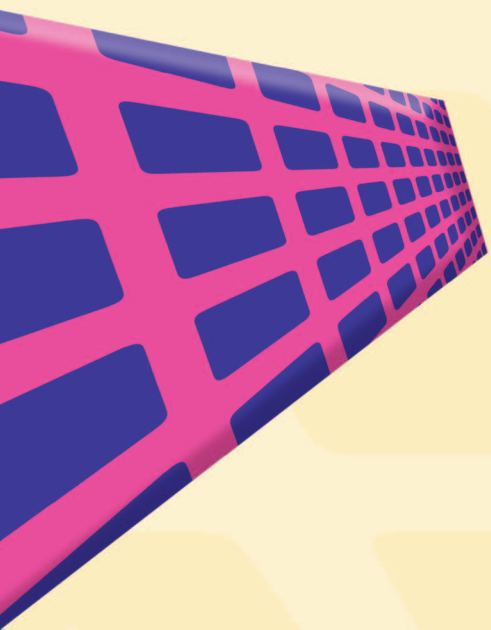
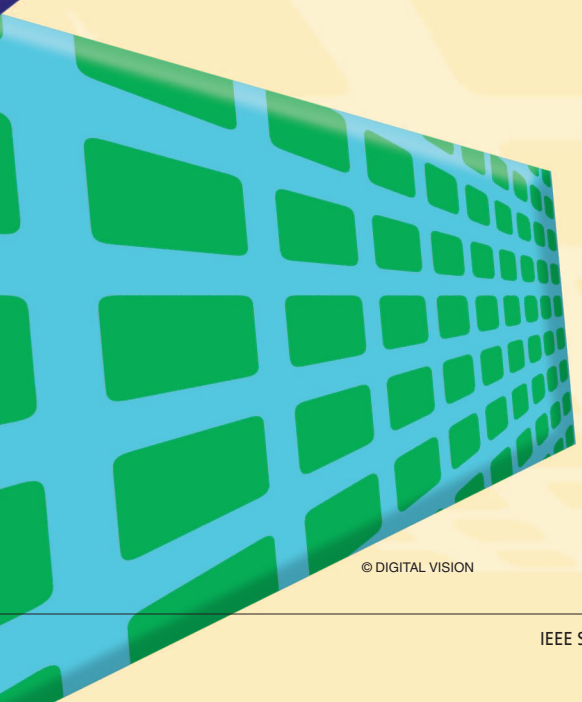
Dirk Wübben, Dominik Seethaler, Joakim Jaldén, and Gerald Matz

Lattice Reduction

A survey with applications in wireless communications





Lattice reduction is a powerful concept for solving diverse problems involving point lattices. Signal processing applications where lattice reduction has been successfully used include global positioning system (GPS), frequency estimation, color space estimation in JPEG pictures, and particularly data detection and precoding in wireless communication systems. In this article, we first provide some background on point lattices and then give a tutorial-style introduction to the theoretical and practical aspects of lattice reduction. We describe the most important lattice reduction algorithms and comment on their performance and computational complexity. Finally, we discuss the application of lattice reduction in wireless communications and statistical signal processing. Throughout the article, we point out open problems and interesting questions for future research.

INTRODUCTION

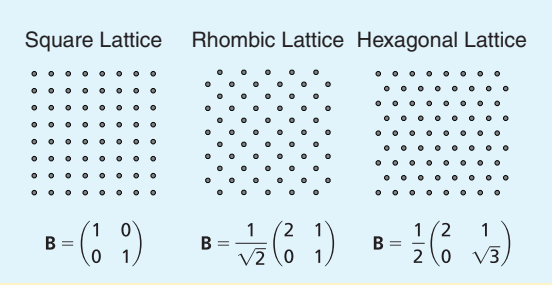
Lattices are periodic arrangements of discrete points (see "Lattices"). Apart from their widespread use in pure mathematics, lattices have found applications in numerous other fields as diverse as cryptography/cryptanalysis, the geometry of numbers, factorization of integer polynomials, subset sum and knapsack problems, integer relations and diophantine approximations, the spigot algorithm for π, materials science and solid-state physics (specifically crystallography), and coding theory. Recently, lattices have also been applied in a variety of signal processing problems.

Digital Object Identifier 10.1109/MSP.2010.938758
Date of publication: 19 April 2011

1053-5888/11/$26.00©2011IEEE

LATTICES

A (point) lattice $\mathcal{L}$ is a periodic arrangement of discrete points. Two-dimensional examples are the square, rhombic, and hexagonal lattices shown in Figure S1.



[FIGS1] Two-dimensional examples of the square, rhombic, and hexagonal lattices.

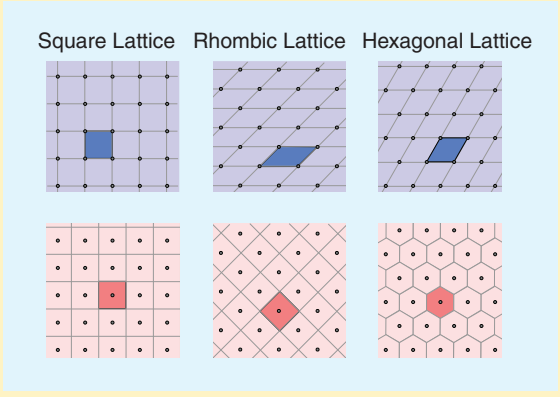
Any lattice can be characterized in terms of a (nonunique) basis $\mathbf{B} = (\mathbf{b}_1 \ldots \mathbf{b}_m)$ that allows any lattice point to be represented as a superposition of integer multiples of the basis vectors $\mathbf{b}_\ell$, i.e., any $\mathbf{x} \in \mathcal{L}$ can be written as

$$\mathbf{x} = \sum_{\ell=1}^{m} z_\ell \, \mathbf{b}_\ell, \quad z_\ell \in \mathbb{Z}.$$

If $\mathbf{x}_1 \in \mathcal{L}$ and $\mathbf{x}_2 \in \mathcal{L}$, then $k\mathbf{x}_1 + \ell\mathbf{x}_2 \in \mathcal{L}$ for any $k, \ell \in \mathbb{Z}$. To any lattice basis, there is an associated fundamental parallelotope $\mathcal{P}(\mathbf{B})$, i.e., the set of points that can be written as

$$\mathbf{x} = \sum_{\ell=1}^{m} \theta_\ell \, \mathbf{b}_\ell, \quad 0 \leq \theta_\ell < 1.$$

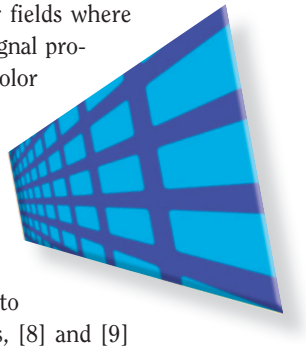
The Voronoi region $\mathcal{V}(\mathcal{L})$ of a lattice is the set of points that are closer to the origin than to any other lattice point. The Voronoi region is a lattice invariant, i.e., it does not depend on a specific lattice basis. Translating either the fundamental parallelotope or the Voronoi region to all lattice points induces a tessellation of Euclidean space. Illustrations of the parallelotope (blue) and the Voronoi region (red) and the associated tessellations for our two-dimensional example lattices are shown in Figure S2.



[FIGS2] Illustrations of the parallelotope and the Voronoi region and the associated tessellations of two-dimensional example lattices.

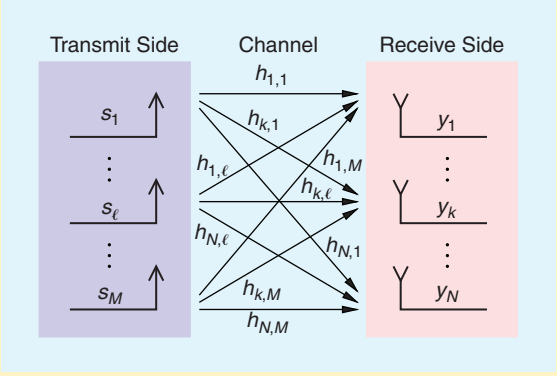
Lattices are used to develop powerful source and channel codes for many communications applications, specifically in scenarios with multiple terminals or with side-information (e.g., [1]). Different from that strand of work, our focus in this article is on the principle of lattice reduction. Lattice reduction is concerned with finding improved representations of a given lattice using algorithms like Lenstra, Lenstra, Lovász (LLL) reduction [2] or Seysen reduction [3]. It is a topic of great interest, both as a theoretical tool and as a practical technique. Since we feel that lattice reduction may be relevant to a much wider class of engineering problems than currently considered, this survey article targets a broader signal processing audience. We give a tutorial-style introduction to lattices and lattice reduction algorithms, discuss the application of lattice reduction in wireless communications and parameter estimation, and as a by-product provide convenient entry points to the literature on the topic. Wherever appropriate, we also point out possible topics for future research.

Lattice reduction plays an important role in the above-mentioned fields as it leads to efficient solutions for several classical problems in lattice theory. For example, lattice reduction is intimately linked to the search for the shortest vector in a lattice, which in turn is related to the development of the now famous sphere decoding algorithm (e.g., [4] and [5]). With regard to signal processing applications, lattice reduction is potentially useful in problems that involve integer programming (i.e., optimization problems with integer variables). An example is provided in [6], where an integer least squares problem—which is NP-hard—is solved in an approximate but efficient manner using lattice reduction techniques. While the application context in that paper was GPS, the authors mention radar imaging and magnetic resonance imaging as other fields where their results could have impact. Another signal processing application of lattice reduction is color space estimation in JPEG images [7]. Here, the quantized discrete cosine transform coefficients in the three color planes are viewed as lattice points in three-dimensional (3-D) space, with the 3-D lattice being determined by the color space transformation matrix. Building on the analogy to simultaneous Diophantine approximations, [8] and [9] use lattice reduction techniques for determining the intercept of multiple periodic pulse trains and for estimating their parameters. A somewhat similar approach was used to develop a lattice-reduction-based implementation of the maximum likelihood (ML) estimator for the frequency of a sinusoid in noise [10]. Lattice reduction has further been applied to various problems in wireless communications, e.g., to the equalization of frequency-selective channels [11], to receiver design for unitarily precoded quadrature amplitude modulation (QAM) transmission over flat fading channels [12], to joint data detection and



MIMO WIRELESS

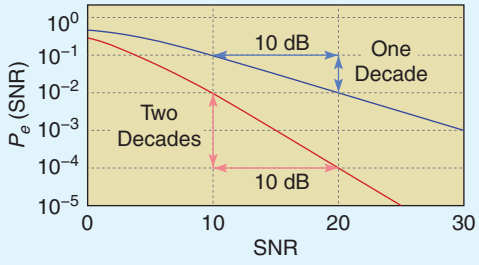
MIMO systems use multiple antennas at both link ends and offer tremendous performance gains without requiring additional bandwidth or transmit power. Two important MIMO gains are the multiplexing gain, which corresponds to an increase of the data rate, and the diversity gain, which corresponds to an increase of the transmission reliability.



[FIGS3] MIMO system using M transmit antennas and N receive antennas.

In this article, we consider MIMO systems with M transmit antennas that transmit parallel data streams s_ℓ, $\ell = 1, \ldots, M$. At the receive side, N receive antennas pick up mixtures of the transmit signals, i.e., $y_k = \sum_{\ell=1}^{M} h_{k,\ell}\, s_\ell + w_k$, where w_k denotes additive Gaussian noise (Figure S3). In the MIMO spatial multiplexing system considered in the section "Lattice Reduction for Data Detection," the receive antennas cooperate and channel state information is available at the receive side. In contrast, the MIMO precoding scheme considered in the section "Lattice Reduction for Precoding" requires channel state information at the transmit side and allows for distributed noncooperative receive antennas.

A useful MIMO performance figure is the (spatial) diversity order, defined as the asymptotic slope of the error probability $P_e(\mathrm{SNR})$ as a function of signal-to-noise ratio (SNR) on a log-log scale (see Figure S4). The diversity order depends on the statistics of the MIMO channel and on the transceiver scheme. For spatial multiplexing the maximum diversity order ("full diversity") equals N, whereas for MIMO precoding the maximum diversity order equals M.



[FIGS4] Illustration of MIMO systems with diversity order 1 (blue) and with diversity order 2 (red).

channel estimation in quasi-synchronous code division multiple access (CDMA) [13], and to equalization in precoded orthogonal frequency division multiplexing (OFDM) systems [14].

Recently (see, e.g., [15] and [16]), lattice reduction (and lattice theory in general) turned out to be extremely useful for detection and precoding in wireless multiple-input multiple-output (MIMO) systems, i.e., systems that use multiple antennas at the transmit and receive side [17] (see "MIMO Wireless" for the basic notions of MIMO systems). The fundamental idea here is to exploit the discrete nature of the digital data and view the channel matrix that acts on the data as a basis (generator) of a point lattice. Via this interpretation, detection and precoding problems can be tackled using tools from lattice theory. Typically, a three-stage procedure is pursued:

1) An improved basis for the lattice induced by the channel is determined via lattice reduction. The original basis and the reduced basis are related via a unimodular matrix.
2) The detection/precoding problem is solved with respect to the reduced basis.
3) The solution is transformed back to the original domain using the unimodular matrix.

Since the reduced basis has nicer mathematical properties (e.g., smaller orthogonality defect and smaller condition number), solving detection and precoding problems with respect to the reduced basis offers advantages with respect to performance and complexity. For example, it was shown recently that in some scenarios even suboptimum detection/precoding techniques can achieve full diversity (see "MIMO Wireless") when preceded by LLL lattice reduction (e.g., [18]–[23]).

While the multitude of applications illustrates the theoretical significance of lattice reduction, its practical importance is corroborated by the fact that very-large-scale integration (VLSI) implementations of lattice reduction algorithms for MIMO systems have recently been presented [24]–[28]. These hardware implementations were motivated by the fact that two core problems in the practical realization of MIMO systems are data detection at the receive side and broadcast precoding at the transmit side, and that several groups have proposed to use lattice reduction algorithms to solve these problems efficiently. In fact, it turned out that lattice reduction approaches have the potential to achieve high performance at low computational complexity. The hardware implementations of lattice reduction algorithms will be applicable to WiMAX, WiFi, and 3GPP/3GPP2 systems, which constitute an economically significant market.

POINT LATTICES

In this section, we discuss the basic concepts relating to point lattices in a real Euclidean space, we provide a simple

motivating example for the use of lattice reduction in MIMO wireless systems, and we discuss briefly the necessary modifications for complex-valued lattices. Further details on the theory of lattices can be found in [4] and [29]–[33].

LATTICES ARE PERIODIC ARRANGEMENTS OF DISCRETE POINTS.

REAL-VALUED LATTICES

A real-valued lattice $\mathcal{L}$ is a discrete additive subgroup of $\mathbb{R}^n$. Any lattice can be characterized in terms of a set of $m \leq n$ linearly independent basis (or generator) vectors $\{\mathbf{b}_1, \ldots, \mathbf{b}_m\}$, $\mathbf{b}_\ell \in \mathbb{R}^n$, as

$$\mathcal{L} \triangleq \left\{ \mathbf{x} \,\middle|\, \mathbf{x} = \sum_{\ell=1}^{m} z_\ell \, \mathbf{b}_\ell, \; z_\ell \in \mathbb{Z} \right\}.$$

Here, $\mathbb{Z}$ denotes the set of integers and m is referred to as the rank or dimension of the lattice. Lattices are periodic arrangements in the sense that translations of the lattice by an arbitrary integer multiple of any lattice point leave the lattice unchanged; formally, for any $\mathbf{x}, \mathbf{y} \in \mathcal{L}$ there is $\mathbf{y} + k\mathbf{x} \in \mathcal{L}$ for all $k \in \mathbb{Z}$. For convenience, we will often arrange the basis vectors into an $n \times m$ matrix $\mathbf{B} = (\mathbf{b}_1 \ldots \mathbf{b}_m)$ and simply call $\mathbf{B}$ the basis of the lattice. Since the basis vectors were assumed linearly independent, $\mathbf{B}$ has full (column) rank, i.e., $\text{rank}(\mathbf{B}) = m$. Any element of the lattice can then be represented as $\mathbf{x} = \mathbf{B}\mathbf{z}$ for some $\mathbf{z} \in \mathbb{Z}^m$.

The simplest lattices are the cubic lattices, obtained for $n = m$ by choosing the basis vectors as $\mathbf{b}_\ell = \mathbf{e}_\ell$, where $\mathbf{e}_\ell$ denotes the ℓth column of the n-dimensional identity matrix $\mathbf{I}_n = (\mathbf{e}_1 \ldots \mathbf{e}_n)$. In this case we have $\mathbf{B} = \mathbf{I}_n$ and $\mathcal{L} = \mathbb{Z}^n$. Note, however, that due to its periodicity, $\mathbb{Z}^n$ can also be generated by the basis vectors $\{\mathbf{e}_1, \mathbf{e}_2 + k\mathbf{e}_1, \ldots, \mathbf{e}_n + k\mathbf{e}_1\}$, $k \in \mathbb{Z}$, i.e., $\tilde{\mathbf{B}} = \mathbf{B} + k(\mathbf{0}\ \mathbf{e}_1 \ldots \mathbf{e}_1)$. While $\mathbf{B} = \mathbf{I}_n$ is an orthogonal basis, the columns of $\tilde{\mathbf{B}}$ become more and more colinear as k increases (i.e., the condition number of $\mathbf{B}$ increases). Obviously, when working with $\mathbb{Z}^n$, the basis $\mathbf{I}_n$ is to be preferred over $\tilde{\mathbf{B}}$. Figure 1 illustrates the case $n = m = 2$ by showing the bases $\mathbf{B} = \mathbf{I}_2$ and $\tilde{\mathbf{B}} = \mathbf{B} + 3(\mathbf{0}\ \mathbf{e}_1)$ along with some lattice characteristics defined later in this section.

The foregoing example revealed that the basis for a given lattice is not unique. Indeed, it can be shown that there exist infinitely many bases for a lattice; it is exactly these degrees of freedom with the choice of a lattice basis that are exploited by lattice reduction algorithms (see the section "Lattice Reduction Techniques"). The requirement that two matrices $\mathbf{B}$ and $\mathbf{BT}$ (with $\mathbf{T}$ an $m \times m$ matrix) span the same lattice is equivalent to $\mathbf{B}\mathbb{Z}^m = \mathbf{BT}\mathbb{Z}^m$ and thus $\mathbb{Z}^m = \mathbf{T}\mathbb{Z}^m$. The last equality holds if and only if $\mathbf{T}$ is invertible and both $\mathbf{T}$ and $\mathbf{T}^{-1}$ have integer elements. Equivalently, $\mathbf{T}$ must be a matrix with integer elements and determinant $|\det(\mathbf{T})| = 1$; in fact, these two properties guarantee that the inverse exists and also is an integer matrix. This follows from the observation that the (k, l)th element of $\mathbf{T}^{-1}$ equals $(-1)^{k+l}\det(\mathbf{T}_{kl})/\det(\mathbf{T}) \in \mathbb{Z}$ ($\mathbf{T}_{kl}$ is obtained by deleting the kth row and lth column of $\boldsymbol{T}$). Matrices satisfying these properties are called unimodular and form a group under matrix multiplication.

