[18] D. S. Hochba, Ed., *Approximation Algorithms for NP-hard Problems*. Boston, Massachusetts , USA: Course Technology / PWS Publishing Company, 1996.

[19] C. M. Bishop, "The curse of dimensionality," in *Pattern Recognition and*

Machine Learning. New York, NY: Springer, 2006, ch. 1.4, pp. 33-38.

[20] J. W. S. Cassels, *An Introduction to the Geometry of Numbers*, 2nd ed. Berlin, Germany: Springer, 1971.

[21] C. Ling, W. H. Mow, and L. Gan, "Dual-lattice ordering and partial lattice reduction for SIC-based MIMO detection," *IEEE Journal of Selec. Topics in Signal Process.*, vol. 3, no. 6, pp. 975-985, December 2009.

[22] E. Agrell, T. Eriksson, A. Vardy, and K. Zeger, "Closest point in lattices," *IEEE Trans. Inf. Theory*, vol. 48, no. 8, pp. 2201-2214, Aug. 2002.

[23] C. Ling, L. Gan, and W. H. Mow, "A dual-lattice view of V-BLAST detection," in *Proc. of ITW' 06, The IEEE Info. Theory Workshop*, Chengdu, China, October 2006, pp. 478-482.

[24] C. Ling and W. H. Mow, "A unified view of sorting in lattice reduction: From V-BLAST to LLL and beyond," in *Proc. of the IEEE Inf. Theory Workshop*, Taormina, Italy, 2009, pp. 529-533.

[25] K. Su and F. R. Kschischang, "Coset-based lattice detection for MIMO systems," in *Proc. of ISIT'07 - IEEE Inter. Symp. on Inf. Theory*, Nice, France, June 2007, pp. 1941 - 1945.

[26] R. Piziak and P. L. Odell, *Matrix Theory - From Generalized Inverses to Jordan Form*. Boca Raton, Florida, USA: Chapman & Hall - CRC, 2007.

[27] K. Su and F. R. Kschischang, "Coset-based lattice detection for MIMO systems," in *Proc. of ISIT'07 - IEEE Inter. Symp. on Information Theory*, Nice, France, June 2007, pp. 1941 - 1945.

[28] C. Ling, "On the proximity factors of lattice reduction-aided decoding," *IEEE Transactions on Signal Processing*, vol. 59, no. 6, pp. 2795-2808, June 2011.

[29] Robert F. H. Fischer, *Precoding and Signal Shaping for Digital Transmission*. Chischester, UK: John Wiley & Sons, 2005.

[30] S. Haykin, *Adaptive Filter Theory*, 3rd ed. Upper Saddle River, New Jersey, EUA: Prentice Hall, 1996.

[31] U. Madhow, *Fundamentals of Digital Communication*. Cambridge, UK: Cambridge University Press, 2008.

[32] G. D. Golden, C. J. Foschini, R. A. Valenzuela, and P. W. Wolniansky, "Detection algorithm and initial laboratory results using V-BLAST space-time communication architecture," *IET Electronics Letters*, vol. 35, no. 1, January 1999.

[33] P. W. Wolniansky, G. J. Foschini, G. D. Golden, and R. A. Valenzuela, "V-BLAST: an architecture for realizing very high data rates over the rich-scattering wireless channel," in *Proc. of URSI Int. Symposium on*, Pisa, Italy, September 1998, pp. 295-300.

[34] L. Babai, "On Lovász' lattice reduction and the nearest lattice," *Combinatoria*, vol. 6, no. 1, pp. 1-13, January 1986.

[35] W. H. Mow, "Universal lattice decoding: principles and recent advances," *Wireless Communications and Mobile Computing*, vol. 3, pp. 553-569, March 2003.

[36] Y. Jiang, Mahesh K. V., and J. Li, "Performance analysis of ZF and MMSE equalizers for MIMO systems: an in-depth study of the high SNR regime," *IEEE Transactions on Information Theory*, vol. 57, no. 4, pp. 2008-2026, April 2011.

[37] Y. Shang and X.-G. Xia, "An improved fast recursive algorithm for V-BLAST with optimal ordered detections," in *Proc of ICC'08 - IEEE Inter. Conference on Communications*, Beijing, China, May 2008, pp. 756-760.

[38] C. Windpassinger, "Detection and precoding for multiple input multiple output channels," PhD thesis, University of Erlangen-Nürnberg, Erlangen, Germany, 2004.

[39] R. A. Trujillo, V. M. Garcia, A. M. Vidal, S. Roger, and A. Gonzalez, "A gradient-based ordering for MIMO decoding," in *Proc. of the 9th IEEE Inter. Symp. on Signal Processing and Information Technology (ISSPIT)*, Ajman, United Arab Emirates, 2009, pp. 5-8.

