Elektrotechnik und Informationstechnik (e&i), vol. 122, no. 3, March 2005, pp. 91–96.

Detection Techniques for MIMO Spatial Multiplexing Systems

Detektionsmethoden für Mehrantennensysteme mit räumlichem Multiplex

Dominik Seethaler, *Harold Artés*+, *and Franz Hlawatsch**

*Institute of Communications and Radio-Frequency Engineering, Vienna University of Technology
Gusshausstrasse 25/389, A-1040 Wien, Austria
Phone: +43 1 58801 38958, Fax: +43 1 58801 38999, E-mail: dominik.seethaler@tuwien.ac.at

+Information Systems Laboratory, Stanford University,
Packard 234, 350 Serra Mall, Stanford, CA 94305-9510, USA

Accepted for publication in e&i, January 21, 2005

Abstract — We discuss and compare the most important detection techniques for MIMO spatial multiplexing wireless systems, focusing on their performance and computational complexity. Our analysis shows that the limited performance of conventional suboptimal detection techniques is primarily caused by their inability to cope with poorly conditioned channels. The recently proposed *sphere projection algorithm* is better suited to these channels and can achieve near-optimal performance.

Index Terms — MIMO, spatial multiplexing, maximum likelihood detection, V-BLAST, sphere decoding, nulling and cancelling.

Zusammenfassung — Wir diskutieren und vergleichen die Arbeitsweise, Leistungsfähigkeit und Komplexität der wichtigsten Detektionsmethoden für Mehrantennen-Funkübertragungssysteme mit räumlichem Multiplex. Unsere Analyse zeigt, dass die begrenzte Leistungsfähigkeit herkömmlicher suboptimaler Detektoren durch schlecht konditionierte Kanäle bedingt ist. Der kürzlich vorgeschlagene *Kugelprojektions-Detektor* ist besser für diese Kanäle geeignet; seine Leistungsfähigkeit kann jener des optimalen Detektors nahekommen.

Stichwörter — Mehrantennensysteme, MIMO, räumlicher Multiplex, optimaler Detektor, V-BLAST, Kugel-Dekodierung, entscheidungsrückgekoppelter Detektor.

List of Abbreviations

Abbreviation	Description
BPSK	binary phase-shift keying
FPSD	Fincke-Phost sphere-decoding
MIMO	multiple-input multiple-output
ML	maximum likelihood
MMSE	minimum mean-square error
NC	nulling-and-cancelling
pdf	probability density function
PSK	phase-shift keying
QAM	quadrature amplitude modulation
SER	symbol error rate
SNR	signal-to-noise ratio
SPA	sphere projection algorithm
V-BLAST	vertical Bell-Labs layered space-time
ZF	zero-forcing

List of Figures

1 Introduction

Multiple-input multiple-output (MIMO) wireless communications systems employ multiple antennas at the transmitter and receiver sides. They can yield significantly increased data rates and improved link reliability (the latter due to spatial diversity) without additional bandwidth [1, 2].

High data rates can be realized by means of *spatial multiplexing* (e.g., the *V-BLAST* scheme [1, 3]), where independent information streams are transmitted in parallel over different transmit antennas. Consider a flat-fading MIMO channel with M_{T} transmit antennas and $M_{\mathrm{R}} \geq M_{\mathrm{T}}$ receive antennas (this will be briefly termed an $(M_{\mathrm{T}}, M_{\mathrm{R}})$ channel). This channel is part of a spatial multiplexing system where the mth data symbol (or *layer*) d_m is directly transmitted on the mth transmit antenna. At a given time instant, this leads to the well-known baseband model

$$\mathbf{r} \,=\, \mathbf{H}\mathbf{d} + \mathbf{w}\,, \tag{1}$$

with the $M_{\mathrm{T}} \times 1$ transmit vector $\mathbf{d} \triangleq \big(d_1 \; d_2 \; \cdots \; d_{M_{\mathrm{T}}}\big)^T$, the $M_{\mathrm{R}} \times M_{\mathrm{T}}$ channel matrix $\mathbf{H}$, the $M_{\mathrm{R}} \times 1$ received vector $\mathbf{r} \triangleq \big(r_1 \; r_2 \; \cdots \; r_{M_{\mathrm{R}}}\big)^T$, and the $M_{\mathrm{R}} \times 1$ noise vector $\mathbf{w} \triangleq \big(w_1 \; w_2 \; \cdots \; w_{M_{\mathrm{R}}}\big)^T$. The (n, m)th entry of $\mathbf{H}$, $H_{n,m} = (\mathbf{H})_{n,m}$, is the complex-valued fading coefficient between the mth transmit antenna and the nth receive antenna. The data symbols d_m are drawn from a complex-valued symbol alphabet $\mathcal{A}$ and are assumed zero-mean and independent with unit variance. The noise components w_m are assumed circularly symmetric complex Gaussian with variance σ_w^2. The channel $\mathbf{H}$ is assumed perfectly known at the receiver.