Any lattice basis also describes a fundamental parallelotope according to

$$\mathcal{P}(\mathbf{B}) \triangleq \left\{ \mathbf{x} \,\middle|\, \mathbf{x} = \sum_{\ell=1}^{m} \theta_\ell \, \mathbf{b}_\ell, \;\; 0 \leq \theta_\ell < 1 \right\}.$$

$\mathcal{P}(\mathbf{B})$ is also a so-called fundamental region, i.e., a region that completely covers the span of $\mathbf{B}$ when shifted to all points of the lattice. Another important fundamental region is the Voronoi region, defined as the set of points in $\mathbb{R}^n$ that are closer to the origin than to any other lattice point,

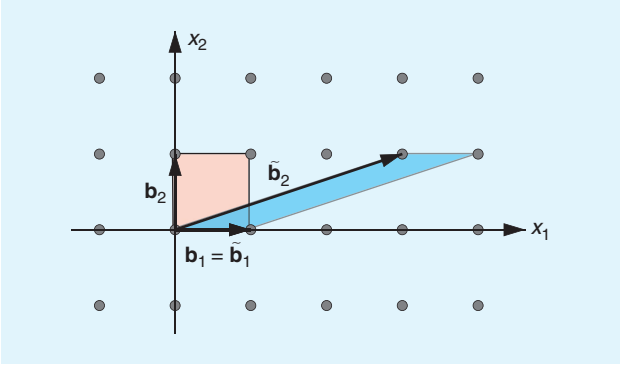
$$\mathcal{V}(\mathcal{L}) \triangleq \{\mathbf{x} \mid \|\mathbf{x}\| \leq \|\mathbf{x} - \mathbf{y}\| \text{ for all } \mathbf{y} \in \mathcal{L}\}.$$

Voronoi regions can be associated to all other lattice points by a simple translation of $\mathcal{V}(\mathcal{L})$. In contrast to the fundamental parallelotope $\mathcal{P}(\mathbf{B})$, the Voronoi region $\mathcal{V}(\mathcal{L})$ is a lattice invariant, i.e., it is independent of the specific choice of a lattice basis.

Clearly, different bases lead to different fundamental parallelotopes. However, the volume (here, volume is defined in the m-dimensional space spanned by the columns of $\mathbf{B}$) of $\mathcal{P}(\mathbf{B})$ is the same for all bases of a given lattice. This volume equals the so-called lattice determinant, which is a lattice invariant defined as the square-root of the determinant of the Gramian $\mathbf{B}^T\mathbf{B}$,

$$|\mathcal{L}| \triangleq \sqrt{\det(\mathbf{B}^T\mathbf{B})}. \tag{1}$$

If the lattice has full rank (i.e., if $n = m$), the lattice determinant equals the magnitude of the determinant of the basis matrix $\mathbf{B}$, i.e., $|\mathcal{L}| = |\det(\mathbf{B})|$. For a transformed basis $\tilde{\mathbf{B}} = \mathbf{BT}$ (with $\mathbf{T}$ being unimodular), we have indeed $\det(\tilde{\mathbf{B}}^T\tilde{\mathbf{B}}) = \det(\mathbf{T}^T\mathbf{B}^T\mathbf{BT}) = \det^2(\mathbf{T})\det(\mathbf{B}^T\mathbf{B}) = |\mathcal{L}|^2$. The "quality" of a



[FIG1] Example lattice $\mathbb{Z}^2$ with bases $\mathbf{B} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ and $\tilde{\mathbf{B}} = \begin{pmatrix} 1 & 3 \\ 0 & 1 \end{pmatrix}$ and associated fundamental parallelograms $\mathcal{P}(\mathbf{B})$ (in red) and $\mathcal{P}(\tilde{\mathbf{B}})$ (in blue). The lattice determinant [defined in (1)] equals $|\mathcal{L}| = 1$ and the orthogonality defects [see (2)] are $\xi(\mathbf{B}) = 1$ and $\xi(\tilde{\mathbf{B}}) = \sqrt{10}$.

lattice basis can be measured in terms of the orthogonality defect, defined as

$$\xi(\mathbf{B}) = \frac{1}{|\mathcal{L}|}\prod_{\ell=1}^{m}\|\mathbf{b}_\ell\|. \tag{2}$$

For any $m \times m$ positive definite matrix $\mathbf{A}$ with elements $a_{k,\ell}$, the Hadamard inequality states that $\det(\mathbf{A}) \leq \prod_{\ell=1}^{m} a_{\ell,\ell}$ with equality if and only if $\mathbf{A}$ is diagonal. Setting $\mathbf{A} = \mathbf{B}^T\mathbf{B}$ this implies that the orthogonality defect is bounded from below as $\xi(\mathbf{B}) \geq 1$, with equality if and only if $\mathbf{B}$ is orthogonal. The fundamental parallelotope, lattice determinant, and orthogonality defect are illustrated for the lattice $\mathbb{Z}^2$ in Figure 1.

To any lattice, there is an associated dual lattice, defined by

$$\mathcal{L}^\star \triangleq \{\mathbf{x}^\star \in \mathrm{span}(\mathbf{B}) |\, \mathbf{x}^T\mathbf{x}^\star \in \mathbb{Z} \quad \text{for all } \mathbf{x} \in \mathcal{L}\}.$$

If $\mathbf{B}$ is a basis for the primal lattice $\mathcal{L}$, then a basis $\mathbf{B}^\star$ for the dual lattice $\mathcal{L}^\star$ can be obtained via the right Moore-Penrose pseudoinverse, i.e.,

$$\mathbf{B}^\star = \mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}. \tag{3}$$

Since $\mathbf{B}^T\mathbf{B}^\star = \mathbf{I}_m$, it follows that the primal and dual basis vectors are bi-orthogonal, i.e., $\mathbf{b}_\ell^T\mathbf{b}_k^\star = 0$ for $\ell \neq k$. Geometrically, this means that the dual basis vector $\mathbf{b}_k^\star$ is orthogonal to the subspace spanned by the primal basis vectors $\mathbf{b}_1, \ldots, \mathbf{b}_{k-1}, \mathbf{b}_{k+1}, \ldots, \mathbf{b}_m$. This is useful since for any $\mathbf{x} = \mathbf{B}\mathbf{z} \in \mathcal{L}$ we can recover the kth integer coefficient via $z_k = \mathbf{x}^T\mathbf{b}_k^\star$. The determinant of the dual lattice is easily seen to be given by $|\mathcal{L}^\star| = 1/|\mathcal{L}|$. The cubic lattice $\mathbb{Z}^n$ is an example of a lattice that is self-dual in the sense that $\mathcal{L} = \mathcal{L}^\star$.

MOTIVATING EXAMPLE

To illustrate the relevance of lattices and lattice reduction to MIMO communications, we consider a system transmitting two integer symbols $\mathbf{s} = \binom{s_1}{s_2} \in \mathbb{Z}^2$ over two transmit antennas (see also "Motivating Example: MIMO Detection"). This example is inspired by a similar one in [34]. At the receiver, two antennas receive the samples $y_1 = h_{1,1}\,s_1 + h_{1,2}\,s_2 + w_1$ and $y_2 = h_{2,1}s_1 + h_{2,2}\,s_2 + w_2$, where $h_{k,\ell}$ denotes the channel coefficient between receive antenna k and transmit antenna ℓ and w_1 and w_2 denotes noise. We rewrite the observation model as $\mathbf{y} = \binom{y_1}{y_2} = \mathbf{x} + \mathbf{w}$ with $\mathbf{x} = \binom{h_{1,1}}{h_{2,1}}s_1 + \binom{h_{1,2}}{h_{2,2}}s_2$ and $\mathbf{w} = \binom{w_1}{w_2}$. Clearly, the noiseless receive vectors $\mathbf{x}$ form a lattice with basis given by the channel matrix $\mathbf{H} = \begin{pmatrix} h_{1,1} & h_{1,2} \\ h_{2,1} & h_{2,2} \end{pmatrix}$, i.e., $\mathbf{x} \in \mathcal{L}(\mathbf{H})$.

If the elements of the noise vector are independent and identically distributed (i.i.d.) Gaussian, optimal ML detection amounts to finding the lattice point $\hat{\mathbf{x}}_{\mathrm{ML}}$ lying closest to $\mathbf{y}$. This leads to decision regions that equal the Voronoi regions of the lattice and hence are independent of the specific lattice basis. While ML detection is optimal, the search for the closest lattice point is computationally hard, especially in high dimensions. As a computationally simpler alternative, we may use a zero-forcing (ZF) detector that first equalizes the channel by computing $\tilde{\mathbf{s}}_{\mathrm{ZF}} = \mathbf{H}^{-1}\mathbf{y} = \mathbf{s} + \mathbf{H}^{-1}\mathbf{w}$ and then uses a simple component-wise slicer (quantizer) that delivers the element in $\mathbb{Z}^2$ closest to $\tilde{\mathbf{s}}_{\mathrm{ZF}}$. The square decision regions for $\tilde{\mathbf{s}}_{\mathrm{ZF}}$ correspond to decisions regions for $\mathbf{y}$ that are given by the fundamental parallelogram $\mathcal{P}(\mathbf{H})$ centered around each lattice point. Unless the channel is orthogonal, the ZF decision regions are different from the ML decision regions, therefore resulting in a larger number of detection errors.

The error probability of a detector is largely determined by the minimum distance of the lattice points from the boundaries of the associated decision region. This distance can be interpreted as the maximum amount of noise that can be tolerated without ending up in an incorrect decision region. This minimum distance is much smaller for ZF detection than for ML detection, thereby explaining the inferiority of ZF detection. The problem with the ZF detector can be seen in the fact that it is based on the (pseudo)inverse of the channel matrix $\mathbf{H}$ and that the component-wise slicer (quantizer) ignores the correlation introduced by the equalizer. This implies that ZF detection is not independent of the basis (induced by the channel matrix), which may often be rather unfavorable. The fundamental idea now is to apply ZF detection with respect to a different, improved basis $\tilde{\mathbf{H}} = \mathbf{H}\mathbf{T}$ (with unimodular $\mathbf{T}$). This is possible since $\mathbf{y} = \mathbf{H}\mathbf{s} + \mathbf{w} = \tilde{\mathbf{H}}\mathbf{z} + \mathbf{w}$ where $\mathbf{z} = \mathbf{T}^{-1}\mathbf{s} \in \mathbb{Z}^2$ due to the unimodularity of $\mathbf{T}$. Lattice-reduction aided ZF detection then performs equalization with the (pseudo)inverse of the improved basis $\tilde{\mathbf{H}}$, followed by component-wise integer quantization (yielding an estimate of $\mathbf{z}$) and multiplication by $\mathbf{T}$ (yielding an estimate of $\mathbf{s}$). This procedure is equivalent to decision regions that correspond to the fundamental parallelogram $\mathcal{P}(\tilde{\mathbf{H}})$. Performance is improved by lattice reduction since $\mathcal{P}(\tilde{\mathbf{H}})$ is more similar to the Voronoi region $\mathcal{V}(\mathcal{L})$ than $\mathcal{P}(\mathbf{H})$. While ZF detection with respect to a reduced basis is generally not equivalent to ML detection, lattice-reduction-aided ZF detection typically results in large performance gains (see the sections "Lattice Reduction for Data Detection" and "Lattice Reduction for Precoding").

COMPLEX-VALUED LATTICES

In many practical problems, e.g., in wireless communications, the quantities involved are complex valued. The previous discussion of real-valued point lattices can be generalized to the complex case in a more or less straightforward manner. Specifically, a complex-valued lattice of rank m in the n-dimension complex space $\mathbb{C}^n$ is defined as

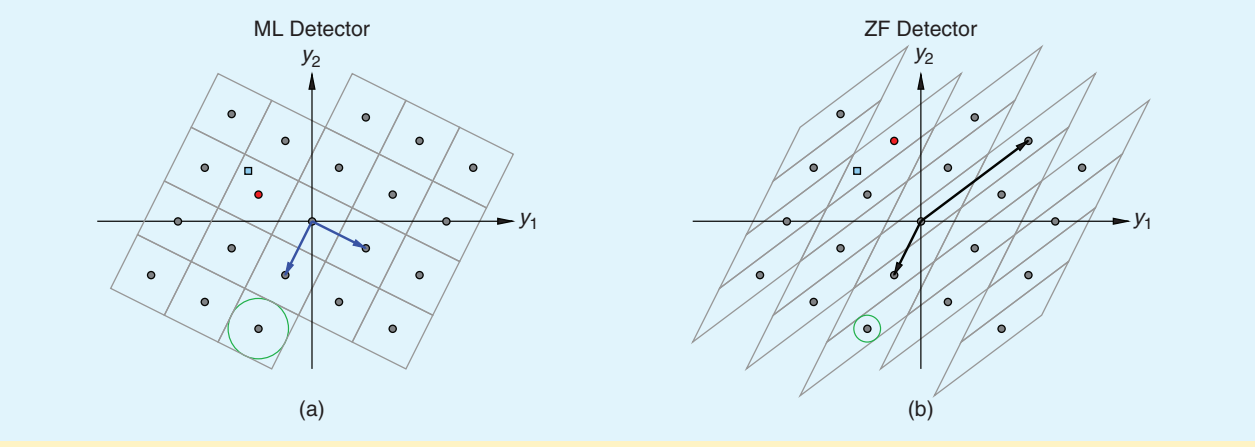
$$\mathcal{L} \triangleq \left\{ \mathbf{x} \middle| \mathbf{x} = \sum_{\ell=1}^{m} z_\ell \mathbf{b}_\ell, \quad z_\ell \in \mathbb{Z}_j \right\},$$

with complex basis vectors $\mathbf{b}_\ell \in \mathbb{C}^n$ and $\mathbb{Z}_j = \mathbb{Z} + j\mathbb{Z}$ denoting the set of complex integers (also known as Gaussian integers). By arranging the basis vectors into an $n \times m$ complex-valued matrix $\mathbf{B}$ and noticing that the complex mapping $\mathbf{x} = \mathbf{B}\mathbf{z}$ can be equivalently expressed as

MOTIVATING EXAMPLE: MIMO DETECTION

Figure S5 illustrates a rotated square receive lattice induced by the channel realization $\mathbf{H} = \begin{pmatrix} -1 & 4 \\ -2 & 3 \end{pmatrix}$ and integer symbols $\mathbf{s} \in \mathbb{Z}^2$. The channel matrix corresponds to the basis vectors $\mathbf{h}_1 = \begin{pmatrix} -1 \\ -2 \end{pmatrix}$ and $\mathbf{h}_2 = \begin{pmatrix} 4 \\ 3 \end{pmatrix}$ (shown in black in (b)), which are seen to be almost collinear. This means that $\mathbf{H}$ is poorly conditioned; its orthogonality defect equals $\xi(\mathbf{H}) = \sqrt{5}$. The ML decision regions, shown in (a), correspond to the Voronoi regions of the lattice and are rotated squares. For the specific receive vector shown as light-blue square, the closest lattice point is $\hat{\mathbf{x}}_{\text{ML}} = \begin{pmatrix} -2 \\ 1 \end{pmatrix}$ (marked as red circle) which corresponds to the ML decision $\hat{\mathbf{s}}_{\text{ML}} = \mathbf{H}^{-1}\hat{\mathbf{x}}_{\text{ML}} = \begin{pmatrix} -2 \\ -1 \end{pmatrix}$. With ZF detection, the decision regions are given by the shifted versions of the fundamental parallelogram $\mathcal{P}(\mathbf{H})$ associated with $\mathbf{H}$ [see (b)]. Due to the poor condition number of $\mathbf{H}$, the ZF and ML decision regions are markedly different. Indeed, the ZF estimates are given by $\hat{\mathbf{x}}_{\text{ZF}} = \begin{pmatrix} -1 \\ 3 \end{pmatrix}$ and $\hat{\mathbf{s}}_{\text{ZF}} = \begin{pmatrix} -3 \\ 1 \end{pmatrix} \neq \hat{\mathbf{s}}_{\text{ML}}$.

In both (a) and (b), green circles illustrate the detector's noise robustness, which is determined by the distance of the lattice points from the associated decision boundaries. In this example, a reduced basis is obtained as $\tilde{\mathbf{h}}_1 = \mathbf{h}_1$ and $\tilde{\mathbf{h}}_2 = \mathbf{h}_2 + 2\mathbf{h}_1$, i.e., $\tilde{\mathbf{H}} = \mathbf{HT}$ with $\mathbf{T} = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$. This new basis [shown in blue in (a)] is optimal in the sense that it consists of two orthogonal shortest vectors (in general, there is no guarantee that an orthogonal basis exists nor that it can be found via lattice reduction). The fundamental parallelogram associated with $\tilde{\mathbf{H}}$ is congruent with the Voronoi region of the lattice. Thus, ZF equalization with respect to $\tilde{\mathbf{H}}$ followed by a $\mathbb{Z}^2$ slicer and a remapping $\mathbf{z} \to \mathbf{s}$ here (but not in general) is equivalent to ML detection. Specifically, the closest lattice point $\tilde{\mathbf{H}}\mathbf{z} = \begin{pmatrix} -2 \\ 1 \end{pmatrix}$ corresponds to the (transformed) symbol vector $\mathbf{z} = \mathbf{T}^{-1}\mathbf{s} = \begin{pmatrix} 0 \\ -1 \end{pmatrix}$ and hence to the symbol decision $\mathbf{s} = \mathbf{Tz} = \begin{pmatrix} -2 \\ -1 \end{pmatrix}$, which coincides with the ML result.



[FIGS5] A rotated square receive lattice induced by the channel realization $\mathbf{H} = \begin{pmatrix} -1 & 4 \\ -2 & 3 \end{pmatrix}$ and integer symbols $\mathbf{s} \in \mathbb{Z}^2$. (a) ML decision regions and (b) decision regions of ZF detector.

$$\begin{pmatrix} \Re\{\mathbf{x}\} \\ \Im\{\mathbf{x}\} \end{pmatrix} = \begin{pmatrix} \Re\{\mathbf{B}\} & -\Im\{\mathbf{B}\} \\ \Im\{\mathbf{B}\} & \Re\{\mathbf{B}\} \end{pmatrix} \begin{pmatrix} \Re\{\mathbf{z}\} \\ \Im\{\mathbf{z}\} \end{pmatrix}, \tag{4}$$

it is seen that any m-dimensional complex-valued lattice in $\mathbb{C}^n$ can be dealt with as a $2m$-dimensional real-valued lattice in $\mathbb{R}^{2n}$. Yet, many of the concepts and algorithms from the real-valued space can be (and have been) formulated directly in the complex domain with minor modifications (see, e.g., [14] and [35]–40]). This is sometimes advantageous since the problem dimension is not increased, the basis matrix need not obey the structure in (4), and the algorithm complexity can be lower.

LATTICE REDUCTION TECHNIQUES

Lattice reduction techniques have a long tradition in mathematics in the field of number theory. The goal of lattice basis reduction is to find, for a given lattice, a basis matrix with favorable properties. Usually, such a basis consists of vectors that are short and therefore this basis is called reduced. Unless stated otherwise, the term "short" is to be interpreted in the usual Euclidean sense. There are several definitions of lattice reduction with corresponding reduction criteria, such as Minkowski reduction [41]–[43], Hermite-Korkine-Zolotareff reduction [44], [45], Gauss reduction [46], LLL reduction [2], [33], Seysen reduction [3], and Brun reduction [47], [48]. The corresponding lattice reduction algorithms yield reduced bases with shorter basis vectors and improved orthogonality; they provide a tradeoff between the quality of the reduced basis and the computational effort required for finding it.