[40] M. Taherzadeh, A. Mobasher, and A. K. Khandani, "LLL reduction achieves the receive diversity in MIMO decoding," *IEEE Transactions on Information Theory*, vol. 53, no. 12, pp. 4801-4805, December 2007.

[41] H. Yao and G.W. Wornell, "Lattice-reduction-aided detectors for MIMO communication systems," in *Proc. of GLOBECOM' 02 - IEEE Global Telecommunications Conference*, Taipei, Taiwan, 2002, pp. 424 -428.

[42] P. Q. Nguyen and B. Vallée, Eds., *The LLL Algorithm*. Berlin, Germany: Springer, 2010.

[43] D. Wübben, D. Seethaler, J. Jaldén, and G. Matz, "Lattice Reduction," *IEEE Signal Processing Magazine*, vol. 28, no. 3, pp. 70-91, May 2011.

[44] J. Jaldén and B. Ottersten, "On the complexity of sphere decoding in digital communications," *IEEE Transactions on Signal Processing*, vol. 53, no. 4, pp. 1474-1484, April 2005.

[45] B. Hassibi and H. Vikalo, "On the sphere-decoding algorithm I. Expected complexity," *IEEE Trans. on signal Process.*, vol. 53, no. 8, pp. 2806-2818, August 2005.

[46] M. Taherzadeh, A. Mobasher, and A. K. Khandani, "LLL reduction achieves the receive diversity in MIMO decoding," *IEEE Trans. on Inf. Theory*, vol. 53, no. 12, pp. 4801-4805, December 2007.

[47] G. D. Forney Jr., "Coset codes-part II: binary lattices and related codes," *IEEE Trans. Inform. Theory*, vol. 34, no. 5, pp. 1152-1187, Sep. 1988.

[48] F. R. Kschischang and V. Sorokine, "On the trellis structure of block codes," *IEEE Trans. Inf. Theory*, vol. 41, no. 6, pp. 1924-1937, Nov. 1995.

[49] J. H. Conway and N. J. A. Sloane, *Sphere Packings, Lattices and Groups*, 3rd ed. New York, New York, USA: Springer, 1999.

[50] N. H. Higham, "Matrix nearness problems and applications," in *Applications of Matrix Theory*, M. J. C. Gover and S. Barnett, Eds. Oxford, UK: Oxford University Press, 1989, pp. 1-27.

[51] J. Lee and S.-C. Park, "MIMO Detector Based on Trellis," *IEICE Trans. on Communications*, vol. E91-B, no. 3, pp. 951-954, March 2008.

[52] A. Vardy, "Trellis Structure of Codes," in *Handbook of Coding Theory*, Vera Pless and W. Cary Huffman, Eds. Amsterdam, The Netherlands: Elsevier, 1998, ch. 24, pp. 1989-2117.

[53] G. D. Forney Jr., "Density / length profiles and trellis complexity of lattices," *IEEE Trans. Inform. Theory*, vol. 40, no. 6, pp. 1753-1772, Nov. 1994.

[54] A. H. Banihashemi, "Decoding complexity and trellis structure of lattices," PhD thesis, University of Waterloo, Ontario, Canada, 1997.

[55] A. H. Banihashemi and F. R. Kschischang, "Tanner graphs for group block codes and lattices: construction and complexity," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 822-834, Feb. 2001.

[56] H. Kan and H. Shen, "The bases associated with trellis of a lattice," *IEICE Trans. on Fundamentals*, vol. E88, no. 7, pp. 2030-2033, July 2005.

[57] L. Lovász, *An Algorithmic Theory of Numbers, Graphs and Convexity*. Philadelphia, Pennsylvania, USA: Society for Industrial and Applied Mathematics (SIAM), 1986, ch. 1, pp. 15-38.

[58] K. A. Berman and J. L. Paul, "Matching and network flow algorithms," in *Algorithms: Sequantial, Parallel and Distribuited*. Boston, Massachussetts: Thomson, 2005, ch. 14.

[59] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information*. Cambridge, UK: Cambridge University Press, 2010, ch. 5, p. 230.

[60] L. G. Barbero and J. S. Thompson, "Fixing the complexity of the sphere decoder for MIMO detection," *Trans. Wireless Commun*, vol. 7, no. 6, pp. 2131-2142, June 2008.

[61] M. Taherzadeh, A. Mobasher, and A. K. Khandani, "LLL reduction achieves the receive diversity in MIMO decoding," *IEEE Trans. on Inf. Theory*, vol. 53, no. 12, pp. 4801-4805, Dec. 2007.

[62] K. Su and Ian J. Wassell, "A new ordering for efficient sphere decoding," in *IEEE Inter. Conf. on Communications*, vol. 3, Seoul, Korea, 2005, pp. 1906-1910.