In a MIMO spatial multiplexing system, the maximum possible diversity is given by the number of receive antennas M_{R}. This maximum diversity is available if all channel coefficients $H_{n,m}$ are independent, because then each data symbol d_m is transmitted over M_{R} independent scalar fading channels $H_{n,m}$, $n = 1, \ldots, M_{\mathrm{R}}$ (cf. (1)). The larger M_{R}, the smaller is the probability that all these channels fade simultaneously, and thus the reliability of data detection can be improved. If the available diversity is M_{R}, the symbol error rate (SER) of the optimal maximum likelihood (ML) detector decays like $\mathrm{SNR}^{-M_{\mathrm{R}}}$ in the high-SNR regime [2, 4]. This corresponds to a slope of $-M_{\mathrm{R}}$ of the double-logarithmic SER-versus-SNR curve. In general, if the SER of some detector decays like $\mathrm{SNR}^{-\delta}$, $\delta \leq M_{\mathrm{R}}$, we say that the detector can exploit δth-order diversity.

The ML detector is optimal and fully exploits the available diversity. Unfortunately, the computational complexity of a direct implementation of the ML detector grows exponentially with the number of transmit antennas M_{T}, and it may be too high already for moderate system and constellation sizes [5]. Several efficient suboptimal detection techniques have therefore been proposed or adapted from the field of multiuser detection [3, 6, 7]. Whereas these techniques are much less computationally demanding than the ML detector, they are often unable to exploit a large part of the available diversity, and thus their performance tends to be significantly poorer than that of ML detection. An interesting alternative to suboptimal techniques is the Fincke-

Phost sphere-decoding algorithm for ML detection [5, 8, 9] (hereafter abbreviated as FPSD). The FPSD implements ML detection with significantly reduced *average* complexity; however, for a *specific* channel realization its complexity may still be very high.

Here, we consider the major detection techniques for MIMO spatial multiplexing systems: ML detection, equalization-based detection, nulling-and-cancelling, and FPSD. We discuss the strengths and weaknesses of these techniques regarding performance and computational complexity. We argue that the inferior performance of conventional suboptimal detection techniques is caused by their inability to cope with "bad" (i.e., poorly conditioned) channel realizations. In addition, bad channels lead to a high computational complexity of the FPSD. Motivated by these insights, we finally consider the recently proposed *sphere projection algorithm* (SPA) [10]. The SPA, by its design, is well suited to bad channels and thus can achieve near-ML performance at a computational complexity comparable with that of nulling-and-cancelling detectors.

Our paper is organized as follows. In Section 2, we discuss conventional detection techniques as well as the SPA. Section 3 provides simulation results for performance evaluation and comparison.

2 MIMO Detection Techniques

In this section, we first discuss major "classical" detection techniques for MIMO spatial multiplexing systems, namely, the (optimal) ML detector, the (suboptimal) equalization-based and nulling-and-cancelling detectors, and the FPSD implementation of ML detection. We then study the effects of bad channels on the performance of suboptimal detectors and discuss the SPA.

2.1 Maximum Likelihood Detection

ML detection is optimal in the sense of minimum error probability when all data vectors are equally likely, and it fully exploits the available diversity. For our system model (1) and with the assumptions made in Section 1, the ML detector is given by

$$\hat{\mathbf{d}}_{\mathrm{ML}} \,=\, \arg\min_{\mathbf{d}\in\mathcal{D}} \left\{ \|\mathbf{r}-\mathbf{H}\mathbf{d}\|^2 \right\}. \tag{2}$$

Here, $\mathcal{D} = \mathcal{A}^{M_{\mathrm{T}}}$ denotes the set of all possible transmitted data vectors $\mathbf{d}$. The cardinality of $\mathcal{D}$ is $|\mathcal{D}| = |\mathcal{A}|^{M_{\mathrm{T}}}$ and thus grows exponentially with M_{T}.

ML detection corresponds to a *nonconvex* optimization problem because $\mathcal{D}$ is not a convex set [11, 12]. Therefore, standard numerical algorithms for convex optimization are not applicable. The straightforward solution of (2) by comparing $\|\mathbf{r}-\mathbf{H}\mathbf{d}\|^2$ for all $\mathbf{d} \in \mathcal{D}$ has computational complexity $\mathcal{O}(|\mathcal{A}|^{M_{\mathrm{T}}})$, and in fact the complexity of ML detection may be excessive already for moderate values of M_{T} and constellation size $|\mathcal{A}|$. The FPSD implementation of ML detection will be discussed in Subsection 2.4.

2.2 Equalization-Based Detection

In linear equalization based detection, an estimate of the transmitted data vector $\mathbf{d}$ is formed as $\mathbf{y} = \mathbf{G}\mathbf{r}$ with an "equalization matrix" $\mathbf{G}$. The detected data vector is then obtained as $\hat{\mathbf{d}} = Q\{\mathbf{y}\}$, where $Q\{\cdot\}$ denotes componentwise quantization according to the symbol alphabet $\mathcal{A}$.