In the following, we discuss the basics of the various lattice reduction approaches and the underlying reduction criteria. MATLAB implementations of some of the algorithms are provided as supplementary material in IEEE *Xplore* (http://ieeexplore.ieee.org). Since most of the lattice reduction techniques are based on an orthogonal decomposition of a lattice basis matrix,

QR AND GRAM-SCHMIDT

The unnormalized Gram-Schmidt orthogonalization of a basis $\mathbf{b}_1, \ldots, \mathbf{b}_m$ is described by the iterative basis updates

$$\hat{\mathbf{b}}_\ell = \mathbf{b}_\ell - \sum_{k=1}^{\ell-1} \mu_{\ell,k}\, \hat{\mathbf{b}}_k, \quad \text{with} \quad \mu_{\ell,k} = \frac{\mathbf{b}_\ell^T \hat{\mathbf{b}}_k}{\|\hat{\mathbf{b}}_k\|^2}.$$

In contrast, the (thin) QR decomposition represents the given basis vectors in terms of orthonormal basis vectors $\mathbf{q}_1, \ldots, \mathbf{q}_m$ as

$$\mathbf{b}_\ell = \sum_{k=1}^{\ell} r_{k,\ell}\, \mathbf{q}_k.$$

Some algebra reveals the following relations between QR decomposition and Gram-Schmidt orthogonalization:

- The ℓth column $\mathbf{q}_\ell$ of $\boldsymbol{Q}$ is the normalized version of the corresponding Gram-Schmidt vector $\hat{\mathbf{b}}_\ell$, i.e., $\mathbf{q}_\ell = \hat{\mathbf{b}}_\ell / \|\hat{\mathbf{b}}_\ell\|$.
- The length of the ℓth Gram-Schmidt vector equals the ℓth diagonal element of the triangular matrix $\mathbf{R}$, i.e., $\|\hat{\mathbf{b}}_\ell\| = r_{\ell,\ell}$.
- The off-diagonal elements of the triangular matrix $\mathbf{R}$ are related to the Gram-Schmidt coefficients according to $\mu_{\ell,k} = r_{k,\ell}/r_{k,k}$.

Using the above analogies, any lattice reduction technique formulated in terms of the QR decomposition can be rephrased using Gram-Schmidt orthogonalization and vice versa.

we first briefly review the QR decomposition. Some texts build on the (unnormalized) Gram-Schmidt orthogonalization instead of the QR decomposition (see "QR and Gram-Schmidt" for a discussion of their relation). In practical implementations, the QR decomposition is preferable since it can be performed efficiently and numerically more stable (using e.g., Givens rotations or Housholder transformations [49]) than Gram-Schmidt.

QR DECOMPOSITION

Consider an $n \times m$ lattice basis $\mathbf{B}$ with $n \geq m$ and $\text{rank}(\mathbf{B}) = m$. The (thin) QR decomposition factorizes $\mathbf{B}$ according to $\mathbf{B} = \mathbf{QR}$, where $\mathbf{Q} = (\mathbf{q}_1 \ldots \mathbf{q}_m)$ is an $n \times m$ column-orthogonal matrix and $\mathbf{R}$ is an $m \times m$ upper triangular matrix with positive diagonal elements. We denote the element of $\mathbf{R}$ in row k and column ℓ by $r_{k,\ell}$. Since $\mathbf{Q}$ has orthogonal columns with unit norm, we have $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}_m$. The QR decomposition amounts to expressing the ℓth column of $\mathbf{B}$ in terms of the orthonormal basis vectors $\mathbf{q}_1, \ldots, \mathbf{q}_\ell$ as

$$\mathbf{b}_\ell = \sum_{k=1}^{\ell} r_{k,\ell}\, \mathbf{q}_k.$$

Here, $\mathbf{q}_k^T \mathbf{b}_\ell = r_{k,\ell}$ characterizes the component of $\mathbf{b}_\ell$ collinear with $\mathbf{q}_k$. Furthermore, $r_{\ell,\ell}$ describes the component of $\mathbf{b}_\ell$ which is orthogonal to the space spanned by $\mathbf{b}_1, \ldots, \mathbf{b}_{\ell-1}$ or, equivalently, by $\mathbf{q}_1, \ldots, \mathbf{q}_{\ell-1}$.

The QR decomposition gives a descriptive explanation for the orthogonality of a basis. A basis vector $\mathbf{b}_\ell$ is almost orthogonal to the space spanned by $\mathbf{b}_1, \ldots, \mathbf{b}_{\ell-1}$, if the absolute values of $r_{1,\ell}, \ldots, r_{\ell-1,\ell}$ are close to zero. If these elements of $\mathbf{R}$ are exactly zero, $\mathbf{b}_\ell$ has no component into the direction of $\mathbf{b}_1, \ldots, \mathbf{b}_{\ell-1}$ and is correspondingly orthogonal to the space spanned by these vectors. However, for general lattices such a strictly orthogonal basis does not exist and one has to settle for a basis satisfying less stringent criteria.

BASIC APPROACH

As discussed in the section "Point Lattices," the columns of the matrix $\mathbf{B}$ define the lattice $\mathcal{L}$. The same lattice is also generated by any matrix that is constructed from $\mathbf{B}$ by the following elementary column operations:

1) *Reflection*: This transformation performs a sign change by multiplying a specific column by -1, i.e., $\widetilde{\mathbf{b}}_\ell = -\mathbf{b}_\ell$; the corresponding unimodular matrix reads

$$\mathbf{T}_{\mathrm{R}}^{(\ell)} = \mathbf{I}_m - 2\mathbf{e}_\ell \mathbf{e}_\ell^T. \tag{5}$$

2) *Swap*: Here, two columns of the basis matrix are interchanged. Swapping column k and ℓ according to $\widetilde{\mathbf{b}}_\ell = \mathbf{b}_k$ and $\widetilde{\mathbf{b}}_k = \mathbf{b}_\ell$ amounts to postmultiplication of the basis matrix with the unimodular matrix

$$\mathbf{T}_{\mathrm{S}}^{(k,\ell)} = \mathbf{I}_m - \mathbf{e}_k\mathbf{e}_k^T - \mathbf{e}_\ell\mathbf{e}_\ell^T + \mathbf{e}_k\mathbf{e}_\ell^T + \mathbf{e}_\ell\mathbf{e}_k^T. \tag{6}$$

3) *Translation:* This operation adds one column to another column, i.e., $\widetilde{\mathbf{b}}_\ell = \mathbf{b}_\ell + \mathbf{b}_k$. Such a translation is characterized by the unimodular matrix

$$\mathbf{T}_{\mathrm{T}}^{(k,\ell)} = \mathbf{I}_m + \mathbf{e}_k\mathbf{e}_\ell^T. \tag{7}$$

A translation by an integer multiple corresponds to repeated application of $\mathbf{T}_{\mathrm{T}}^{(k,\ell)}$, i.e., $\widetilde{\mathbf{b}}_\ell = \mathbf{b}_\ell + \mu\mathbf{b}_k$ with $\mu \in \mathbb{Z}$ is obtained via postmultiplication with $[\mathbf{T}_{\mathrm{T}}^{(k,\ell)}]^\mu = \mathbf{I}_m + \mu\mathbf{e}_k\mathbf{e}_\ell^T$. Subtraction of integer multiples of a column can be achieved by reflecting the corresponding column before and after the translation, i.e., $\widetilde{\mathbf{b}}_\ell = \mathbf{b}_\ell - \mu\mathbf{b}_k$ corresponds to the unimodular matrix $\mathbf{T}_{\mathrm{R}}^{(k)}[\mathbf{T}_{\mathrm{T}}^{(k,\ell)}]^\mu\mathbf{T}_{\mathrm{R}}^{(k)} = [\mathbf{T}_{\mathrm{T}}^{(k,\ell)}]^{-\mu} = \mathbf{I}_m - \mu\mathbf{e}_k\mathbf{e}_\ell^T$.

We note that only a translation actually impacts the quality of the basis directly, whereas reflections and column swaps help to systematically perform the appropriate translations. Since unimodular matrices form a group under multiplication, any sequential combination of the above three elementary operations corresponds to post-multiplying $\mathbf{B}$ with a specific unimodular matrix $\mathbf{T}$. The goal of lattice reduction algorithms is to determine a sequence of elementary column operations (equivalently, a unimodular matrix $\mathbf{T}$) that transforms the given basis $\mathbf{B}$ into a reduced basis $\widetilde{\mathbf{B}} = \mathbf{BT}$ according to the specific requirements of the corresponding reduction criterion. In the following, we discuss various lattice reduction approaches. Hereafter, the QR factors of $\widetilde{\mathbf{B}}$ will be denoted by $\widetilde{\mathbf{Q}}$ and $\widetilde{\mathbf{R}}$, i.e., $\widetilde{\mathbf{B}} = \widetilde{\mathbf{Q}}\widetilde{\mathbf{R}}$.

MINKOWSKI AND HERMITE-KORKINE-ZOLOTAREFF REDUCTION

H. Minkowski [41]–[43] developed the field of geometry of numbers and exploited the concept of point lattices to formulate the

corresponding theory. He introduced a very strong reduction criterion that requires that the first vector $\widetilde{\mathbf{b}}_1$ of the ordered basis $\widetilde{\mathbf{B}}$ is a shortest nonzero vector in $\mathcal{L}(\widetilde{\mathbf{B}})$. In the following, when we speak of "shortest vectors," this is meant to implicitly exclude the trivial all-zeros vector. All subsequent vectors $\widetilde{\mathbf{b}}_\ell$ for $2 \leq \ell \leq m$ have to be shortest vectors such that the set of vectors $\widetilde{\mathbf{b}}_1, \ldots, \widetilde{\mathbf{b}}_\ell$ can be extended to a basis of $\mathcal{L}(\widetilde{\mathbf{B}})$. Thus, $\widetilde{\mathbf{b}}_\ell$ is a shortest vector in $\mathcal{L}(\widetilde{\mathbf{B}})$ that is not a linear combination of $\widetilde{\mathbf{b}}_1, \ldots, \widetilde{\mathbf{b}}_{\ell-1}$.

The definition of a Hermite-Korkine-Zolotareff-reduced basis is related to that of Minkowski. It requires that the projection of the basis vectors $\widetilde{\mathbf{b}}_\ell$ onto the orthogonal complement of the space spanned by $\{\widetilde{\mathbf{b}}_1, \ldots, \widetilde{\mathbf{b}}_{\ell-1}\}$ are the shortest vectors of the corresponding projected lattices [44], [45]. Consequently, the first vector $\widetilde{\mathbf{b}}_1$ is again a shortest vector of $\mathcal{L}(\widetilde{\mathbf{B}})$.

Due to their high computational complexity (see the section "Complexity of Lattice Reduction"), the Minkowski and Hermite-Korkine-Zolotareff reductions will not be considered in the context of MIMO detection and precoding (cf. the sections "Lattice Reduction for Data Detection" and "Lattice Reduction for Precoding").

SIZE REDUCTION

A rather simple but not very powerful criterion is given by the so-called size reduction. A basis $\widetilde{\mathbf{B}}$ is called size reduced if the elements of the corresponding upper triangular matrix $\widetilde{\mathbf{R}}$ satisfy the condition

$$|\widetilde{r}_{k,\ell}| \leq \tfrac{1}{2}\,|\widetilde{r}_{k,k}| \quad \text{for } 1 \leq k < \ell \leq m\,. \tag{8}$$

Thus, the component of any vector $\widetilde{\mathbf{b}}_\ell$ into the direction of $\widetilde{\mathbf{q}}_k$ for $k < \ell$ is not longer than half of the length of $\widetilde{\mathbf{b}}_k$ perpendicular to $\text{span}\{\widetilde{\mathbf{q}}_1, \ldots, \widetilde{\mathbf{q}}_{k-1}\}$. If this condition is not fulfilled for an index pair (k,ℓ), the length of $\widetilde{\mathbf{b}}_\ell$ can be reduced by subtracting a multiple of $\widetilde{\mathbf{b}}_k$ where the integer multiplication factor is given by $\mu = \lceil \widetilde{r}_{k,\ell}/\widetilde{r}_{k,k} \rfloor$ with $\lceil \cdot \rfloor$ denoting the rounding operation. This subtraction corresponds to applying the unimodular translation matrix $[\mathbf{T}_{\mathrm{T}}^{(k,\ell)}]^{-\mu}$ [cf. (7)]. We note that size reduction does not involve any column swaps. A basis fulfilling (8) is often called weakly reduced.

GAUSS REDUCTION

In contrast to the other reduction approaches, the reduction method introduced by C.F. Gauss in the context of binary quadratic forms is restricted to lattices of rank $m = 2$, i.e., $\mathbf{B} = (\mathbf{b}_1\ \mathbf{b}_2) \in \mathbb{R}^{n\times 2}$ [46]. For such two-dimensional lattices, Gauss reduction constructs a basis that fulfills the reduction criteria introduced by Minkowski and Hermite-Korkine-Zolotareff.

In addition to size reduction, Gauss reduction also includes column swapping operations that correspond to applying the unimodular matrices $\mathbf{T}_{\mathrm{S}}^{(k,\ell)}$ defined in (6). In particular, after first size reducing the given basis matrix $\mathbf{B}$, the columns of the resulting basis $\widetilde{\mathbf{B}} = (\widetilde{\mathbf{b}}_1\ \widetilde{\mathbf{b}}_2)$ are swapped if the length of $\widetilde{\mathbf{b}}_1$ is larger than that of $\widetilde{\mathbf{b}}_2$ and the resulting basis is again size reduced. This process of successive size reduction and column swapping operations is repeated until the length of $\widetilde{\mathbf{b}}_1$ is shorter—after the preceding size reduction step—than that of $\widetilde{\mathbf{b}}_2$, which implies that no further column swapping operation is performed (see "Gauss Reduction"). After a finite number of iterations, this algorithm provides a Gauss-reduced basis $\widetilde{\mathbf{B}}$, where $\|\widetilde{\mathbf{b}}_1\| \leq \|\widetilde{\mathbf{b}}_2\|$ and the properties of a size-reduced basis are satisfied. In particular, $\widetilde{\mathbf{b}}_1$ and $\widetilde{\mathbf{b}}_2$ are the two shortest vectors in the lattice $\mathcal{L}$ that form a basis for $\mathcal{L}$. "Example: Size and Gauss Reduction" illustrates these two reduction techniques with a simple example.

GAUSS REDUCTION

Below, we provide pseudocode for Gauss reduction. The algorithm takes a two-dimensional basis matrix $\mathbf{B}$ as input and successively performs column swaps and size reductions until a Gauss reduced basis is obtained.

1: $\widetilde{\mathbf{B}} \leftarrow \mathbf{B}$
2: **repeat**
3: $\quad \widetilde{\mathbf{b}}_1 \leftrightarrow \widetilde{\mathbf{b}}_2$
4: $\quad \widetilde{\mathbf{b}}_2 \leftarrow \widetilde{\mathbf{b}}_2 - \left\lceil \dfrac{\widetilde{\mathbf{b}}_1^T \widetilde{\mathbf{b}}_2}{\|\widetilde{\mathbf{b}}_1\|} \right\rfloor \widetilde{\mathbf{b}}_1$
5: **until** $\|\widetilde{\mathbf{b}}_1\| \leq \|\widetilde{\mathbf{b}}_2\|$

LLL REDUCTION

A powerful and famous reduction criterion for arbitrary lattice dimensions was introduced by A.K. Lenstra, H.W. Lenstra, and L. Lovász in [2], and the algorithm they proposed is known as the LLL (or L^3) algorithm (see also [29] and [33]). It can be interpreted as an extension of Gauss reduction to lattices of rank $m > 2$.

In particular, a basis $\widetilde{\mathbf{B}}$ with QR decomposition $\widetilde{\mathbf{B}} = \widetilde{\mathbf{Q}}\widetilde{\mathbf{R}}$ is called LLL reduced with parameter $1/4 < \delta \leq 1$, if

$$|\widetilde{r}_{k,\ell}| \leq \tfrac{1}{2}|\widetilde{r}_{k,k}|, \ \text{for } 1 \leq k < \ell \leq m, \tag{9a}$$

$$\delta\,|\widetilde{r}_{\ell-1,\ell-1}|^2 \leq |\widetilde{r}_{\ell,\ell}|^2 + |\widetilde{r}_{\ell-1,\ell}|^2, \ \text{for } \ell = 2, \ldots, m. \tag{9b}$$

The choice of the parameter δ affects the quality of the reduced basis and the computational complexity. Larger δ results in a better basis at the price of higher complexity (see also the section "LLL Complexity"). A common choice is $\delta = 3/4$. The inequality in (9a) is the condition for a size-reduced basis [cf. (8)] and (9b) is the so-called Lovász condition. The quantities $|\widetilde{r}_{\ell-1,\ell-1}|^2$ and $|\widetilde{r}_{\ell,\ell}|^2 + |\widetilde{r}_{\ell-1,\ell}|^2$ in the Lovász condition equal the squared lengths of the components of $\widetilde{\mathbf{b}}_{\ell-1}$ and $\widetilde{\mathbf{b}}_\ell$ that are orthogonal to the space spanned by $\widetilde{\mathbf{b}}_1, \ldots, \widetilde{\mathbf{b}}_{\ell-2}$. If the Lovász condition is not fulfilled for two vectors $\widetilde{\mathbf{b}}_{\ell-1}$ and $\widetilde{\mathbf{b}}_\ell$, these vectors are swapped (similar to Gauss reduction) and the resulting basis matrix is again QR-decomposed and size reduced. This process of subsequent size reduction and column swapping operations [which corresponds to the application of the reflections, swaps, and translations defined in (5)–(7)] is repeated until—after a finite number of iterations—an LLL-reduced basis is obtained (see "The LLL Algorithm" and [2],

EXAMPLE: SIZE AND GAUSS REDUCTION

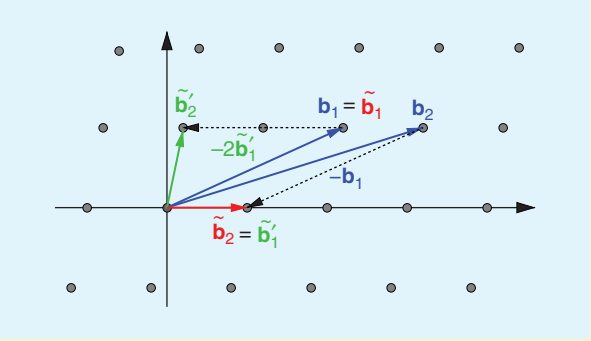
We illustrate size reduction and Gauss reduction with a small example. Figure S6 shows a two-dimensional lattice $\mathcal{L}$ spanned by the basis vectors $\mathbf{b}_1 = \begin{pmatrix} 2.2 \\ 1 \end{pmatrix}$ and $\mathbf{b}_2 = \begin{pmatrix} 3.2 \\ 1 \end{pmatrix}$ (shown in blue). The orthogonality defect of this basis equals $\xi(\mathbf{B}) = 8.1$.

Size Reduction

The QR decomposition of this basis yields $\mathbf{R} = \begin{pmatrix} 2.417 & 3.327 \\ 0 & 0.414 \end{pmatrix}$. Since $|r_{1,2}| = 3.327 > |r_{1,1}|/2 = 1.208$, the basis is not size reduced. The corresponding size reduction step is to replace $\mathbf{b}_2$ with $\tilde{\mathbf{b}}_2 = \mathbf{b}_2 - \mu\mathbf{b}_1$ where $\mu = \lceil 3.327/2.417 \rfloor = 1$, leading to the new basis vector $\tilde{\mathbf{b}}_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ (shown in red). The first basis vector remains unchanged, i.e., $\tilde{\mathbf{b}}_1 = \mathbf{b}_1$. Since $\tilde{r}_{1,2} = r_{1,2} - r_{1,1} = 0.91 < |r_{1,1}|/2 = 1.208$, the new basis is size reduced; its orthogonality defect is $\xi(\tilde{\mathbf{B}}) = 2.42$.

Gauss Reduction

After the size reduction, $\tilde{\mathbf{b}}_1$ is more than twice as long as $\tilde{\mathbf{b}}_2$. This suggests to perform a column swap $\tilde{\mathbf{b}}_1 \leftrightarrow \tilde{\mathbf{b}}_2$ that leads to the basis matrix $\begin{pmatrix} 1 & 2.2 \\ 0 & 1 \end{pmatrix}$ that is actually upper triangular. Since $|\tilde{r}_{1,2}| = 2.2 > |\tilde{r}_{1,1}|/2 = 1/2$, we perform another size reduction step, i.e., $\tilde{\mathbf{b}}_2' = \tilde{\mathbf{b}}_2 - \lceil 2.2/1 \rfloor\, \tilde{\mathbf{b}}_1 = \begin{pmatrix} 0.2 \\ 1 \end{pmatrix}$ and $\tilde{\mathbf{b}}_1' = \tilde{\mathbf{b}}_2$. Since $\tilde{\mathbf{b}}_1'$ is shorter than $\tilde{\mathbf{b}}_2'$ no further column swap or size reduction is possible and a Gauss-reduced basis has been obtained, i.e., the basis $\tilde{\mathbf{B}}' = \begin{pmatrix} 1 & 0.2 \\ 0 & 1 \end{pmatrix}$ consists of two shortest vectors that span $\mathcal{L}$. The orthogonality defect of this basis is $\xi(\tilde{\mathbf{B}}') = 1.02$.



[FIGS6] Illustration of size reduction (red) and Gauss reduction (green) for a two-dimensional lattice $\mathcal{L}$ spanned by the basis vectors $\mathbf{b}_1 = \begin{pmatrix} 2.2 \\ 1 \end{pmatrix}$ and $\mathbf{b}_2 = \begin{pmatrix} 3.2 \\ 1 \end{pmatrix}$ (shown in blue).

We note that the above sequence of size reduction, column swap, and size reduction can be described in terms of a unimodular matrix $\mathbf{T}$ that decomposes into the elementary column operations [see (5)–(7)]

$$\mathbf{T} = \mathbf{T}_{\mathrm{R}}^{(1)}\,\mathbf{T}_{\mathrm{T}}^{(1,2)}\,\mathbf{T}_{\mathrm{R}}^{(1)}\,\mathbf{T}_{\mathrm{S}}^{(1,2)}\,\mathbf{T}_{\mathrm{R}}^{(1)}\,\mathbf{T}_{\mathrm{T}}^{(1,2)}\,\mathbf{T}_{\mathrm{T}}^{(1,2)}\,\mathbf{T}_{\mathrm{R}}^{(1)}$$
$$= \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & -2 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} -1 & 3 \\ 1 & -2 \end{pmatrix}.$$

[50], and [51] for more details). LLL reduction results in many interesting properties of the corresponding reduced basis $\tilde{\mathbf{B}}$. For example, it can be shown that (9a) and (9b) imply that the length of the shortest vector $\tilde{\mathbf{b}}_1$ in an LLL-reduced basis is upper bounded in terms of the length of a shortest vector $\mathbf{x}_{\min}$ in the lattice, i.e., $\|\tilde{\mathbf{b}}_1\| \leq \sqrt{\alpha^{m-1}}\|\mathbf{x}_{\min}\|$ with $\alpha = \left(\delta - 1/4\right)^{-1}$ (it was observed that in practice LLL reduction provides a much better approximation of the shortest vector than suggested by this bound). Furthermore, the orthogonality defect of an LLL-reduced basis is bounded (though not necessarily minimal), i.e., $\xi(\tilde{\mathbf{B}}) \leq \sqrt[4]{\alpha^{m(m-1)}}$. This property has been used in [18] to show that suboptimum detectors aided by LLL lattice reduction achieve full diversity (see also [14] and [19]–[23]). We refer to the section "LLL Complexity" for a discussion of the complexity of LLL.

DUAL LATTICE REDUCTION

Instead of applying a particular lattice reduction algorithm to the basis $\mathbf{B}$, a reduction of the dual basis $\mathbf{B}^\star$ defined in (3) can be performed. We note that a reduction of $\mathbf{B}^\star$ by a unimodular matrix $\mathbf{T}$ corresponds to a reduction of the primal basis $\mathbf{B}$ by the unimodular matrix $\mathbf{T}^{-1}$. For lattice reduction based on the LLL algorithm, this dual lattice-reduction approach was proposed in [18] to improve the performance of lattice-reduction-based linear equalization schemes (see the section "Lattice Reduction for Data Detection").

SEYSEN'S LATTICE REDUCTION ALGORITHM

The basic principle of Seysen's lattice reduction algorithm [3], [52] lies in the simultaneous reduction of the basis $\mathbf{B}$ and the dual basis $\mathbf{B}^\star$. M. Seysen defined the orthogonality criterion

$$S(\mathbf{B}) = \sum_{\ell=1}^{m} \|\mathbf{b}_\ell\|^2 \|\mathbf{b}_\ell^\star\|^2,$$

which achieves its minimum, $S(\mathbf{B}) = m$, if and only if the basis $\mathbf{B}$ is orthogonal. A basis $\mathbf{B}$ is called S-reduced if $S(\mathbf{B}) \leq S(\mathbf{BT})$ holds for all possible unimodular transformation matrices $\mathbf{T}$. Thus, for an S-reduced basis, no unimodular transformation matrix $\mathbf{T}$ leads to a further reduction of the corresponding Seysen's measure. The determination of an S-reduced basis in general is computationally too expensive. Hence, one typically considers relaxed versions of lattice reduction based on Seysen's measure that successively perform elementary column operations with respect to just two basis vectors [cf. (5)–(7)] to find a local minimum of $S(\mathbf{B})$. The greedy variant that selects the two columns involved in the basis update such that $S(\mathbf{B})$ is maximally decreased is referred to as S_2 reduction. A MATLAB implementation of S_2 reduction is provided as supplementary material in IEEE *Xplore* (http://ieeexplore.ieee.org). The application of S_2 reduction for data detection has recently been proposed in [36] (see also the section "Lattice Reduction for Data Detection").

BRUN'S ALGORITHM

In 1919, V. Brun [47], [48] proposed an efficient algorithm for finding approximate integer relations (see also [8]), i.e., finding

THE LLL ALGORITHM

We provide pseudocode for the LLL algorithm, summarizing the main algorithmic steps. The algorithm inputs are the basis matrix $\mathbf{B}$ and the reduction parameter δ.

1: $\widetilde{\mathbf{B}} \leftarrow \mathbf{B}$
2: $[\widetilde{\mathbf{Q}}, \widetilde{\mathbf{R}}] \leftarrow \mathrm{qr}(\widetilde{\mathbf{B}})$
3: $\ell \leftarrow 2$
4: **repeat**
5: $\quad \widetilde{\mathbf{b}}_\ell \leftarrow \widetilde{\mathbf{b}}_\ell - \left\lceil \dfrac{\widetilde{r}_{\ell-1,\ell}}{\widetilde{r}_{\ell-1,\ell-1}} \right\rfloor \widetilde{\mathbf{b}}_{\ell-1}$
6: $\quad$ **if** $\delta\,|\widetilde{r}_{\ell-1,\ell-1}|^2 > |\widetilde{r}_{\ell,\ell}|^2 + |\widetilde{r}_{\ell-1,\ell}|^2$ **then**
7: $\quad\quad \widetilde{\boldsymbol{b}}_{\ell-1} \leftrightarrow \widetilde{\boldsymbol{b}}_\ell$
8: $\quad\quad \ell \leftarrow \max(\ell - 1, 2)$
9: $\quad$ **else**
10: $\quad\quad$ **for** $k = \ell - 2$ to 1 **do**
11: $\quad\quad\quad \widetilde{\mathbf{b}}_\ell \leftarrow \widetilde{\mathbf{b}}_\ell - \left\lceil \dfrac{\widetilde{r}_{k,\ell}}{\widetilde{r}_{k,k}} \right\rfloor \widetilde{\boldsymbol{b}}_k$
12: $\quad\quad$ **end for**
13: $\quad\quad \ell \leftarrow \ell + 1$
14: $\quad$ **end if**
15: **until** $\ell > m$

In lines 5 and 11, a size reduction is performed, whereas line 7 implements a column swap. In the above pseudocode, we omitted the updates of the QR decomposition, which essentially consist of simple Givens rotations and are necessary after each basis change (lines 5, 7, and 11). Actual implementations of the LLL algorithm (e.g., [50]) operate directly on the QR factors and additionally provide the corresponding unimodular matrix $\mathbf{T}$ as output. MATLAB code for such an implementation is made available as supplementary material in IEEE *Xplore* (http://ieeexplore.ieee.org).

integer vectors $\mathbf{t}_\ell \in \mathbb{Z}^n$ that are (almost) orthogonal to a given vector $\mathbf{u} \in \mathbb{R}^n$ while being as short as possible. It has been realized in [35] that Brun's algorithm can also be used for lattice reduction. In particular, let us consider a transformed bases $\widetilde{\mathbf{B}} = (\widetilde{\mathbf{b}}_1 \ldots \widetilde{\mathbf{b}}_m) = \mathbf{B}\mathbf{T}$ where $\mathbf{T} = (\mathbf{t}_1 \ldots \mathbf{t}_m)$ is unimodular. The orthogonality defect of $\widetilde{\mathbf{B}}$ is determined by the lengths of the basis vectors $\widetilde{\mathbf{b}}_\ell$, which equal

$$\|\widetilde{\mathbf{b}}_\ell\|^2 = \|\mathbf{B}\mathbf{t}_\ell\|^2 = \sum_{k=1}^{m} \lambda_k (\mathbf{u}_k^T \mathbf{t}_\ell)^2, \tag{10}$$

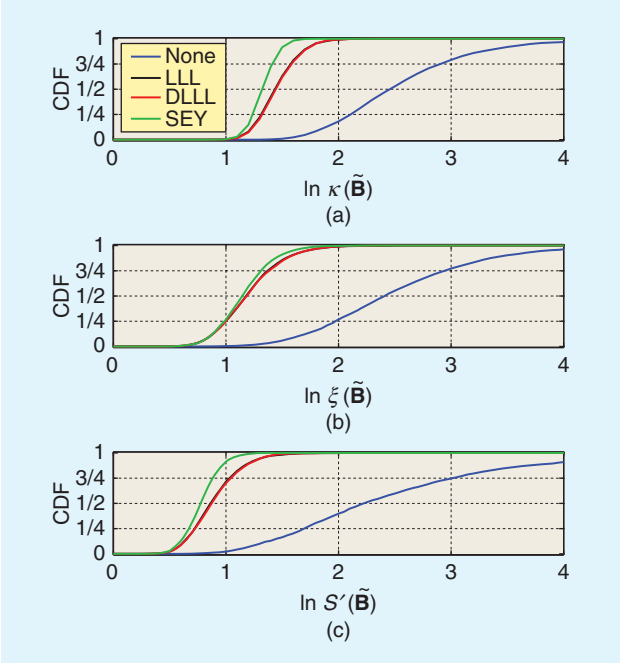
where $\lambda_k > 0$ and $\mathbf{u}_k$, $k = 1, \ldots, m$, denote the eigenvalues and eigenvectors of the Gramian $\mathbf{B}^T\mathbf{B}$. Assuming that λ_1 is the largest eigenvalue and that $\lambda_1 \gg \lambda_k$, $k \neq 1$, (10) suggests that decreasing the orthogonality defect by reducing $\|\widetilde{\mathbf{b}}_\ell\|^2$ requires $\mathbf{t}_\ell$ to be as orthogonal as possible to $\mathbf{u}_1$; this is exactly the approximate integer relation problem solved by Brun's algorithm. The assumption $\lambda_1 \gg \lambda_k$ is particularly well justified for precoding in wireless MIMO systems; see the section "Brun's Algorithm for Lattice-Reduction-Assisted Precoding."

Lattice reduction based on Brun's algorithm performs iterative column updates (translations/size reductions) such that $|\mathbf{u}_1^T \mathbf{t}_\ell|$ decreases (the column pair and the size reduction parameter can be determined very efficiently). This process is repeated as long as the orthogonality defect of the corresponding reduced basis decreases. Compared to LLL and Seysen reduction, Brun reduction performs poorer but has significantly lower complexity. A MATLAB implementation of Brun reduction is available as supplementary material in IEEE *Xplore* (http://ieeexplore.ieee.org).

REDUCTION PERFORMANCE

In the following, the quality of the different reduction schemes is compared by means of the condition number $\kappa(\widetilde{\mathbf{B}})$ (the ratio of the maximum and the minimum singular value of $\widetilde{\mathbf{B}}$), the orthogonality defect $\xi(\widetilde{\mathbf{B}})$, and the normalized Seysen criterion $S'(\widetilde{\mathbf{B}}) = S(\widetilde{\mathbf{B}})/m$. These three performance metrics are lower-bounded by one and should be as small as possible. The comparison here does not take into account the complexity of the various lattice reduction methods. While the preceding discussion of lattice reduction methods was restricted to real-valued basis matrices, we here compare the respective extensions to complex-valued basis matrices (cf. the section "Complex-Valued Lattices") for the case of complex Gaussian matrices whose elements are i.i.d. with zero mean and unit variance. For the LLL algorithm, we used the common choice $\delta = 3/4$. Better results at the expense of higher complexity can be achieved with larger δ.

For matrices of dimension 6×6, Figure 2 shows the cumulative distribution functions (CDFs) of the different performance



[FIG2] Performance assessment of LLL, dual LLL, and Seysen S_2 reduction in terms of the CDF of (a) $\ln \kappa(\widetilde{\mathbf{B}})$, (b) $\ln \xi(\widetilde{\mathbf{B}})$, and (c) $\ln S'(\widetilde{\mathbf{B}})$ (i.i.d. complex Gaussian basis matrices with zero mean and unit variance, $m = n = 6$). LLL and dual LLL lie practically on top of each other.

metrics (on a log scale) for the LLL-reduced basis, for the S_2-reduced basis (labeled "SEY"), for dual LLL ("DLLL"—LLL applied to the dual basis $\mathbf{B}^{\star}$), and for the unreduced basis. All lattice reduction algorithms scale the three performance metrics down significantly. Furthermore, SEY marginally outperforms LLL and DLLL (which perform identically).

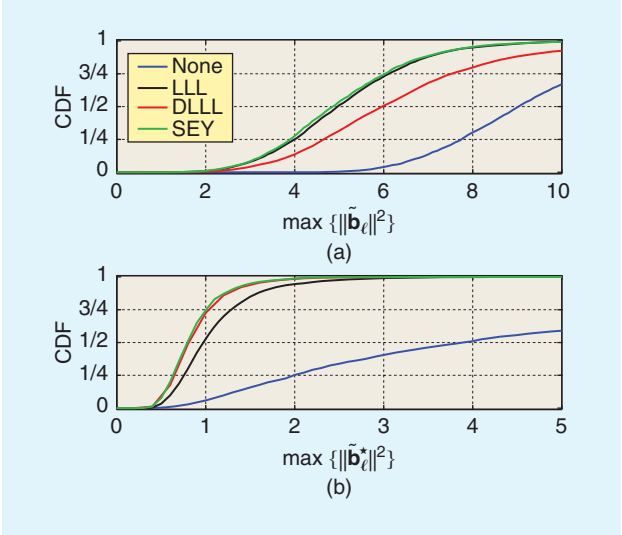
Figure 3 shows the CDFs of the squared length of the longest primal basis vector $\widetilde{\mathbf{b}}_\ell$ and of the longest dual basis vector $\widetilde{\mathbf{b}}_\ell^{\star}$. It is seen that for the primal basis SEY and LLL perform identically and superior to DLLL; in contrast, for the dual basis SEY and DLLL achieve equivalent results with LLL being inferior. This conforms that SEY reduces the primal and dual basis simultaneously while LLL only reduces the primal basis and DLLL only reduces the dual basis.

The metrics considered in this section give an indication of the performance of the various algorithms; they are not necessarily in one-to-one correspondence, though, with other performance metrics like bit error rate (BER) in a communication system.

COMPLEXITY OF LATTICE REDUCTION

COMPLEXITY OF MINKOWSKI AND HERMITE-KORKINE-ZOLOTAREFF REDUCTION

The Minkowski and Hermite-Korkine-Zolotareff reductions are the strongest but also the computationally most demanding to obtain. In both the Minkowski and the Hermite-Korkine-Zolotareff-reduced lattice basis the first vector $\widetilde{\mathbf{b}}_1$ of the reduced basis $\widetilde{\mathbf{B}}$ corresponds to the shortest vector in the lattice. This implies that the computation of the Minkowski and Hermite-Korkine-Zolotareff-reduced bases are at least as complex as the computation of the shortest lattice vector, a problem known to be NP-hard (under randomized reductions) [53]. In fact, the computation of a Hermite-Korkine-Zolotareff-reduced basis essentially consists of solving m shortest vector problems, and any method for computing the shortest lattice vector is thus applicable for use in the Hermite-Korkine-Zolotareff reduction [4]. The sphere decoder [4] may be used to solve the shortest vector problems. Since the sphere decoder has a (worst case and average) complexity that grows exponentially with the dimension m of the lattice [54], it is practically feasible only for lattices of moderate dimension. In this context, it is interesting to note that the sphere decoding algorithm of Fincke and Pohst was in fact presented as a solution to the shortest vector problem in [5], rather than the closest vector problem to which it is typically applied in the context of detection and precoding.



[FIG3] The CDF of maximum squared length of reduced basis vectors for (a) the primal basis and (b) the dual basis (i.i.d. complex Gaussian basis matrices with zero mean and unit variance, $m = n = 6$).

LLL COMPLEXITY

The complexity of the LLL algorithm is traditionally assessed by counting the number of LLL iterations, denoted K, where an iteration is started whenever the Lovász condition in (9b) is tested (this corresponds to one repeat loop in the pseudocode provided in "The LLL Algorithm"). Thus, an iteration may consist of the selection of a new pair of columns to operate on, or a column swap followed by a size reduction (cf. the section "LLL Reduction"). The number of LLL iterations required to reduce a given basis was treated already in [2], which made clever use of the quantity

$$D \triangleq \prod_{k=1}^{n-1} \prod_{\ell=1}^{k} |\widetilde{r}_{\ell,\ell}|^2.$$

It can be shown that D is decreased by (at least) a multiplicative factor δ [cf. (9b)] whenever two columns are swapped by the LLL algorithm and left unchanged by the size reduction. If the original basis matrix $\mathbf{B}$ is integer-valued, then $D \geq 1$ throughout the execution of the LLL algorithm [2], which in turn implies that the number of swap operation carried out by the LLL algorithm is finite and that the algorithm always terminates. In fact, given integer basis vectors of bounded Euclidean length V, the algorithm terminates after at most

$$K \leq m^2 \log_t V + m$$

iterations [2], [55], where $t \triangleq 1/\sqrt{\delta} > 1$. This also implies that the complexity (in terms of binary operations) of the LLL algorithm is polynomial in the (binary) description length of the basis $\mathbf{B}$—a widely celebrated result. For an integer-valued basis $\mathbf{B}$, the binary description length is the number of bits required to specify the elements of $\mathbf{B}$. A similar statement can be made also in the extreme case where $\delta = 1$, although the proof of this statement requires some additional work [56].