For the *zero-forcing* (ZF) equalizer, $\mathbf{G}$ is given by the pseudo-inverse [13] of $\mathbf{H}$, i.e., $\mathbf{G} = \mathbf{H}^{\#} = (\mathbf{H}^H\mathbf{H})^{-1}\mathbf{H}^H$. (For the last expression, we assumed that $M_{\mathrm{R}} \geq M_{\mathrm{T}}$ and that $\mathbf{H}$ has full rank.) Thus, the result of ZF equalization (before quantization) is

$$\mathbf{y}_{\mathrm{ZF}} \,=\, \mathbf{H}^{\#}\mathbf{r} \,=\, (\mathbf{H}^H\mathbf{H})^{-1}\mathbf{H}^H\mathbf{r} \,=\, \mathbf{d} + \tilde{\mathbf{w}}\,, \tag{3}$$

which is the transmitted data vector $\mathbf{d}$ corrupted by the transformed noise $\tilde{\mathbf{w}} = \mathbf{H}^{\#}\mathbf{w}$. This means that the interference caused by the channel $\mathbf{H}$ is completely removed ("forced to zero"). However, in general the transformed noise $\tilde{\mathbf{w}} = \mathbf{H}^{\#}\mathbf{w}$ is larger than $\mathbf{w}$ ("noise enhancement"); this will be analyzed in Subsection 2.5. The ZF-equalized received vector $\mathbf{y}_{\mathrm{ZF}}$ can be seen as the solution to a *relaxed* ML problem (cf. (2)) where the data set $\mathcal{D}$ underlying ML detection is relaxed to the convex set $\mathbb{C}^{M_{\mathrm{T}}}$ [12]:

$$\mathbf{y}_{\mathrm{ZF}} \,=\, \arg\min_{\mathbf{y}\in\mathbb{C}^{M_{\mathrm{T}}}} \left\{\|\mathbf{r} - \mathbf{H}\mathbf{y}\|^2\right\}.$$

The noise enhancement effect plaguing the ZF equalizer can be reduced by using the *minimum mean-square error* (MMSE) equalizer $\mathbf{G} = \big(\mathbf{H}^H\mathbf{H} + \sigma_w^2\mathbf{I}\big)^{-1}\mathbf{H}^H$, which is the $\mathbf{G}$ minimizing the mean-square error $\mathrm{E}\big\{\|\mathbf{G}\mathbf{r} - \mathbf{d}\|^2\big\}$ [14]. Thus, the result of MMSE equalization is

$$\mathbf{y}_{\mathrm{MMSE}} \,=\, \big(\mathbf{H}^H\mathbf{H} + \sigma_w^2\mathbf{I}\big)^{-1}\mathbf{H}^H\mathbf{r}\,.$$

This can again be seen as the solution to a relaxed ML problem, with the distance $\|\mathbf{r} - \mathbf{H}\mathbf{y}\|^2$ augmented by a penalty term $\sigma_w^2\|\mathbf{y}\|^2$ that prevents $\mathbf{y}$ from growing too large [12]:

$$\mathbf{y}_{\mathrm{MMSE}} \,=\, \arg\min_{\mathbf{y}\in\mathbb{C}^{M_{\mathrm{T}}}} \left\{\|\mathbf{r} - \mathbf{H}\mathbf{y}\|^2 + \sigma_w^2\|\mathbf{y}\|^2\right\}.$$

There also exist more sophisticated detection techniques based on the principle of relaxing the ML problem (e.g., *semidefinite relaxation* as proposed in [12] for multiuser detection).

While ZF or MMSE equalization alone does not, in general, imply a loss of information (i.e., an optimal detector could still be based on $\mathbf{y}_{\mathrm{ZF}}$ or $\mathbf{y}_{\mathrm{MMSE}}$), the subsequent componentwise quantization of $\mathbf{y}_{\mathrm{ZF}}$ or $\mathbf{y}_{\mathrm{MMSE}}$ is suboptimal since it does not take into account the correlation of the components of the transformed noise $\tilde{\mathbf{w}}$. In fact, ZF or MMSE detection can only exploit a diversity of order $M_{\mathrm{R}} - M_{\mathrm{T}} + 1$ [4]. On the other hand, the computational complexity is rather low. The task with highest complexity is the calculation of the equalizer matrix $\mathbf{G}$. Thus, if we assume $M_{\mathrm{T}} = M_{\mathrm{R}}$ for simplicity, the complexity behaves as $\mathcal{O}(M_{\mathrm{T}}^3)$. Note that MMSE detection is different from ML or ZF detection in that it requires an estimate of the noise variance.

2.3 Nulling-and-Cancelling

Nulling-and-cancelling (NC) is a recursive detection technique using the decision-feedback principle [3]. At each detection step, a single data vector component is detected and the corresponding contribution to the received vector $\mathbf{r}$ is subtracted from $\mathbf{r}$; the other components that have not yet been detected are "nulled out" (equalized) using a ZF or MMSE equalizer.