The complexity analysis for the case of general (real- or complex-valued) $\mathbf{B}$ is complicated by the fact that it is no longer necessarily true that $D \geq 1$. However, other strictly positive lower (and upper) bounds on D can be found in this case. Based on such bounds, it was shown in [55] that the number of LLL iterations in general is upper bounded as

$$K \leq m^2 \log_t \frac{A}{a} + m, \tag{11}$$

where $A \triangleq \max_\ell |\widetilde{r}_{\ell,\ell}|$ and $a \triangleq \min_\ell |\widetilde{r}_{\ell,\ell}|$. The bound (11) implies that the LLL algorithm terminates for arbitrary bases. The expression in (11) was used in [55] to prove polynomial average complexity of the LLL algorithm for the case where the basis vectors are uniformly distributed inside the unit sphere in $\mathbb{R}^m$. This result has subsequently been extended to i.i.d. real- and complex-valued Gaussian bases in the context of MIMO communications [40], [57].

LATTICE REDUCTION IS EXTREMELY USEFUL FOR DETECTION AND PRECODING IN WIRELESS MIMO SYSTEMS.

It is possible to further upper bound (11) based on the condition number of B according to [57]

$$K \leq m^2 \log_t \kappa(\mathbf{B}) + m. \tag{12}$$

This bound is valuable specifically because it applies to the LLL reduction of both the primal and the dual basis due to $\kappa(\mathbf{B}) = \kappa(\mathbf{B}^\star)$. It further follows from (12) that a large number of LLL iterations can occur only when the original basis matrix B is poorly conditioned. Although the opposite is not true (there are arbitrarily poorly conditioned bases that are simultaneously LLL reduced), it has been shown (by explicit construction of corresponding bases) that the number of LLL iterations can be arbitrarily large [57]. This means that there is no universal upper bound on the number of LLL iterations in the MIMO context, or equivalently that the worst-case complexity is unbounded. This said, it should be noted that (in addition to the polynomial average complexity) the probability mass function of K has an exponential tail [57], which implies that atypically large values of K are rare. Numerical experiments also indicate that the complexity of the algorithm is relatively low in practice.

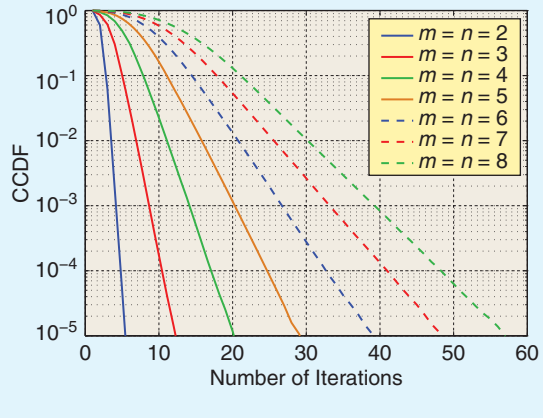
COMPLEXITY OF SEYSEN AND BRUN REDUCTION

There are much fewer analytical results for the complexity of Seysen and Brun reduction than for the LLL algorithm. That the number of S_2 reductions performed by Seysen's algorithm must be finite follows by [3, Corollary 9] although the bound given there is too loose to be useful in a practical complexity analysis. There is, to our knowledge, no similar statement available for Brun's algorithm. It is likely that an argument similar to [3] could be used to show that Brun's algorithm also always terminates. It is also likely that (similar to the LLL algorithm) there are bases that require an arbitrarily large number of iterations to reduce, both for Brun's algorithm and Seysen's algorithm.

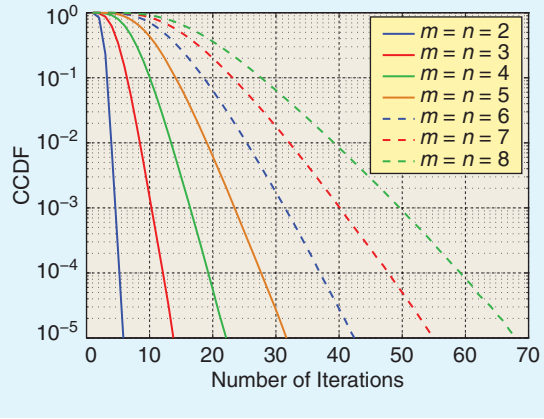
Like the LLL algorithm, Seysen's and Brun's algorithm have a data-dependent complexity, which makes numerical studies of their complexity relevant. Such investigations have found that Brun's algorithm tends to be an order of magnitude less complex than LLL and Seysen's algorithm (at the expense of reduced performance) [35]. The overall complexity (in terms of floating point operations) of reducing real-valued bases is slightly lower for the LLL algorithm than for Seysen's algorithm [58]. No comparable analysis has been made for the complex-valued case.

NUMERICAL RESULTS

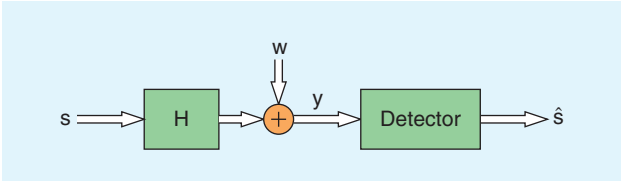
We illustrate the effective complexity of LLL and Seysen lattice reduction for random $n \times m$ bases whose elements are i.i.d. complex Gaussian with zero mean and unit variance. In the remainder of this article, such matrices will, for brevity, be referred to as i.i.d. Gaussian matrices. Figures 4 and 5 show the complementary cumulative distribution function (CCDF) of the number of (complex) LLL and Seysen iterations. For Seysen reduction, we used the greedy S_2 implementation in which a pair of basis vectors is always selected so that the largest reduction in the Seysen measure over any pair of two basis vectors is obtained [3], [52]; here, an iteration is defined as the selection of a pair of basis vectors, followed by their S_2 reduction. In Figure 4, one can clearly see that the probability of experiencing an atypically large



[FIG4] The CCDF of the number of LLL iterations when reducing $n \times m$ random bases with i.i.d. complex Gaussian elements with zero mean and unit variance.



[FIG5] The CCDF of the number of Seysen iterations when reducing $n \times m$ random bases with i.i.d. complex Gaussian elements with zero mean and unit variance.



[FIG6] Block diagram for a MIMO spatial multiplexing system.

number of LLL iterations vanishes exponentially fast, and also that the number of iterations increases with the dimension of the basis. The number of Seysen iterations shows a remarkable similarity to the LLL case; this observation has yet to be confirmed analytically.

TOPICS FOR FUTURE RESEARCH

There are a number of open problems concerning the complexity of lattice reduction. One is the lack of analytical results regarding the complexity of Seysen's and Brun's algorithm. Another important open problem is the development of better (probabilistic) bounds on the complexity of the LLL algorithm. In essence, bounds of the form (11) and (12) refer to the worst case complexity over some set of matrices, e.g., a set of matrices with bounded condition number. However, as noted previously, given a specific condition number there are LLL-reduced bases with this condition number. This means that for some matrices the bound in (12) is overly pessimistic. Numerical evidence also indicates that the average complexity is significantly lower than what is suggested by currently available bounds.

LATTICE REDUCTION FOR DATA DETECTION

The application of lattice reduction for efficient near-optimum data detection in MIMO wireless systems has attracted a lot of attention over recent years (see "MIMO Wireless" and, e.g., [15], [16], [18], [20], [21], [23] [36], [50], [51], and [59]). Here, parallel data streams are transmitted using multiple antennas to increase the spectral efficiency at the cost of increased complexity for data detection at the receiver. Optimal ML detection achieving full diversity can be performed using sphere decoding [4], [5], [60], whose complexity has been analyzed, e.g., in [54], [61], and [62]. Less complex suboptimal detection schemes are abundant but mostly do not achieve full diversity. Interestingly though, suboptimal detectors augmented with lattice reduction perform close to optimal and have the potential to achieve full diversity. This is of practical interest since lattice-reduction-aided detection also has two major complexity advantages over sphere decoding: 1) lattice reduction complexity is low (LLL has polynomial average complexity as opposed to exponentially growing average complexity of sphere decoding) and 2) with lattice reduction, the main computational burden accrues only when the channel changes whereas with sphere decoding the computationally expensive steps have to be performed for each new data vector.

SYSTEM MODEL AND BACKGROUND

In the following, we consider a spatial multiplexing MIMO system with M transmit and $N \geq M$ receive antennas as shown in Figure 6. Here, the lattice concepts and lattice reduction algorithms apply with $n = N$ and $m = M$.

At the transmitter the data is demultiplexed into M parallel data streams that are mapped to symbols from the QAM alphabet $\mathbb{S} \subset \mathbb{Z}_j$ and simultaneously transmitted over the M antennas. For QAM constellations, the condition $\mathbb{S} \subset \mathbb{Z}_j$ can easily be satisfied by an appropriate combination of scaling and translation. Consider $\mathbb{S} = \{-(1+j)/\sqrt{2}\,, (1-j)/\sqrt{2}, -(1-j)/\sqrt{2}, (1+j)/\sqrt{2}\}$, i.e., power-normalized 4-QAM. By adding $(1+j)/\sqrt{2}$ and then multiplying by $1/\sqrt{2}$, this constellation is transformed to $\{0,1,j,1+j\} \subset \mathbb{Z}_j$. The fixed additive offset can be moved into the observation whereas the multiplicative scaling can be incorporated into the channel (see, e.g., [16] and [50] for details). We consider one time slot of the discrete-time complex MIMO baseband model. Let $\mathbf{s}$ denote the complex-valued transmitted data vector with i.i.d. elements, each of which has power normalized to one. The corresponding received vector $\mathbf{y}$ is given by (see, e.g., [17])

$$\mathbf{y} = \mathbf{H}\mathbf{s} + \mathbf{w} \tag{13}$$

with the $N \times M$ (tall) channel matrix $\mathbf{H}$ and the noise vector $\mathbf{w}$. The noise is assumed to be spatially white (in case of spatially correlated noise, a spatial whitening filter can be used to obtain an equivalent model with white noise) complex Gaussian with variance σ_w^2.

CONVENTIONAL MIMO DETECTORS

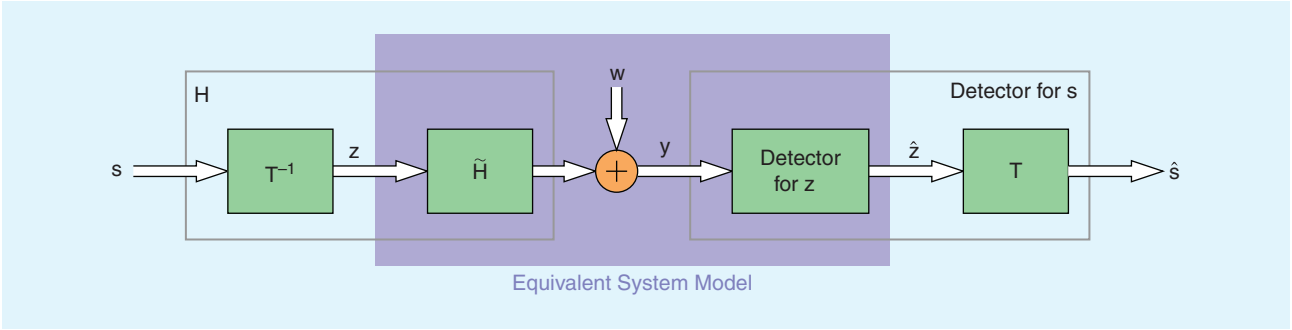
The ML detector for detecting the transmitted data vector based on (13) amounts to

$$\hat{\mathbf{s}}_{\mathrm{ML}} = \arg\min_{\mathbf{s} \in \mathbb{S}^M} \|\mathbf{y} - \mathbf{H}\mathbf{s}\|^2, \tag{14}$$

where $\mathbb{S}^M$ denotes the set of possible data vectors. For a reasonable number of antennas and modulation order, solving this closest vector problem via exhaustive search is typically not feasible. Consequently, various more efficient optimum and suboptimum data detection algorithms have been proposed in the literature. In particular, this includes efficient ML implementations based on sphere decoding (see, e.g., [4] and [60]) and low-complexity suboptimum algorithms based on linear equalization and successive interference cancelation (SIC) (see, e.g., [17]).

With linear equalization, the received vector is first passed through a linear spatial filter and the resulting filter output is then quantized (sliced) component-wise with respect to the symbol alphabet $\mathbb{S}$. For linear ZF equalization, the filter is designed to completely suppress spatial interference in the filter output signal. To this end, the received vector is multiplied with the (left) Moore-Penrose pseudoinverse of the channel matrix $\mathbf{H}^+ = (\mathbf{H}^H\mathbf{H})^{-1}\mathbf{H}^H$,

$$\tilde{\mathbf{s}}_{\mathrm{ZF}} = \mathbf{H}^+\mathbf{y} = \mathbf{s} + \mathbf{H}^+\mathbf{w}. \tag{15}$$



[FIG7] **Block diagram for a MIMO spatial multiplexing system with lattice-reduction-aided data detection.**

Note that this is the unconstrained least squares solution of (14). ZF detection is optimal (i.e., equivalent to ML) for orthogonal channel matrices but otherwise suffers from noise enhancement. An alternative to ZF equalization with improved performance is the minimum mean-square error (MMSE) equalizer, which minimizes the overall error consisting of residual spatial interference and noise at the filter output. As shown in [63] and [64], MMSE detection is mathematically equivalent to ZF detection with respect to an extended system model with the $(N+M)\times M$ channel matrix $\underline{\mathbf{H}}=\begin{pmatrix}\mathbf{H}\\ \sigma_w\mathbf{I}\end{pmatrix}$ and length-$(N+M)$ receive vector $\underline{\mathbf{y}}=\begin{pmatrix}\mathbf{y}\\ \mathbf{0}\end{pmatrix}$, i.e.,

$$\tilde{\mathbf{s}}_{\mathrm{MMSE}}=\underline{\mathbf{H}}^{+}\underline{\mathbf{y}}=(\mathbf{H}^{H}\mathbf{H}+\sigma_w^2\mathbf{I})^{-1}\mathbf{H}^{H}\mathbf{y}.$$

Even better performance can be achieved by nonlinear detection schemes such as SIC (or decision-feedback) detection [17], [64], [65]. With these methods, data symbols are detected successively by canceling the effect of previously detected symbols. Hence, the order in which the symbols are detected affects the performance and can be optimized. For example, the V-BLAST detection ordering algorithm [65] finds the ordering that leads to the maximum postequalization SNR at each detection step. Several computationally efficient algorithms have been proposed for this task in the literature, e.g., [63]–[66]. During SIC detection, the yet undetected data symbols have to be suppressed ("nulled"), which can be done again using either a ZF or an MMSE approach (see [64] for more details).

LATTICE-REDUCTION-AIDED DETECTION

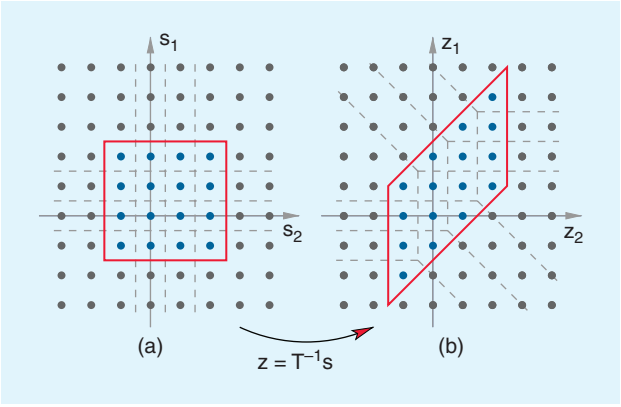
For common suboptimal detection approaches like linear equalization and SIC detection (see the previous section), the performance strongly depends on the specific channel realization (see, e.g., [50] and [67]). For example, ZF detection is optimal if the channel realization $\mathbf{H}$ happens to be orthogonal but results in poor performance otherwise. It is therefore natural to try to apply the detector not directly to $\mathbf{H}$, but rather to a transformed system model with a "more orthogonal" channel matrix $\tilde{\mathbf{H}}$ [16], [50], which can be obtained with lattice reduction. To this end, the channel matrix $\mathbf{H}$ is interpreted as a basis for the discrete lattice of noise-free received vectors given by $\mathbf{Hs}$. To determine a corresponding reduced basis $\tilde{\mathbf{H}}$ with better properties, one can use one of the lattice reduction algorithms described in the section "Lattice Reduction Techniques" with $\mathbf{B}=\mathbf{H}$. Since $\mathbf{H}$ is in general complex-valued, we note that either the corresponding lattice reduction algorithm has to be adapted to the complex-valued case or the equivalent real-valued system model has to be used (cf. the section "Complex-Valued Lattices"). Figure 7 illustrates the application of lattice reduction to data detection.

We first rewrite the system model (13) as

$$\mathbf{y}=\mathbf{H}\mathbf{T}\mathbf{T}^{-1}\mathbf{s}+\mathbf{w}=\tilde{\mathbf{H}}\mathbf{z}+\mathbf{w}, \tag{16}$$

where $\tilde{\mathbf{H}}=\mathbf{HT}$ and $\mathbf{z}=\mathbf{T}^{-1}\mathbf{s}$. Assuming for the moment that the symbol alphabet is given by the complex integers, i.e., $\mathbb{S}=\mathbb{Z}_j$, it follows from the unimodularity of $\mathbf{T}$ that $\mathbf{z}\in\mathbb{Z}_j^M$. Note that $\mathbf{Hs}$ and $\tilde{\mathbf{H}}\mathbf{z}$ describe the same lattice point but the reduced matrix $\tilde{\mathbf{H}}$ is more orthogonal than the original channel matrix $\mathbf{H}$. Thus, equalizing with the pseudoinverse $\tilde{\mathbf{H}}^{+}$ leads to $\tilde{\mathbf{z}}_{\mathrm{ZF}}=\tilde{\mathbf{H}}^{+}\mathbf{y}=\mathbf{z}+\tilde{\mathbf{H}}^{+}\mathbf{w}$, which suffers less noise amplification than conventional equalization according to (15). Consequently, a quantization based on $\tilde{\mathbf{z}}_{\mathrm{ZF}}$ is more reliable than that based on $\tilde{\mathbf{s}}_{\mathrm{ZF}}$. Denoting the quantization result obtained with the reduced channel $\tilde{\mathbf{H}}$ by $\hat{\mathbf{z}}_{\mathrm{ZF}}$, the final detection result is obtained as $\mathbf{T}\hat{\mathbf{z}}_{\mathrm{ZF}}$.

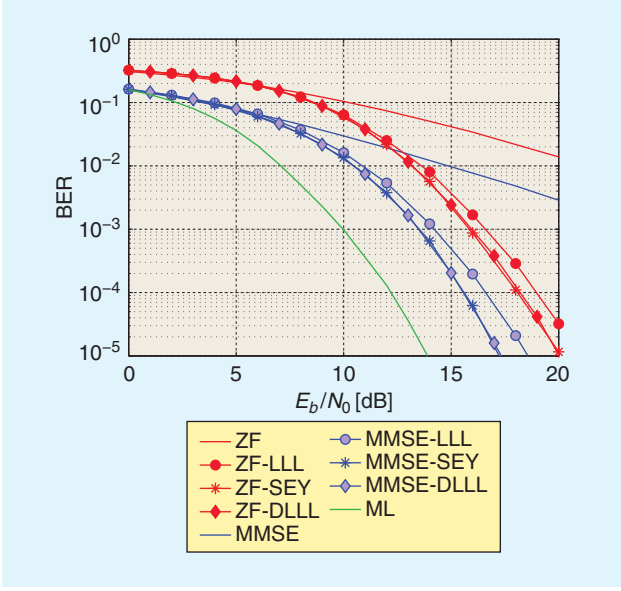
For practical systems, the symbol alphabet is a finite subset of the infinite set of integers, i.e., $\mathbb{S}\subset\mathbb{Z}_j$ (recall that we assumed $\mathbb{S}$ to be appropriately scaled and translated). Consequently, the domain of the transformed symbols $\mathbf{z}$ is also a subset of $\mathbb{Z}_j^M$. Appropriately taking into account the constellation boundary in the transformed domain leads to a shaping problem, which is illustrated in Figure 8 for a two-dimensional real-valued pulse amplitude modulation (PAM) constellation. For the original constellation, the optimum quantizer is equivalent to rounding (quantization with respect to $\mathbb{Z}_j^M$) followed by clipping to enforce the constellation boundary. In contrast, the decision regions for the transformed constellation differ from those of $\mathbb{Z}_j^M$ at the constellation boundary. Implementing these decision regions would again be computationally very expensive [68]. A simple but suboptimal alternative consists of the following three steps [16], [68]: 1) quantize (i.e., round) $\tilde{\mathbf{z}}$ with respect to $\mathbb{Z}_j^M$; 2) return to the original symbol domain by multiplying with $\mathbf{T}$; and 3) requantize (i.e., clip) the result with



[FIG8] Illustration of the shaping problem: (a) original two-dimensional 4-PAM constellation $\mathbb{S}^2 = \{-1, 0, 1, 2\}^2$ (blue dots) embedded in $\mathbb{Z}^2$; (b) constellation resulting with the unimodular transformation matrix $\mathbf{T}^{-1} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$. Slicer boundaries are indicated by dashed lines.

respect to $\mathbb{S}^M$ (the last step need not be performed if the intermediate result already belongs to the symbol constellation). This simple approach entails a noticeable performance loss [68], specifically for small constellation size.

The extension of linear lattice-reduction-aided detection according to the MMSE criterion is achieved by applying lattice reduction to the extended channel matrix $\underline{\mathbf{H}} = \begin{pmatrix} \mathbf{H} \\ \sigma_w \mathbf{I} \end{pmatrix}$ [50], [51]. In addition to improved performance, MMSE-based lattice reduction has a significantly smaller complexity than ZF-based lattice reduction; this can be explained by the fact that (due to the scaled identity matrix at the bottom) the extended channel matrix $\underline{\mathbf{H}}$ is more orthogonal and better conditioned than $\mathbf{H}$. Furthermore, as shown in [69], applying the MMSE criterion also limits the performance loss when quantizing with respect to $\mathbb{Z}_j$ instead of $\mathbb{S}_j$. To further improve the detection performance, lattice reduction can be combined with SIC detection [50], [51]. The resulting method is equivalent to the algorithm for finding the so-called Babai point [70]. While the reduced channel generally does not fulfill the V-BLAST ordering criterion, such an ordering can be achieved by using a postsorting algorithm [64]. We note that lattice-reduction-aided SIC detection can be interpreted as generalization of V-BLAST in which not only column swaps but also translation/size-reduction steps are performed.



[FIG9] BER versus E_b/N_0 for a MIMO system using ZF detection (red) and MMSE detection (blue) and their lattice-reduction-aided variants. ML detection is shown as an ultimate benchmark. Note that the SEY and DLLL variants lie virtually on top of each other. The MIMO system used $N = M = 6$ antennas, a 4-QAM symbol constellation, and an i.i.d. Gaussian channel.

PERFORMANCE RESULTS

In the following, we compare the performance of various lattice-reduction-aided data detection algorithms. Gauss reduction will not be considered, since it only applies to lattices of rank two. Furthermore, Hermite-Korkine-Zolotareff reduction is not considered as it results only in marginal performance improvements compared to other lattice reduction methods but features a much higher computational complexity [71], [72]. To compare the different combinations, the BER performance is investigated for a MIMO system with $M = N = 6$ antennas employing 4-QAM. The SNR is given by E_b/N_0 with E_b denoting the average receive energy per bit.

Figure 9 shows BER versus E_b/N_0 for the ML, ZF, and MMSE detectors along with their lattice-reduction-aided versions. Lattice reduction was achieved using LLL, S_2 reduction (referred to as SEY), and DLLL (LLL applied to the dual basis). It can be observed that lattice-reduction-aided linear equalization achieves full diversity order (in this case six), leading to strong performance improvements compared to conventional linear equalization that achieves a diversity order of only one. For the case of V-BLAST transmission over an i.i.d. Gaussian channel, the property of achieving full diversity was proven for LLL and DLLL in [18] (see also [14], and [19]–[23]). For general channel statistics and space-time mappings, full diversity of lattice-reduction-aided detection based on the LLL algorithm and the MMSE criterion follows from recent results in [69]. Furthermore, we can observe that SEY and DLLL noticeably outperform LLL. This can be explained as follows (cf. [59]). For ZF equalization, the layer with the worst postequalization SNR dominates the overall performance. The postequalization SNR of layer ℓ equals $1/(\sigma_w^2 \|\mathbf{h}_\ell^\star\|^2)$ and is thus inversely proportional to the Euclidean length of the dual basis vector $\mathbf{h}_\ell^\star$ [59]. Thus, for lattice-reduction-aided linear equalization, the goal is to find a reduced basis for which the longest dual vector is as short as possible so that the worst postequalization SNR is maximized. Both, SEY and DLLL reduce the dual basis $\mathbf{H}^\star$ and thus achieve a strong reduction of the longest dual basis vector [cf. Figure 3(b)], thereby explaining the performance advantage over LLL-aided linear equalization.

Figure 10 shows BER versus E_b/N_0 for plain SIC detectors based on ZF and MMSE and with V-BLAST ordering (referred to as OSIC) and for their lattice-reduction-aided variants. It can be observed that lattice reduction significantly improves the

performance of SIC detectors. Furthermore, OSIC-LLL, OSIC-SEY, and OSIC-DLL perform almost identically. OSIC-LLL offers the advantage that the QR decomposition required for SIC detection can be directly provided by the LLL algorithm.

TOPICS FOR FUTURE RESEARCH

The application of lattice reduction to MIMO detection is an active research area with several interesting open problems. Specifically, the impact of imperfect channel state information on lattice reduction aided detection is little understood. Numerical evidence (e.g., [50]) suggests that lattice reduction gains are preserved with imperfect channel knowledge. However, there are neither analytical performance results for this case nor specifically tailored detector designs.

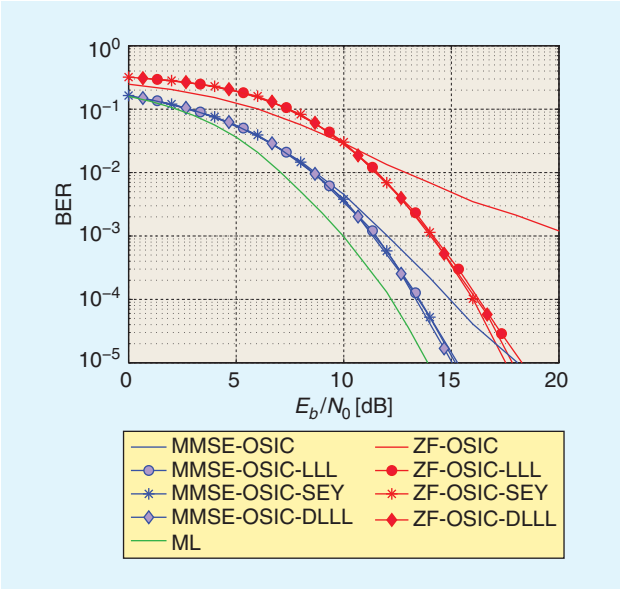
Another topic of high practical relevance is the extension of lattice-reduction-aided hard-output detectors (as discussed in this article) to the soft-output case. This is relevant for coded MIMO systems where iterative detection and decoding is enabled by exchanging soft information (e.g., log-likelihood ratios). Apart from list-based approaches (e.g., [73]–[75]), extensions into this direction seem difficult since the required quantization has to be performed in a transformed domain, where the information about the bit labels is not explicit.

Furthermore, no analytical performance results along the lines of [18], [20], and [21] (i.e., diversity order and SNR gap for LLL) are known for Seysen's lattice reduction algorithm. Finally, it would be very important to develop efficient hardware architectures for the various lattice-reduction-aided data detection algorithms (first results in this direction have been reported in [25] and [26]).

LATTICE REDUCTION FOR PRECODING

Another promising application of lattice reduction in wireless communications is efficient precoding. Precoding is used to realize MIMO gains in the following scenarios: 1) the receive antennas are not colocated (e.g., if they belong to distinct users) so that no spatial receiver processing is possible and 2) as much computational complexity as possible shall be shifted from the receiver to the transmitter (base station). The main prerequisite for precoding is channel state information at the transmitter. We will focus on precoding schemes that consist of pre-equalization in conjunction with a perturbation of the data vector to minimize the transmit power [76]–[78]. Such vector perturbation (VP) precoding schemes enable simple per-antenna (or, per-user) receiver processing.

As with data detection (see the section "Lattice Reduction for Data Detection"), the problem of finding the optimum (in the sense of minimizing the transmit power) perturbation vector is equivalent to a closest lattice point problem [cf. (14)], which can be solved using sphere-decoding approaches (in the precoding context, this is often referred to as sphere-encoding [77]). However, since this optimum approach may be computationally too expensive, several efficient approximate (i.e., suboptimum) VP techniques have been developed in the literature. Examples include linear precoding, Tomlinson-Harashima precoding (THP) [77], [78], and variants involving lattice reduction [16], [79]. Recently, it was shown that approximate VP preceded by lattice reduction using the LLL algorithm can achieve full diversity [18].

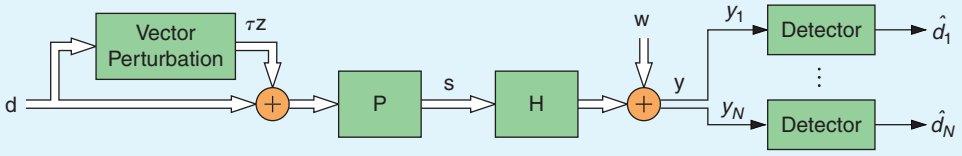


[FIG10] BER versus E_b/N_0 for a MIMO system using optimally ordered SIC (OSIC) based on ZF (red) and MMSE (blue) and their lattice-reduction-aided variants (all of which lie practically on top of each other). ML detection is shown as an ultimate benchmark. The MIMO system used $N = M = 6$ antennas, a 4-QAM symbol constellation, and an i.i.d. Gaussian channel.

SYSTEM MODEL AND BASIC APPROACH

We again consider the linear input/output relation (13), where the transmitter is equipped with M transmit antennas and there are N single-antenna users (or a single user equipped with N receive antennas). However, in contrast to the section "Lattice Reduction for Data Detection," we now assume $N \leq M$ (i.e., there are more transmit antennas than receive antennas/users and the channel $\mathbf{H}$ is a fat matrix). In this context, the discussion from the sections "Point Lattices" and "Lattice Reduction Techniques" applies with $n = M$ and $m = N$ to the lattice generated by the $M \times N$ right pseudoinverse $\mathbf{P} = \mathbf{H}^H(\mathbf{H}\mathbf{H}^H)^{-1}$ of the channel $\mathbf{H}$. Note that this right pseudoinverse $\mathbf{P}$ is different from the left pseudoinverse $\mathbf{H}^+$ used in the context of MIMO detection. At each time instant, the base station transmits N data symbols. The ℓth data symbol $d_\ell \in \mathbb{S}$ is intended for user (or receive antenna) ℓ. With perfect channel state information at the transmitter, interference free transmission to each user is achieved by using ZF precoding/pre-equalization (e.g., [76]). For simplicity, we do not consider MMSE-based precoding (e.g., [76]), even though it generally performs better than ZF precoding.

Here the transmit vector is obtained by multiplying the data vector $\mathbf{d} = (d_1 \ldots d_N)^T$ with the pseudoinverse $\mathbf{P}$. The main problem of this linear ZF precoding scheme lies in the fact that

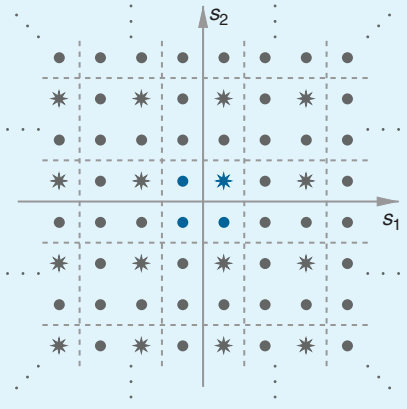


[FIG11] Block diagram for a MIMO precoding system using VP.

the transmit power $\|\mathbf{Pd}\|^2$ of the pre-equalized signal can become very large. This happens specifically if $\mathbf{H}$ is poorly conditioned and $\mathbf{d}$ is oriented along a channel singular vector associated to a small singular value. This power enhancement can be combatted by VP [77] (see Figure 11 for a block diagram). With VP, the transmit vector is formed by adding a (scaled) integer perturbation vector $\mathbf{z} \in \mathbb{Z}_j^N$ to the data vector before pre-equalization, i.e., $\mathbf{s} = \mathbf{P}(\mathbf{d} + \tau\mathbf{z})$, with τ a real-valued constant. This allows the data vector to be reoriented into a more favorable signal space direction. VP amounts to using a periodically extended symbol constellation $\mathbb{S} + \tau\mathbb{Z}_j^N$ in which all possible perturbed versions of a symbol vector constitute an equivalence class that actually carries the information (see Figure 12 for an illustration). The real-valued constant τ is chosen such that all translates of the symbol alphabet are nonoverlapping. The integer perturbation vector $\mathbf{z} \in \mathbb{Z}_j^N$ is designed to minimize the transmit power, i.e.,

$$\mathbf{z}_{\text{opt}} = \arg \min_{\mathbf{z} \in \mathbb{Z}_j^N} \|\mathbf{P}(\mathbf{d} + \tau\mathbf{z})\|^2. \tag{17}$$

The receive vector $\mathbf{y} = \mathbf{Hs} + \mathbf{w} = \mathbf{d} + \tau\mathbf{z} + \mathbf{w}$ is free of spatial interference and hence allows for per-antenna detection based on $y_\ell = d_\ell + \tau z_\ell + w_\ell$. The perturbation can be removed by subjecting the receive values to a modulo-τ operation (i.e., shifting the translated constellation back to the origin). Data detection then amounts to slicing $y_\ell \bmod \tau$ with respect to the constellation $\mathbb{S}$. We note that VP precoding has been shown to achieve the full diversity order of M.



[FIG12] Illustration of the main idea of VP for the real-valued constellation $\mathbb{S}^2 = \{-1,1\}^2$ (shown in blue). The periodic extension of $\mathbb{S}^2$ is obtained with $\tau = 4$ (boundaries are shown as dashed lines). The equivalence class for the symbol vector $\binom{1}{1}$ is indicated with stars; this equivalence class contains the vectors $\binom{1}{1} + \tau\binom{-k}{k} = \binom{-4k+1}{4k+1}$, which become increasingly orthogonal to $\binom{1}{1}$ as $|k|$ increases.

Finding the optimum perturbation vector according to (17) amounts to a closest vector problem that can be solved by sphere-encoding [77]. In fact, we look for $\mathbf{z} \in \mathbb{Z}_j^N$ such that $\tau\mathbf{Pz}$ is closest in Euclidean distance to the vector $-\mathbf{Pd}$. This problem is very similar to the ML detection problem (just replace $\mathbf{H}$, $\mathbf{s}$, and $\mathbf{y}$ in (14) with $-\tau\mathbf{P}$, $\mathbf{z}$, and $\mathbf{Pd}$, respectively), with the difference that the perturbation vector $\mathbf{z}$ comes from the infinite lattice $\mathbb{Z}_j^N$, whereas the symbol vector $\mathbf{s}$ is taken from the finite constellation $\mathbb{S}^M$.

Several suboptimal precoding techniques can be interpreted as approximations to (17). Specifically, (17) can be solved approximately by first computing the (unconstrained) least-squares solution by applying the pseudoinverse of $-\tau\mathbf{P}$ to $\mathbf{Pd}$ and then rounding the result to the nearest integer. This leads to the perturbation vector

$$\mathbf{z}_{\text{ZF}} = \left\lceil (-\tau\mathbf{P})^+ \mathbf{Pd} \right\rfloor = \left\lceil -\frac{1}{\tau}\mathbf{d} \right\rfloor = \mathbf{0} \tag{18}$$

(the last equality follows from the fact that the real part and the imaginary part of d_ℓ/τ are less than 1/2). We thus reobtain plain ZF precoding, i.e., $\mathbf{s}_{\text{ZF}} = \mathbf{Pd}$. The diversity order achieved by ZF precoding is only $M - N + 1$.

Alternatively, using a decision feedback approach (similar to SIC detection) for finding the elements of the perturbation vector $\mathbf{z}$ leads to nonlinear THP [78]. THP can be interpreted as successive optimization of the elements of the perturbation vector instead of a joint optimization. It performs better than ZF precoding, but its diversity order also equals only $M - N + 1$.

LATTICE-REDUCTION-AIDED PRECODING

As with lattice-reduction-aided data detection (see the section "Lattice Reduction for Data Detection"), using lattice reduction for precoding is motivated by the fact that optimum and good approximate choices of the integer perturbation can be found much more efficiently if the lattice basis that appears in the closest vector problem (17) is more orthogonal. In contrast to lattice-reduction-aided data detection, lattice reduction for precoding does not suffer from the shaping problem, i.e., the relaxation from a finite to an infinite lattice. Hence, in the context of precoding lattice reduction itself does not imply any performance loss (cf. [68]).

We now briefly review the basic concepts of lattice-reduction-aided precoding [16], [18], [79]. In the following, the relation between the original precoding matrix $\mathbf{P}$ and the corresponding reduced matrix $\widetilde{\mathbf{P}}$ is given by $\widetilde{\mathbf{P}} = \mathbf{PT}$, where $\mathbf{T}$ is an unimodular transformation matrix obtained by a certain lattice reduction algorithm (see the section "Lattice Reduction Techniques" with $\mathbf{B} = \mathbf{P}$).