Let us consider the first detection step. Equalization-based detection of the m_1th data vector component ($m_1 \in \{1, \ldots, M_\mathrm{T}\}$) yields $\hat{d}_{m_1} = Q\big\{(\mathbf{Gr})_{m_1}\big\}$, where $\mathbf{G}$ is the ZF or MMSE equalizer matrix. NC then attempts to clean $\mathbf{r}$ from the interference caused by d_{m_1} by forming

$$\mathbf{r}^{(2)} \,=\, \mathbf{r} - \mathbf{h}_{m_1}\hat{d}_{m_1}\,,$$

where $\mathbf{h}_{m_1}$ denotes the m_1th column of the channel matrix $\mathbf{H}$. If $\hat{d}_{m_1}$ is correct, i.e., $\hat{d}_{m_1} = d_{m_1}$, we obtain the *reduced system model*

$$\mathbf{r}^{(2)} \,=\, \mathbf{H}^{(2)}\mathbf{d}^{(2)} + \mathbf{w}\,, \tag{4}$$

where the reduced channel matrix $\mathbf{H}^{(2)}$ of size $M_\mathrm{R} \times (M_\mathrm{T} - 1)$ is $\mathbf{H}$ with the m_1th column removed and the reduced data vector $\mathbf{d}^{(2)}$ of size $M_\mathrm{T} - 1$ is $\mathbf{d}$ with the m_1th component removed. Because in the reduced system model (4) M_T is replaced by $M_\mathrm{T} - 1$, equalization-based detection applied to $\mathbf{r}^{(2)}$ can now exploit one additional degree of diversity.

At the second decoding step, another data vector component d_{m_2} is detected by applying equalization-based detection to $\mathbf{r}^{(2)}$. The result is $\hat{d}_{m_2} = Q\big\{(\mathbf{G}^{(2)}\mathbf{r}^{(2)})_{m_2}\big\}$, where $\mathbf{G}^{(2)}$ denotes the ZF or MMSE equalizer matrix corresponding to $\mathbf{H}^{(2)}$. The interference caused by $\hat{d}_{m_2}$ is then subtracted from $\mathbf{r}^{(2)}$ by forming

$$\mathbf{r}^{(3)} \,=\, \mathbf{r}^{(2)} - \mathbf{h}_{m_2}\hat{d}_{m_2}\,.$$

This recursive detection-and-interference-cancellation procedure is continued until all M_T data vector components have been detected.

It is seen that NC attempts to progressively clean $\mathbf{r}$ from the interference caused by the components already detected. At each new detection step, additional degrees of diversity become available provided that all previous decisions were correct. The performance of NC depends crucially on the *order* in which the data vector components are processed. To minimize error propagation effects and to improve the detection of unreliable components through the additional degrees of diversity that become available in the reduced system models, *more reliable data vector components should be detected first.* Usually, at each decoding step the row norms of the equalizers $\mathbf{G}^{(l)}$ or the componentwise post-equalization SNRs are used as measures of reliability [3, 15, 16]. The performance of NC is significantly improved by such a *layer-sorting* procedure, although it is still poorer than that of ML detection. A "dynamic" layer-sorting strategy that leads to a further substantial performance improvement has recently been proposed [17].

The maximal diversity order that can be exploited by NC is $M_{\mathrm{R}} - M_{\mathrm{T}} + 1$, the same as for equalization-based detection [4, 18]. This is because the diversity exploited by NC is limited by the first step of the detection process, which is equalization-based detection applied to the full system model and thus has diversity order $M_{\mathrm{R}} - M_{\mathrm{T}} + 1$. In the practically relevant SNR regime (not extremely high SNRs), however, higher slopes of the SER-versus-SNR curve can be achieved. The computational complexity of NC is dominated by the calculation of the equalizers $\mathbf{G}^{(l)}$, $l = 1, \ldots, M_{\mathrm{T}}$. If we assume $M_{\mathrm{T}} = M_{\mathrm{R}}$ for simplicity, the complexity of a straightforward implementation of NC behaves as $\mathcal{O}(M_{\mathrm{T}}^4)$. However, this can be reduced to $\mathcal{O}(M_{\mathrm{T}}^3)$ by using the algorithm of [15] or the recursive implementation of [19].

2.4 The Fincke-Phost Sphere Decoding Algorithm

The Fincke-Phost sphere decoding (FPSD) algorithm [8] for ML detection allows the efficient determination of all data vectors $\mathbf{d} \in \mathcal{D}$ for which $\mathbf{Hd}$ lies within a hypersphere with given radius r about the received vector $\mathbf{r}$, i.e.,

$$\|\mathbf{r} - \mathbf{Hd}\|^2 < r^2. \tag{5}$$

If there are any $\mathbf{d}$'s inside this hypersphere, the ML solution $\hat{\mathbf{d}}_{\mathrm{ML}}$ must be one of them because $\mathbf{H}\hat{\mathbf{d}}_{\mathrm{ML}}$ is closest to $\mathbf{r}$ (cf. (2)). To find $\hat{\mathbf{d}}_{\mathrm{ML}}$, it then suffices to calculate and minimize $\|\mathbf{r} - \mathbf{Hd}\|^2$ for the data vectors produced by the FPSD, which implies a substantial reduction of complexity.