The cost function in (17) can be rewritten in terms of the reduced matrix $\widetilde{\mathbf{P}}$ as

$$\begin{aligned} \|\mathbf{P}(\mathbf{d} + \tau\,\mathbf{z})\|^2 &= \|\mathbf{PTT}^{-1}(\mathbf{d} + \tau\,\mathbf{z})\|^2 \\ &= \|\widetilde{\mathbf{P}}(\widetilde{\mathbf{d}} + \tau\,\widetilde{\mathbf{z}})\|^2, \end{aligned} \tag{19}$$

where $\widetilde{\mathbf{d}} = \mathbf{T}^{-1}\mathbf{d}$ and $\widetilde{\mathbf{z}} = \mathbf{T}^{-1}\mathbf{z}$. The transformed perturbation vector $\widetilde{\mathbf{z}}$ is still integer-valued since $\mathbf{T}$ is unimodular. Hence, we arrive at a cost function that is completely equivalent to that used in the original formulation (17). Consequently, the optimum perturbation vector in (17) can also be found by first minimizing the lattice-reduction-transformed cost function (19) over $\widetilde{\mathbf{z}}$ and then applying the transformation $\mathbf{T}$ to the corresponding result $\widetilde{\mathbf{z}}_{\text{opt}}$, i.e., $\mathbf{z}_{\text{opt}} = \mathbf{T}\,\widetilde{\mathbf{z}}_{\text{opt}}$. Clearly, lattice reduction itself does not imply any loss of optimality. However, the optimum solution can be found more efficiently via sphere-encoding since $\widetilde{\mathbf{P}}$ is more orthogonal than $\mathbf{P}$. Furthermore, using conventional approximation techniques (like linear ZF or nonlinear THP) results in a better approximation quality when preceded by lattice reduction.

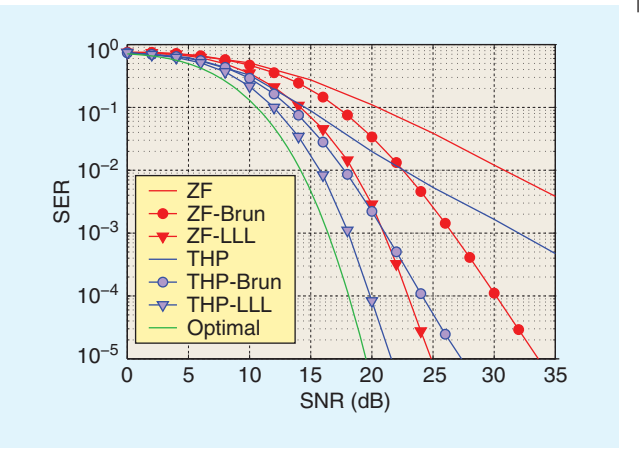
As an example, let us consider lattice-reduction-aided ZF precoding. Here, the reduced basis $\widetilde{\mathbf{P}}$ is used to compute the (unconstrained) least squares solution of (19) followed by a rounding operation [cf. (18)],

$$\widetilde{\mathbf{z}}_{\text{ZF}} = \left\lceil (-\tau\widetilde{\mathbf{P}})^{+}\,\widetilde{\mathbf{P}}\,\widetilde{\mathbf{d}} \right\rfloor = \left\lceil -\frac{1}{\tau}\,\widetilde{\mathbf{d}} \right\rfloor.$$

In general, $\widetilde{\mathbf{z}}_{\text{ZF}} \neq \mathbf{0}$ since $\widetilde{\mathbf{d}} = \mathbf{T}^{-1}\mathbf{d}$ is an integer linear combination of data symbols and thus can have elements that are larger than τ. Finally, the perturbation vector obtained by lattice-reduction-aided ZF precoding is given by $\mathbf{T}\,\widetilde{\mathbf{z}}_{\text{ZF}}$. If $\mathbf{T}$ is obtained by the LLL algorithm (see the section "LLL Reduction"), this approach can be shown to achieve the full diversity order of M [18].

BRUN'S ALGORITHM FOR LATTICE-REDUCTION-ASSISTED PRECODING

The use of Brun's algorithm for lattice reduction algorithm in the context of precoding in wireless systems has been proposed in [35]. This was motivated by the fact that under the usual i.i.d. Gaussian model, the channel matrix $\mathbf{H}$ typically has just one small singular value associated with the left singular vector $\mathbf{u}_1$ [67], [76]. When reducing $\mathbf{P}$, this means that $\mathbf{P}^H\mathbf{P} = (\mathbf{HH}^H)^{-1}$ has one dominating eigenvalue with associated eigenvector $\mathbf{u}_1$. As explained in the section "Brun's Algorithm," $\mathbf{P}$ can be reduced by finding approximate integer relations for $\mathbf{u}_1$ via Brun's algorithm. Full details are provided in [35]. A high-throughput VLSI implementation of Brun's algorithm for lattice-reduction-aided precoding (including fixed-point considerations) can be found in [24].



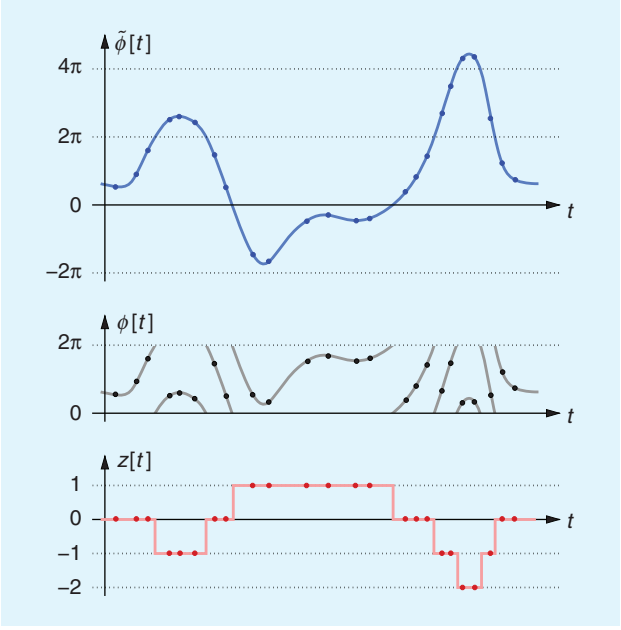
[FIG13] SER versus SNR for ZF precoding (red) and THP (blue) without lattice reduction, with Brun lattice reduction, and with LLL lattice reduction. Optimum VP (green) is shown as a performance benchmark. The results were obtained for $N = M = 8$ antennas, a 4-QAM symbol constellation, and an i.i.d. Gaussian channel.

PERFORMANCE

Figure 13 illustrates the symbol-error-rate (SER) versus SNR performance for the various precoding schemes with $M = 8$ transmit antennas, $N = 8$ users, 4-QAM symbols, and i.i.d. Gaussian channels. We compared linear ZF precoding and nonlinear THP with and without lattice reduction. We considered the LLL algorithm and Brun's algorithm for lattice reduction. The corresponding lattice-reduction-aided precoding schemes are referred to as ZF-LLL, THP-LLL, ZF-Brun, and THP-Brun. As a performance benchmark, we also considered the exact (optimum) solution of (17) via sphere-encoding. We can observe that lattice reduction significantly improves the performance of the approximate (i.e., ZF and THP) precoding schemes. With lattice reduction using the LLL algorithm, close to optimum performance is achieved and the full diversity order (in this case, $M = 8$) is obtained. The lattice-reduction-aided precoding schemes based on the simple Brun's algorithm suffer from a performance loss as compared to the corresponding schemes assisted by the LLL algorithm, but are still able to achieve a large part of the available diversity. This is in contrast to ZF and THP without lattice reduction, which just achieve a diversity order of one (cf. [76]).

TOPICS FOR FUTURE RESEARCH

There are various interesting directions for future research dealing with lattice-reduction-aided precoding. In particular, no analytical results on the performance (i.e., diversity order and SNR gap) of lattice-reduction-aided precoding using Brun's algorithm (complementing the analytical findings of [18] for the LLL algorithm) are available. Furthermore, little is known about the computational complexity of sphere-encoding (with and without lattice reduction), which is in contrast to sphere-decoding techniques for data detection, where various results on the complexity of sphere-decoding (without lattice reduction) have been provided in the literature (see [54], [61], and [62]). Indeed, a



[FIG14] Illustration of the phase unwrapping problem with unwrapped phase (blue), wrapped phase (gray), and integer unwrapping function (red); sampling instants are marked with dots.

theoretical underpinning of the complexity savings achieved with lattice reduction for sphere-encoding is completely missing. Here, only numerical complexity results are available (see, e.g., [79] for sphere-encoding aided by the LLL algorithm). Finally, similar to MIMO detection, little is known about lattice-reduction-aided precoding in situations with imperfect channel state information.

PHASE UNWRAPPING

In the previous sections, we have studied applications of lattice reduction in wireless communications in considerable detail. We next briefly discuss the usefulness of lattice reduction for a signal processing application outside MIMO wireless. Specifically, we focus on the phase unwrapping problem, which was used to formulate lattice-reduction-based estimators of the frequency and phase of a single sinusoid [10] and of the parameters of polynomial-phase signals [80]. In line with the suggestions in [10, Sec. 7], we discuss a more general setup with a nonlinear phase and nonuniformly spaced samples. This case has not been dealt with explicitly up to now and is intended to stimulate further research. The model we consider is potentially useful in applications like radar, sonar, geophysics, biomedical signal processing, sensor networks, and communications.

Consider the complex discrete-time signal $x[t] = e^{j\psi[t]} + w[t]$, where $\psi[t]$ is a phase function and $w[t]$ is additive noise. The phase function is modeled using a basis expansion, i.e.,

$$\psi[t] = \sum_{k=1}^{K} u_k[t]\,\theta_k,$$

where $\{u_k[t]\}_{k=1}^{K}$ is a prescribed set of linearly independent basis functions and $\boldsymbol{\theta} = (\theta_1 \ldots \theta_K)^T$ is an unknown parameter vector that we want to estimate.

We assume that we are given measurements of the wrapped phase $\phi[t] \in [0, 2\pi[$ of $x[t]$ at L irregularly spaced time instants $t_1 \ \ldots \ t_L$, i.e., $\varphi[t_\ell] = \sum_{k=1}^{K} u_k[t_\ell]\,\theta_k + \varepsilon[t_\ell] \bmod 2\pi$; here, $\varepsilon[t]$ denotes the phase error caused by the additive noise $w[t]$. The case of polynomial phase signals with uniform sampling ($t_\ell = \ell$) considered in [80] is obtained with the model $u_k[t_\ell] = \ell^{k-1}$ (frequency and phase estimation for a single sinusoid amounts to the special case $K = 2$ [10]).

By defining the length-L vectors $\boldsymbol{\phi} = (\phi[t_1] \ldots \phi[t_L])^T$, $\mathbf{u}_k = (u_k[t_1] \ldots u_k[t_L])^T$, $\boldsymbol{\epsilon} = (\varepsilon[t_1] \ldots \varepsilon[t_L])^T$, and the $L \times K$ matrix $\mathbf{U} = (\mathbf{u}_1 \ldots \mathbf{u}_K)$, we obtain the equivalent matrix-vector model

$$\boldsymbol{\phi} = \mathbf{U}\boldsymbol{\theta} + \boldsymbol{\epsilon} \mod 2\pi.$$

If we had the unwrapped phase $\tilde{\boldsymbol{\phi}} = \mathbf{U}\boldsymbol{\theta} + \boldsymbol{\epsilon}$ to our disposal, the parameter vector $\boldsymbol{\theta}$ could be estimated according to a simple least squares approach. The wrapped and unwrapped phase are related as $\tilde{\boldsymbol{\phi}} = \boldsymbol{\phi} - 2\pi\mathbf{z}$ where the integer vector $\mathbf{z} \in \mathbb{Z}^L$ models the unknown unwrapping (see Figure 14). We use an extended least squares approach to jointly estimate $\boldsymbol{\theta}$ and $\mathbf{z}$ according to

$$(\hat{\boldsymbol{\theta}}, \hat{\mathbf{z}}) = \arg \min_{\boldsymbol{\theta} \in \mathbb{R}^K,\, \mathbf{z} \in \mathbb{Z}^L} \|\boldsymbol{\phi} - 2\pi\mathbf{z} - \mathbf{U}\boldsymbol{\theta}\|^2. \tag{20}$$

For any given unwrapping vector $\mathbf{z}$, the associated parameter estimate equals

$$\hat{\boldsymbol{\theta}}(\mathbf{z}) = \mathbf{U}^{+}(\boldsymbol{\phi} - 2\pi\mathbf{z}). \tag{21}$$

Here, $\mathbf{U}^{+}$ denotes the (left) pseudoinverse of $\mathbf{U}$. By inserting $\hat{\theta}(\mathbf{z})$ into the cost function (20), we obtain after some algebra

$$\hat{\mathbf{z}} = \arg \min_{\mathbf{z} \in \mathbb{Z}^L} \|\mathrm{P}_{\mathbf{U}}^{\perp}(\boldsymbol{\phi} - 2\pi\mathbf{z})\|^2. \tag{22}$$

Here, $\mathrm{P}_{\mathbf{U}}^{\perp} = \mathbf{I} - \mathbf{U}\mathbf{U}^{+}$ denotes the rank-m ($m = L - K$) orthogonal projection matrix onto the orthogonal complement of the column span of $\mathbf{U}$. The final parameter estimate is obtained as $\hat{\boldsymbol{\theta}}(\hat{\mathbf{z}})$ according to (21) with $\hat{\mathbf{z}}$ the solution to (22).

The integer least squares problem (22) is recognized to be a closest vector problem with respect to the m-dimensional lattice induced by $\mathrm{P}_{\mathbf{U}}^{\perp}$, similar to the MIMO detection and precoding problems. Hence, all lattice-reduction-aided approximate solution techniques discussed in the MIMO context can be applied to the phase unwrapping problem as well. The main difference is that with MIMO the lattice is determined by the random channel matrix whereas here it depends on the underlying basis $\{u_k[t]\}_{k=1}^{K}$ and the sampling instants $t_1, \ldots, t_L$. This difference prompts several questions regarding the performance and complexity of lattice-reduction-aided estimation of the parameter vector $\boldsymbol{\theta}$, which are left for future research. Furthermore, lattice reduction algorithms are particularly interesting in the context of phase

unwrapping since here the lattice dimension grows linearly with the number of sampling points, which in practice can be much larger than the lattice dimension (i.e., number of antennas) for MIMO.

LATTICE REDUCTION AND ITS APPLICATION TO WIRELESS COMMUNICATIONS REMAINS AN ACTIVE RESEARCH AREA WITH SEVERAL INTERESTING OPEN PROBLEMS.

CONCLUSIONS

In this article, we provided a survey of lattice reduction techniques and their application to wireless communications and parameter estimation. After reviewing the basic concepts of lattices, we described the various lattice reduction algorithms that have been proposed in the literature for solving classical problems in lattice theory (such as the shortest vector problem). MATLAB code for some of the lattice reduction algorithms has been made available as supplementary material in IEEE *Xplore* (http://ieeexplore.ieee.org). We then discussed how the lattice reduction principle can be applied to simplify the detection and precoding problem in wireless communications with emphasis on multiple antenna systems. These communications problems were complemented by a discussion of the usefulness of lattice reduction algorithms in parameter estimation problems involving phase unwrapping. In all of these applications, the fundamental approach is as follows: 1) use a lattice reduction algorithm to determine an improved (i.e., "more orthogonal") basis for the lattice of interest; 2) solve the given problem with respect to the reduced basis; and 3) transform the solution back to the original domain.

One of the key results in the literature showed that in certain setups lattice reduction using the LLL algorithm allows suboptimum detection and precoding techniques to achieve full diversity. By means of numerical results we demonstrated that the various lattice reduction approaches (such as LLL, dual LLL, and Seysen) offer different advantages depending on the specific performance target (such as the orthogonality defect of the lattice basis or the error rate of a subsequent data detector). In particular, we showed that suboptimum linear and nonlinear detection and precoding schemes aided by lattice reduction are able to achieve excellent error rate performance.

We reviewed the known analytical results about the complexity of the various lattice reduction algorithms that were complemented by numerical complexity assessments. Here, one of the main results is that for lattices resulting from i.i.d. Gaussian channels (a model often used in wireless communications) the LLL algorithm has an average complexity that scales polynomially with the lattice dimension. This is in contrast to optimum detection and precoding approaches (e.g., sphere decoding), which feature an expected complexity that scales exponentially with the lattice dimension. We conclude that the tools provided by lattice theory are very powerful and at the same time easily applied to simplify the hard detection and precoding problems in wireless communications. The practical feasibility and importance of lattice reduction is corroborated by recent hardware implementations of various lattice reduction algorithms.

Lattice reduction and its application to wireless communications remains an active research area with several interesting open problems, e.g., the analysis of performance and complexity of lattice reduction in specific applications and practical hardware implementations. Our hope is that this survey article provides a convenient entry point to this exciting field and motivates a larger number of researchers to apply lattice reduction algorithms to a wider class of signal processing problems.

ACKNOWLEDGMENTS

We thank N. Sidiropoulos and the anonymous reviewers for their careful and critical reading and for their numerous comments and suggestions, which resulted in a significant improvement of this article. This work was supported in part by the STREP project MASCOT (IST-026905) within the Sixth Framework of the European Commission, by grant S10606 of the Austrian Science Fund (FWF), by grant ICA08-0046 of the Swedish Foundation for Strategic Research (SSF), and by the FP7 grant 228044 of the European Research Council (ERC).

AUTHORS

Dirk Wübben (wuebben@ant.uni-bremen.de) received the Dipl.-Ing. (FH) degree in electrical engineering from the University of Applied Science Münster, Germany, in 1998, and the Dipl.-Ing. (Uni) degree and the Dr.-Ing. degree in electrical engineering from the University of Bremen, Germany, in 2000 and 2005, respectively. He was a visiting student at the Daimler Benz Research Departments in Palo Alto, California, in 1997, and in Stuttgart, Germany, in 1998. From 1998 to 1999 he was with the Research and Development Center of Nokia Networks, Düsseldorf, Germany. In 2001, he joined the Department of Communications Engineering, University of Bremen, Germany, where he is currently a senior researcher and lecturer. His research interests include wireless communications, signal processing for multiple antenna systems, cooperative communication systems, and channel coding. He was a technical program cochair of the 2010 International ITG Workshop on Smart Antennas and has been member of the program committee of several international conferences. His doctoral dissertation on reduced complexity detection algorithms for MIMO systems received the Bremer Studienpreis in 2006. He is a Member of the IEEE.

Dominik Seethaler (dominik.seethaler@gmail.com) received the Dipl.-Ing. and Dr. techn. degrees in electrical/communication engineering from Vienna University of Technology, Austria, in 2002 and 2006, respectively. From 2002 to 2007, he was a research and teaching assistant at the Institute of Communications and Radio Frequency Engineering, Vienna University of Technology. From 2007 to

2009, he was a postdoctoral researcher at the Communication Technology Laboratory, ETH Zurich, Switzerland. Since 2009, he has been freelancing in the area of home robotics.