We can rewrite (5) as [15]

$$(\mathbf{d} - \mathbf{y}_{\mathrm{ZF}})^H \mathbf{H}^H \mathbf{H} (\mathbf{d} - \mathbf{y}_{\mathrm{ZF}}) < \tilde{r}^2, \qquad \text{with } \tilde{r}^2 \triangleq r^2 - \|\mathbf{r}\|^2 + \|\mathbf{H}\mathbf{y}_{\mathrm{ZF}}\|^2, \tag{6}$$

where, as before, $\mathbf{y}_{\mathrm{ZF}} = (\mathbf{H}^H\mathbf{H})^{-1}\mathbf{H}^H\mathbf{r}$ is the result of ZF equalization. With the QR decomposition [13] $\mathbf{H} = \mathbf{QR}$, where $\mathbf{Q}$ is unitary and $\mathbf{R}$ is upper triangular, (6) becomes

$$(\mathbf{d} - \mathbf{y}_{\mathrm{ZF}})^H \mathbf{R}^H \mathbf{R} (\mathbf{d} - \mathbf{y}_{\mathrm{ZF}}) = \|\mathbf{R}(\mathbf{d} - \mathbf{y}_{\mathrm{ZF}})\|^2 < \tilde{r}^2,$$

or equivalently

$$\sum_{m=1}^{M_{\mathrm{T}}} \left| \sum_{i=m}^{M_{\mathrm{T}}} R_{m,i}(d_i - y_{\mathrm{ZF},i}) \right|^2 < \tilde{r}^2, \tag{7}$$

with $R_{m,i} = (\mathbf{R})_{m,i}$. Evidently, a set of *necessary* conditions for (7) is given by the fact that all the partial sums of the outer sum in (7) must be smaller than $\tilde{r}^2$. One of these necessary conditions involves just the data vector component $d_{M_{\mathrm{T}}}$:

$$R^2_{M_{\mathrm{T}},M_{\mathrm{T}}} |d_{M_{\mathrm{T}}} - y_{\mathrm{ZF},M_{\mathrm{T}}}|^2 < \tilde{r}^2.$$

If for a specific $d_{M_{\mathrm{T}}} = a \in \mathcal{A}$ this condition is not satisfied, then we can discard all $\mathbf{d}$'s with $d_{M_{\mathrm{T}}} = a$ (the corresponding vectors $\mathbf{Hd}$ lie outside the hypersphere). If the condition is satisfied, however, then all $\mathbf{d}$'s with $d_{M_{\mathrm{T}}} = a$ remain possible candidates, and we can invoke another necessary condition that involves also $d_{M_{\mathrm{T}}-1}$:

$$\left|R_{M_{\mathrm{T}}-1,M_{\mathrm{T}}-1}(d_{M_{\mathrm{T}}-1} - y_{\mathrm{ZF},M_{\mathrm{T}}-1}) + R_{M_{\mathrm{T}}-1,M_{\mathrm{T}}}(a - y_{\mathrm{ZF},M_{\mathrm{T}}})\right|^2 < \tilde{r}^2 - R^2_{M_{\mathrm{T}},M_{\mathrm{T}}}|a - y_{\mathrm{ZF},M_{\mathrm{T}}}|^2 \, .$$

Again, if for a specific $d_{M_{\mathrm{T}}-1} = a' \in \mathcal{A}$ this condition is not satisfied, we can discard all $\mathbf{d}$'s with $d_{M_{\mathrm{T}}-1} = a'$, etc. This procedure is continued until the last condition (which is the necessary and sufficient condition (7) itself) is checked. The $\mathbf{d}$'s that survive this last check are all the data vectors inside the hypersphere. They represent a reduced search set for the ML solution, i.e., minimizing $\|\mathbf{r} - \mathbf{Hd}\|^2$ over this set yields the ML solution. For a convenient tree representation of this procedure see e.g. [20].

An appropriate choice of the hypersphere radius r is of crucial importance. If r is too small, we will not find any data vector inside the hypersphere; the reduced search set thus is empty and the ML solution cannot be provided. If r is too large, the reduced search set will be very large and thus the computational complexity of the subsequent search for the ML solution will be excessive. Usually, r is adjusted according to the noise variance (e.g. [5, 21, 22]). Each time the FPSD produces an empty search set, it has to be restarted with a larger radius.

The computational complexity of the FPSD implementation of the ML detector is strongly dependent on the channel realization and the SNR. The *average* complexity (i.e., averaged over a sufficient number of channel realizations) is exponential in M_{T} [9] just as the complexity of the straightforward implementation of the ML detector (cf. Section 2.1); however, for sufficiently large SNR it behaves polynomially in M_{T} as long as M_{T} is not too large [5]. Unfortunately, for *specific* channel realizations the complexity may still be very high.

2.5 The Sphere Projection Algorithm

In the case of a "good" MIMO channel—i.e., a channel whose condition number[1] is near to one—, the performance of the suboptimal detectors (ZF, MMSE, and NC) is very close to ML performance and the FPSD implementation of the ML detector is very efficient. However, for a "bad"—i.e., poorly conditioned—MIMO channel, the suboptimal detectors perform quite poorly and the FPSD has high complexity [8]. The *sphere projection algorithm* (SPA) [10] is motivated by the observation that the inferior average performance of the standard suboptimal detectors is mainly caused by the occurrence of bad channel realizations.

The effect of bad channels is best seen from the ZF-equalized vector $\mathbf{y}_{\mathrm{ZF}} = \mathbf{d} + \tilde{\mathbf{w}}$ (see (3)). The transformed noise vector $\tilde{\mathbf{w}}$ is correlated according to the covariance matrix $\mathbf{R}_{\tilde{\mathbf{w}}} = \sigma_w^2(\mathbf{H}^H\mathbf{H})^{-1}$, which implies a *distortion* of the noise probability density function (pdf) relative to the spherical pdf geometry of the original noise $\mathbf{w}$. More specifically, the contour surfaces of the pdf of $\tilde{\mathbf{w}}$

[1] The condition number of the channel matrix $\mathbf{H}$ is the ratio of the largest and smallest singular values of $\mathbf{H}$.

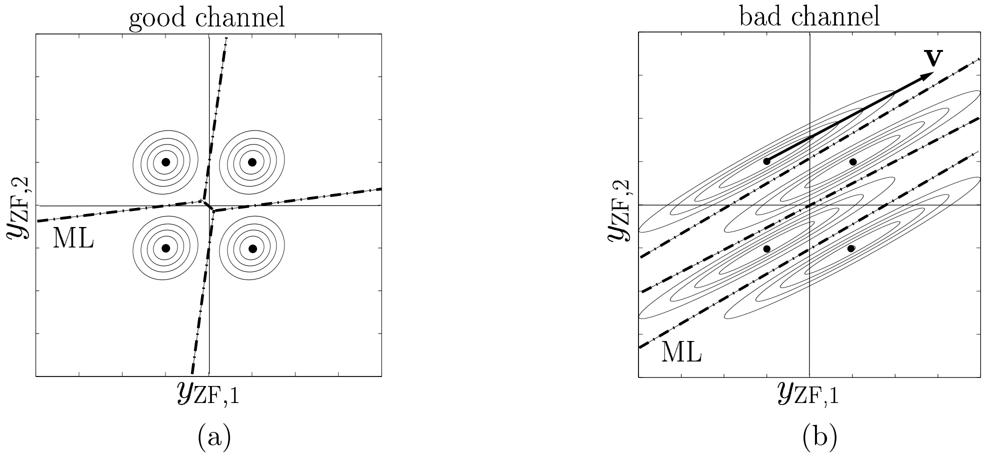

Figure 1: *Contour lines of the noise pdf as well as the ZF and ML decision regions in the ZF-equalized domain for a* $(2,2)$ *channel and BPSK modulation: (a) "Good" channel realization with condition number* 1.1*, (b) "bad" channel realization with condition number* 10.5 *(the vector* $\mathbf{v}$ *indicates the dominant noise direction).*

are *hyperellipsoids* [23] instead of hyperspheres as for $\mathbf{w}$. This distortion becomes stronger with increasing channel condition number. For illustration, Fig. 1 shows the pdf of $\mathbf{y}_{\mathrm{ZF}}$ for a good and a bad realization of a real-valued $(2,2)$ channel. The modulation format is BPSK. Also shown are the ZF decision regions (the four quadrants) and the ML decision regions (indicated by dash-dotted lines). For the good channel, the ZF and ML decision regions are quite similar; for the bad channel, however, they are very different. The decision regions of the MMSE and NC detectors (again represented in the ZF domain, not shown in Fig. 1) are somewhat better matched to the distorted pdf of $\tilde{\mathbf{w}}$ than the ZF decision regions but still very different from the ML decision regions [10].

The SPA is a simple nonlinear add-on to an arbitrary suboptimal detector that significantly increases that detector's robustness to bad channels. The basic idea is as follows. Recall from (2) that the ML detector would minimize $\|\mathbf{r}-\mathbf{Hd}\|^2$ over the entire data vector set $\mathcal{D}$. Let $\hat{\mathbf{d}}_{\mathrm{sub}} \in \mathcal{D}$ denote the result of the given suboptimal detector. This result can be expected to be reasonably good for good channels. In order to improve the result for bad channels, we use an additional set $\mathcal{D}_+ \subset \mathcal{D}$ of valid data vectors $\mathbf{d}$ that are potentially better than $\hat{\mathbf{d}}_{\mathrm{sub}}$ in the sense of smaller $\|\mathbf{r}-\mathbf{Hd}\|^2$. We then minimize $\|\mathbf{r}-\mathbf{Hd}\|^2$ over the combined search set $\mathcal{D}_{\mathrm{SP}}$ consisting of $\hat{\mathbf{d}}_{\mathrm{sub}}$ and all data vectors in $\mathcal{D}_+$:

$$\hat{\mathbf{d}}_{\mathrm{SP}} \triangleq \arg\min_{\mathbf{d}\in\mathcal{D}_{\mathrm{SP}}} \left\{\|\mathbf{r}-\mathbf{Hd}\|^2\right\}, \qquad \text{with } \mathcal{D}_{\mathrm{SP}} \triangleq \{\hat{\mathbf{d}}_{\mathrm{sub}}\} \cup \mathcal{D}_+ \,. \tag{8}$$

For the construction of $\mathcal{D}_+$, it is assumed that in the ZF domain there is a single *dominant* noise component, as indicated in Fig. 1(b) by the vector $\mathbf{v}$. Since a bad channel is assumed, this dominant noise component in the direction of $\mathbf{v}$ is much larger than all other noise components. Therefore, it makes sense to exclude the dominant noise direction in all distance calculations.