Joakim Jaldén (jalden@kth.se) received the M.Sc. and Ph.D. degrees in electrical engineering from the Royal Institute of Technology (KTH), Stockholm, Sweden, in 2002 and 2007, respectively. From 2007 to 2009, he held a postdoctoral research position at the Vienna University of Technology, Austria. He also studied at Stanford University, California, from 2000 to 2002, and worked at ETH, Zürich, Switzerland, as a visiting researcher, in 2008. In 2009, he joined the Signal Processing Lab within the School of Electrical Engineering at KTH, Stockholm, Sweden, as an assistant professor. He won the IEEE Signal Processing Society's 2006 Young Author Best Paper Award for his work on MIMO communications, and in 2007, he won first prize in the Student Paper Contest at the International Conference on Acoustics, Speech and Signal Processing. He also received the 2009 Ingvar Carlsson Award by the Swedish Foundation for Strategic Research.

Gerald Matz (gmatz@nt.tuwien.ac.at) received the Dipl.-Ing. and Dr. techn. degrees in electrical engineering in 1994 and 2000, respectively, and the Habilitation degree for communication systems in 2004, all from Vienna University of Technology, Austria. Since 1995, he has been with the Institute of Communications and Radio-Frequency Engineering, Vienna University of Technology, where he is currently associate professor. In 2004 and 2005, he was an Erwin Schrödinger Fellow with the Laboratoire des Signaux et Systèmes, Supélec, France. In 2007, he was a guest researcher with the Communication Theory Lab at ETH Zurich, Switzerland. He has directed or actively participated in several research projects funded by the Austrian Science Fund (FWF), the Vienna Science and Technology Fund (WWTF), and the European Union. He has published more than 130 papers in international journals, conference proceedings, and edited books. His research interests include wireless communications, statistical signal processing, and information theory. He serves on the IEEE Signal Processing Society (SPS) Technical Committee on Signal Processing for Communications and Networking and on the IEEE SPS Technical Committee on Signal Processing Theory and Methods. He is an associate editor for *IEEE Transactions on Signal Processing*. He was an associate editor for *IEEE Signal Processing Letters* (2004–2008) and for the EURASIP journal *Signal Processing* (2007–2010). He was technical program cochair of EUSIPCO 2004 and has been on the Technical Program Committee of numerous international conferences. He received the 2006 Kardinal Innitzer Most Promising Young Investigator Award and is a Senior Member of the IEEE.

REFERENCES

[1] R. Zamir, S. Shamai, and U. Erez, "Nested linear/lattice codes for structured multiterminal binning," *IEEE Trans. Inform. Theory*, vol. 48, no. 6, pp. 1250–1276, June 2002.

[2] A. K. Lenstra, H. W. Lenstra, and L. Lovász, "Factoring polynomials with rational coefficients," *Mathematische Annalen*, vol. 261, no. 4, pp. 515–534, 1982.

[3] M. Seysen, "Simultaneous reduction of a lattice basis and its reciprocal basis," *Combinatorica*, vol. 13, no. 3, pp. 363–376, 1993.

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

[5] U. Fincke and M. Pohst, "Improved methods for calculating vectors of short length in a lattice, including a complexity analysis," *Math. Comput.*, vol. 44, no. 170, pp. 463–471, 1985.

[6] A. Hassibi and S. Boyd, "Integer parameter estimation in linear models with applications to GPS," *IEEE Trans. Signal Processing*, vol. 46, no. 11, pp. 2938–2952, Nov. 1998.

[7] R. Neelamani, R. G. Baraniuk, and R. de Queiroz, "Compression color space estimation of JPEG images using lattice basis reduction," in *Proc. IEEE Int. Conf. Image Processing (ICIP),* Thessaloniki, Greece, Oct. 2001, vol. I, pp. 890–893.

[8] I. V. L. Clarkson, "Approximation of linear forms by lattice points with applications to signal processing," Ph.D. dissertation, Australian Nat. Univ., Canberra, Australia, 1997.

[9] I. V. L. Clarkson, S. D. Howard, and I. M. Y. Mareels, "Estimating the period of a pulse train from a set of sparse, noisy measurements," in *Proc. Int. Symp. Signal Processing and Its Applications*, Aug. 1996, vol. 2, pp. 885–888.

[10] I. V. L. Clarkson, "Frequency estimation, phase unwrapping and the nearest lattice point problem," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP),* Phoenix, AZ, Mar. 1999, pp. 1609–1612.

[11] W. H. Mow, "Maximum likelihood sequence estimation from the lattice viewpoint," Ph.D. dissertation, The Chinese Univ. of Hong Kong, June 1991.

[12] G. Rekaya, J.-C. Belfiore, and E. Viterbo, "A very efficient lattice reduction tool on fast fading channels," in *Proc. IEEE Int. Symp. Information Theory and Its Applications (ISITA),* Parma, Italy, Oct. 2004, pp. 714–717.

[13] K. J. Kim and R. A. Iltis, "Joint constrained data detection and channel estimation algorithms for QS-CDMA signals," in *Proc. Asilomar Conf. Signals, Systems and Computers,* Pacific Grove, CA, 2001, vol. 1, pp. 394–398.

[14] X. Ma, W. Zhang, and A. Swami, "Lattice-reduction aided equalization for OFDM systems," *IEEE Trans. Wireless Commun.*, vol. 8, no. 4, pp. 1608–1613, Apr. 2009.

[15] H. Yao and G. W. Wornell, "Lattice-reduction-aided detectors for MIMO communication systems," in *Proc. IEEE Global Communications Conf. (GLOBECOM),* Taipei, Taiwan, Nov. 2002.

[16] C. Windpassinger and R. F. H. Fischer, "Low-complexity near-maximum-likelihood detection and precoding for MIMO systems using lattice reduction," in *Proc. IEEE Information Theory Workshop (ITW),* Paris, France, Mar. 2003, pp. 345–348.

[17] A. Paulraj, R. Nabar, and D. Gore, *Introduction to Space-Time Wireless Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[18] M. Taherzadeh, A. Mobasher, and A. K. Khandani, "LLL reduction achieves the receive diversity in MIMO decdoing," *IEEE Trans. Inform. Theory*, vol. 53, no. 12, pp. 4801–4805, Dec. 2007.

[19] M. Taherzadeh, A. Mobasher, and A. K. Khandani, "Communication over MIMO broadcast channels using lattice-basis reduction," *IEEE Trans. Inform. Theory*, vol. 53, no. 12, pp. 4567–4582, Dec. 2007.

[20] C. Ling, "Towards characterizing the performance of approximate lattice decoding in MIMO communications," in *Proc. Int. ITG Conf. Source and Channel Coding (SCC),* Munich, Germany, Apr. 2006.

[21] C. Ling, "Approximate lattice decoding: Primal versus dual basis reduction," in *Proc. IEEE Int. Symp. Information Theory (ISIT),* Seattle, WA, July 2006, pp. 1–5.

[22] J. Jaldén and P. Elia, "LR-aided MMSE lattice decoding is DMT optimal for all approximately universal codes," in *Proc. IEEE Int. Symp. Information Theory (ISIT),* Seoul, Korea, June 2009, pp. 1263–1267.

[23] X. Ma and W. Zhang, "Performance analysis for MIMO systems with lattice-reduction aided linear equalization," *IEEE Trans. Commun. Technol.*, vol. 56, no. 2, pp. 309–318, Feb. 2008.

[24] A. Burg, D. Seethaler, and G. Matz, "VLSI implementation of a lattice-reduction algorithm for multi-antenna broadcast precoding," in *Proc. IEEE Int. Symp. Circuits and Systems (ISCAS)*, May 2007, pp. 673–676.

[25] B. Gestner, W. Zhang, X. Ma, and D. V. Anderson, "VLSI implementation of a lattice reduction algorithm for low-complexity equalization," in *Proc. IEEE Int. Conf. Circuits and Systems for Communications (ICCSC)*, May 2008, pp. 643–647.

[26] D. Wu, J. Eilert, and D. Liu, "A programmable lattice-reduction aided detector for MIMO-OFDMA," in *Proc. IEEE Int. Conf. Circuits and Systems for Communications (ICCSC)*, Shanghai, China, May 2008, pp. 293–297.

[27] L. G. Barbero, D. L. Milliner, T. Ratnarajah, J. R. Barry, and C. F. N. Cowan, "Rapid prototyping of Clarkson's lattice reduction for MIMO detection," in *Proc. IEEE Int. Conf. Communications (ICC)*, Dresden, Germany, June 2009, pp. 1–5.

[28] L. Bruderer, C. Studer, M. Wenk, D. Seethaler, and A. Burg, "VLSI implementation of a low-complexity LLL lattice reduction algorithm for MIMO

detection," in *Proc. IEEE Int. Symp. Circuits and Systems (ISCAS)*, Paris, France, May 2010, pp. 3745–3748.

[29] H. Cohen, *A Course in Computational Algebraic Number Theory,* 3rd ed. Berlin, Germany: Springer-Verlag, 1996.

[30] J. H. Conway and N. J. A. Sloane, *Sphere Packings, Lattices and Groups,* 3rd ed. New York: Springer-Verlag, 1998.

[31] N. D. Elkies, "Lattices, linear codes, and invariants, Part I," *Notices Amer. Math. Soc.*, vol. 47, no. 10, pp. 1238–1245, Nov. 2000.

[32] N. D. Elkies, "Lattices, linear codes, and invariants, Part II," *Notices Amer. Math. Soc.*, vol. 47, no. 11, pp. 1382–1391, Dec. 2000.

[33] P. Q. Nguyen and B. Vallée, Eds., *The LLL Algorithm: Survey and Applications*. Berlin, Germany: Springer-Verlag, 2010.

[34] H. Yao, "Effcient signal, code, and receiver designs for MIMO communication systems," Ph.D. dissertation, Dept. Elect. Eng. and Comp. Sci., Massachusetts Inst. Technol., Cambridge, MA, June 2003.

[35] D. Seethaler and G. Matz, "Efficient vector pertubation in multi-antenna multi-user systems based on approximate integer relations," in *Proc. European Signal Processing Conf. (EUSIPCO),* Florence, Italy, Sept. 2006.

[36] D. Seethaler, G. Matz, and F. Hlawatsch, "Low-complexity MIMO detection using Seysen's lattice reduction algorithm," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP),* Honolulu, HI, Apr. 2007, pp. 53–56.

[37] Y. H. Gan, C. Ling, and W. H. Mow, "Complex lattice reduction algorithm for low-complexity full-diversity MIMO detection," *IEEE Trans. Signal Processing*, vol. 57, no. 7, pp. 2701–2710, July 2009.

[38] H. W. Mow, "Universal lattice decoding: A review and some recent results," in *Proc. IEEE Int. Conf. Communications (ICC)*, Paris, France, June 2004, vol. 5, pp. 2842–2846.

[39] Y. H. Gan and H. W. Mow, "Complex lattice reduction algorithms for low-complexity MIMO detection," in *Proc. IEEE Global Communications Conf. (GLOBECOM),* St. Louis, MI, Nov. 2005, pp. 2953–2957.

[40] C. Ling and N. Howgrave-Graham, "Effective LLL reduction for lattice decoding," in *Proc. IEEE Int. Symp. Information Theory (ISIT),* Nice, France, June 2007, pp. 196–200.

[41] H. Minkowski, "Ueber positive quadratische Formen," *Journal für die reine und angewandte Mathematik*, vol. 1886, no. 99, pp. 1–9, 1886.

[42] H. Minkowski, "Ueber die positiven quadratischen Formen und über kettenbruchähnliche Algorithmen," *Journal für die reine und angewandte Mathematik*, vol. 1891, no. 107, pp. 278–297, 1891.

[43] H. Minkowski, *Geometrie der Zahlen.* Leipzig, Germany: Teubner Verlag, 1896.

[44] C. Hermite, "Extraits de lettres de M. Ch. Hermite á M. Jacobi sur différents objects de la théorie des nombres," *Journal für die reine und angewandte Mathematik*, vol. 1850, no. 40, pp. 279–290, 1850.

[45] A. Korkine and G. Zolotareff, "Sur les formes quadratiques," *Mathematische Annalen*, vol. 6, no. 3, pp. 366–389, 1873.

[46] C. F. Gauss, Untersuchungen über höhere Arithmetik, (Disquisitiones Arithmeticae). Berlin, Germany: Springer-Verlag, 1889.

[47] V. Brun, "En generalisation av kjedebrøken I," *Skr. Vidensk. Selsk. Kristiana, Mat. Nat. Klasse*, vol. 6, pp. 1–29, 1919.

[48] V. Brun, "En generalisation av kjedebrøken II," *Skr. Vidensk. Selsk. Kristiana, Mat. Nat. Klasse*, vol. 6, pp. 1–24, 1920.

[49] G. Golub and C. van Loan, *Matrix Computations,* 2nd ed. London: The John Hopkins Univ. Press, 1993.

[50] D. Wübben, R. Böhnke, V. Kühn, and K. D. Kammeyer, "MMSE-based lattice reduction for near-ML detection of MIMO systems," in *Proc. ITG Workshop Smart Antennas (WSA),* Munich, Germany, Mar. 2004.

[51] D. Wübben, R. Böhnke, V. Kühn, and K. D. Kammeyer, "Near-maximum-likelihood detection of MIMO systems using MMSE-based lattice reduction," *in Proc. IEEE Int. Conf. Communications (ICC)*, Paris, France, June 2004, vol. 2, pp. 798–802.

[52] B. A. LaMacchia, "Basis reduction algorithms and subset sum problems," Master's thesis, Massachusetts Inst. Technol., May 1991.

[53] M. Ajtai, "The shortest vector problem in L_2 is NP-hard for randomized reductions," in *Proc. 30th Annu. ACM Symp. Theory of Computing,* Dallas, TX, May 1998, pp. 10–19.

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

[55] H. Daudée and B. Vallée, "An upper bound on the average number of iterations of the LLL algorithm," *Theor. Comput. Sci.*, vol. 123, no. 1, p. 95115, Jan. 1994.

[56] A. Akhavi, "The optimal LLL algorithm is still polynomial in fixed dimension," *Theor. Comput. Sci.*, vol. 297, no. 1, pp. 3–23, Mar. 2003.

[57] J. Jaldén, D. Seethaler, and G. Matz, "Worst- and average-case complexity of LLL lattice reduction in MIMO wireless systems," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, NV, Apr. 2008, pp. 2685–2688.

[58] L. G. Barbero, T. Ratnarajah, and C. Cowan, "A comparison of complex lattice reduction algorithms for MIMO detection," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, NV, Apr. 2008, pp. 2705–2708.

[59] D. Wübben and D. Seethaler, "On the performance of lattice reduction schemes for MIMO data detection," in *Proc. Asilomar Conf. Signals, Systems and Computers,* Pacific Grove, CA, Nov. 2007, pp. 1534–1538.

[60] A. D. Murugan, H. E. Gamal, M. O. Damen, and G. G. Caire, "A unified framework for tree search decoding: Rediscovering the sequential decoder," *IEEE Trans. Inform. Theory*, vol. 52, no. 3, pp. 933–953, 2007.

[61] B. Hassibi and H. Vikalo, "On the sphere decoding algorithm I. Expected complexity," *IEEE Trans. Signal Processing*, vol. 53, no. 8, pp. 2806–2818, Aug. 2005.

[62] H. Vikalo and B. Hassibi, "On the sphere decoding algorithm II. Generalizations, second-order statistics, and applications to communications," *IEEE Trans. Signal Processing*, vol. 53, no. 8, pp. 2819–2834, Aug. 2005.

[63] B. Hassibi, "An efficient square-root algorithm for BLAST," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Istanbul, Turkey, June 2000, pp. 5–9.

[64] D. Wübben, R. Böhnke, V. Kühn, and K. D. Kammeyer, "MMSE extension of V-BLAST based on sorted QR decomposition," in *Proc. IEEE Vehicular Technology Conf. (VTC)*, Orlando, FL, Oct. 2003, pp. 508–512.

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

[66] D. Wübben, R. Böhnke, J. Rinas, V. Kühn, and K. D. Kammeyer, "Efficient algorithm for decoding layered space-time codes," *Electron. Lett.*, vol. 37, no. 22, pp. 1348–1350, Oct. 2001.

[67] H. Artés, D. Seethaler, and F. Hlawatsch, "Efficient detection algorithms for MIMO channels: A geometrical approach to approximate ML detection," *IEEE Trans. Signal Processing*, vol. 51, no. 11, pp. 2808–2820, Nov. 2003.

[68] C. Studer, D. Seethaler, and H. Bölcskei, "Finite lattice-size effects in MIMO detection," in *Proc. Asilomar Conf. Signals, Systems and Computers*, Oct. 2008, pp. 2032–2037.

[69] J. Jaldén and P. Elia, "DMT optimality of LR-aided linear decoders for a general class of channels, lattice designs, and system models," *IEEE Trans. Infom. Theory,* vol. 56, no. 10, pp. 4765–4780, Oct. 2010.

[70] L. Babai, "On Lovász lattice reduction and the nearest lattice point problem," *Combinatorica*, vol. 6, no. 1, pp. 1–13, 1986.

[71] C. Windpassinger and R. F. H. Fischer, "Optimum and sub-optimum lattice-reduction-aided detection and precoding for MIMO communications," in *Proc. Canadian Workshop Information Theory*, Waterloo, ON, Canada, May 2003, pp. 88–91.

[72] D. Wübben, "Effiziente Detektionsverfahren für Multilayer-MIMO-Systeme," (in German), Ph.D. dissertation, Universität Bremen, Arbeitsbereich Nachrichtentechnik, Shaker Verlag, Germany, Feb. 2006.

[73] P. Silvola, K. Hooli, and M. Juntti, "Suboptimal soft-output MAP detector with lattice reduction," *IEEE Signal Processing Lett.*, vol. 13, no. 6, pp. 321–324, June 2006.

[74] X.-F. Qi and K. Holt, "A lattice-reduction-aided soft demapper for high-rate coded MIMO-OFDM systems," *IEEE Signal Processing Lett.*, vol. 14, no. 5, pp. 305–308, May 2007.

[75] W. Zhang and X. Ma, "Approaching optimal performance by lattice-reduction aided soft detectors," in *Proc. Conf. Information Sciences and Systems (CISS),* Baltimore, MD, Mar. 2007, pp. 818–822.

[76] C. B. Peel, B. M. Hochwald, and A. L. Swindlehurst, "A vector-perturbation technique for near-capacity multiantenna multiuser communication—Part I: Channel inversion and regularization," *IEEE Trans. Commun.*, vol. 53, no. 1, pp. 195–202, Jan. 2005.

[77] B. M. Hochwald, C. B. Peel, and A. L. Swindlehurst, "A vector-perturbation technique for near-capacity multiantenna multiuser communication—Part II: Perturbation," *IEEE Trans. Commun.*, vol. 53, no. 3, pp. 537–544, Mar. 2005.

[78] C. Windpassinger, R. F. H. Fischer, T. Vencel, and J. B. Huber, "Precoding in multiantenna and multiuser communications," *IEEE Trans. Wireless Commun.*, vol. 3, no. 4, pp. 1305–1316, July 2004.

[79] C. Windpassinger, R. F. H. Fischer, and J. B. Huber, "Lattice-reduction-aided broadcast precoding," *IEEE Trans. Commun.*, vol. 52, no. 12, pp. 2057–2060, Dec. 2004.

[80] R. G. McKilliam, I. V. L. Clarkson, B. G. Quinn, and B. Moran, "Polynomial-phase estimation, phase unwrapping and the nearest lattice point problem," in *Proc. Asilomar Conf. Signals, Systems, and Computers,* Pacific Grove, CA, Oct. 2009, pp. 483–485.

[SP]