Furthermore, to simplify things, we will do as if all the other (nondominant) noise components had equal variance. With these approximations, it can be shown [10] that a data vector $\mathbf{d}$ has a small distance $\|\mathbf{r} - \mathbf{H}\mathbf{d}\|^2$ if and only if it is close to the straight line

$$\mathcal{L}: \quad \mathbf{y}(k) \triangleq k\,\mathbf{v} + \mathbf{y}_{\mathrm{ZF}}\,, \qquad k \in \mathbb{C}\,,$$

i.e., the line parallel to the dominant noise axis $\mathbf{v}$ that passes through $\mathbf{y}_{\mathrm{ZF}}$. In fact, if all nondominant noise components were exactly zero, $\mathcal{L}$ would pass right through the transmitted data vector $\mathbf{d}$.

Let us assume a constant-modulus symbol alphabet such as a PSK constellation (the SPA can be extended to more general alphabets but its complexity will be larger). All data vectors $\mathbf{d}$ then are located on an M_{T}-dimensional *data hypersphere* $\mathcal{H}$ about the origin. According to the above discussion, the data vectors in $\mathcal{D}_+$ should be close to the line $\mathcal{L}$. The intersection between $\mathcal{L}$ and $\mathcal{H}$—if it is nonempty—is a circle in the real plane corresponding to the (complex) line $\mathcal{L}$. We choose $\mathcal{D}_+$ to consist of all data vectors whose ZF decision (or quantization) regions are pierced by this intersection circle, since these vectors are close to $\mathcal{L}$. If $\mathcal{L}$ and $\mathcal{H}$ do not intersect, we choose $\mathcal{D}_+$ to consist of the data vector closest to $\mathcal{L}$ and all its nearest neighbors. A detailed discussion of the construction of $\mathcal{D}_+$ is given in [10].

Simulation results (see Section 3) indicate that the SPA combined with a suitable suboptimal detection technique exploits a large part of the available diversity. In fact, for practically relevant system sizes the SPA can get very close to ML performance. The SPA is more complex than the standard suboptimal techniques due to the construction of $\mathcal{D}_+$ and the minimization in (8). However, the extra complexity is moderate because these additional operations can be performed very efficiently [10]. When combined with the ZF or MMSE detector or an efficient implementation of the NC detector [15, 19], the complexity of the SPA is $\mathcal{O}(M_{\mathrm{T}}^3)$ [10].

3 Simulation Results

Next, we present simulation results in order to assess and compare the error-rate performance and computational complexity of the various detection techniques. We used 4-QAM modulation and a $(4, 4)$ MIMO channel with iid Gaussian entries. Fig. 2 shows the SER-versus-SNR performance[2] of the various detectors, with part (a) focusing on the comparison of ZF and MMSE detection, unsorted NC, and sorted NC (using the sorting strategy proposed in [3]) and part (b) comparing the performance of various suboptimal techniques (including the ZF-based SPA) with ML performance. The following conclusions can be drawn from these results.

[2] The SNR is defined as $\mathrm{E}\{\|\mathbf{H}\mathbf{d}\|^2\}/\mathrm{E}\{\|\mathbf{w}\|^2\} = M_{\mathrm{T}}\,\sigma_d^2$ (recall that $\sigma_w^2 = 1$).

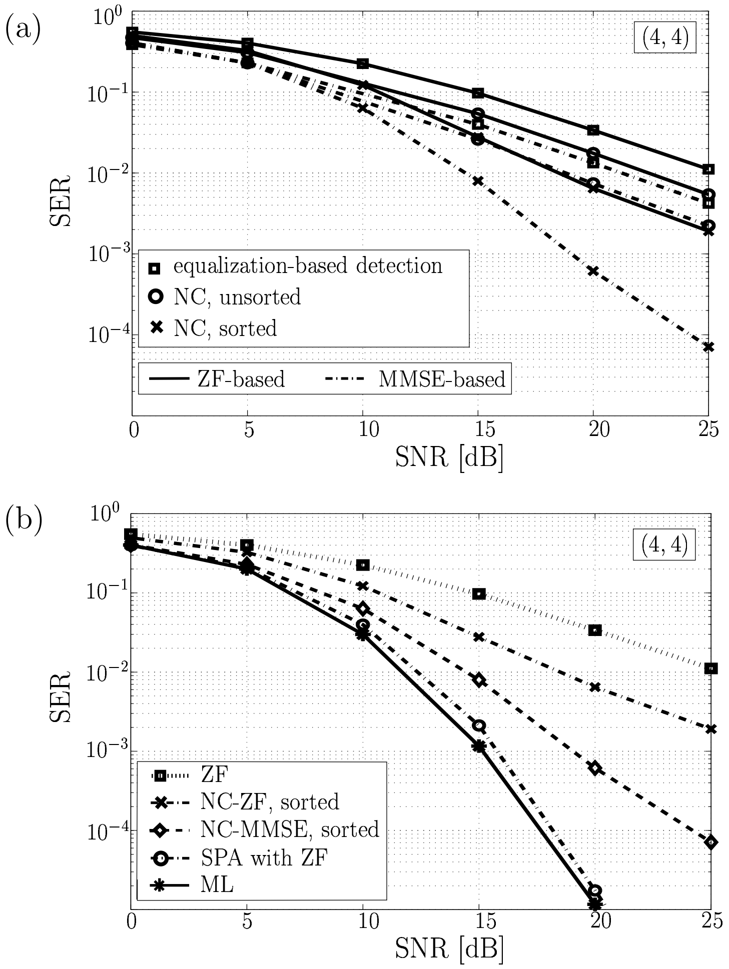

Figure 2: *SER-versus-SNR performance of various detectors for a* $(4, 4)$ *MIMO channel using 4-QAM data modulation: (a) Equalization-based detection in comparison with sorted/unsorted NC, (b) SPA based on ZF detection in comparison with ZF, sorted NC, and ML.*

- Unsorted NC based on ZF and MMSE equalization outperforms pure ZF and MMSE detection, respectively.
- Component sorting for NC can yield large performance improvements for SNRs above 10 dB, especially for MMSE-based NC. Nevertheless, sorted NC cannot exploit all of the available diversity.
- ML detection fully exploits the available diversity of order 4.
- The ZF-based SPA almost achieves ML performance and exploits almost all of the available diversity.

A rough picture of the complexity of the various detectors that complements the asymptotic $\mathcal{O}(\cdot)$ results presented in Section 2 can be obtained by measuring the kflops required by MATLAB V5.3 implementations. In the simulation study described above, the direct implementation of

ML detection required 43 kflops. The FPSD implementation of ML detection required only 5.2 kflops on average (at an SNR of 10 dB); however, the worst-case complexity (measured during 10000 simulation runs) was 26.7 kflops. Hence, on average the FPSD is far more efficient than a direct ML implementation, but in the worst case the computational savings may be moderate. Pure ZF and MMSE detection (2.3 kflops) and ZF- and MMSE-based NC using the efficient implementation in [19] (2.9 kflops) are much less complex than the FPSD. Finally, the average complexity of the ZF-based SPA (4.1 kflops) is similar to the average complexity of the FPSD; however, its worst-case complexity (5.1 kflops) is roughly 5 times smaller than the worst-case complexity of the FPSD even though the SPA achieves near-ML performance. For system sizes larger than the $(4,4)$ case considered, the suboptimal techniques achieve larger computational savings relative to the ML detector but their performance loss becomes larger as well; this is also true for the SPA.

4 Conclusions

We have studied and compared the most important classical detection techniques for MIMO spatial multiplexing systems, focusing on their error performance and computational complexity. In particular, we have highlighted the degrading influence that "bad" (i.e., poorly conditioned) channels exert on classical suboptimal techniques: either the performance is reduced, or the computational complexity is increased. This motivated a discussion of the recently introduced *sphere projection algorithm* (SPA), which is an add-on to conventional suboptimal techniques resulting in significantly improved robustness to bad channels. Simulation results allowed a performance comparison of the various detectors and showed that SPA-enhanced techniques can almost achieve optimal performance and exploit a large part of the available diversity.

References

[1] G. J. Foschini, "Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas," *Bell Labs Tech. J.*, vol. 1, no. 2, pp. 41–59, 1996.

[2] V. Tarokh, N. Seshadri, and A. R. Calderbank, "Space-time codes for high data rate wireless communications: Performance criterion and code construction," *IEEE Trans. Inf. Theory*, vol. 44, pp. 744–765, March 1998.

[3] 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. on Signals, Systems and Electronics*, (Pisa, Italy), pp. 295–300, Sept. 1998.

[4] L. Zheng and D. Tse, "Diversity and multiplexing: A fundamental tradeoff in multiple antenna channels," *IEEE Trans. Inf. Theory*, vol. 49, pp. 1073–1096, May 2003.

[5] B. Hassibi and H. Vikalo, "On the expected complexity of sphere decoding," in *Proc. 35th Asilomar Conf. Signals, Systems, Computers*, (Pacific Grove, CA), pp. 1051–1055, Nov. 2001.

[6] G. Ginis and J. Cioffi, “On the relation between V-BLAST and GDFE,” *IEEE Comm. Letters*, vol. 5, pp. 364–366, Sept. 2001.

[7] S. Verdú, *Multiuser Detection.* Cambridge (UK): Cambridge Univ. Press, 1998.

[8] U. Fincke and M. Phost, “Improved methods for calculating vectors of short length in a lattice, including a complexity analysis,” *Math. Comp.*, vol. 44, pp. 463–471, April 1985.

[9] J. Jaldén and B. Ottersten, “An exponential lower bound on the expected complexity of sphere decoding,” in *Proc. IEEE ICASSP 2004*, vol. IV, (Montreal, Canada), pp. 393–396, May 2004.

[10] H. Artés, D. Seethaler, and F. Hlawatsch, “Efficient detection algorithms for MIMO channels: A geometrical approach to approximate ML detection,” *IEEE Trans. Signal Processing, Special Issue on MIMO Communications Systems*, vol. 51, pp. 2808–2820, Nov. 2003.

[11] S. Boyd and L. Vandenberghe, *Convex Optimization.* Cambridge University Press, Dec. 2004.

[12] Wing-Kin Ma, T. Davidson, Kon Max Wong, Zhi-Quan Luo, and Pak-Chung Ching, “Quasi-maximum-likelihood multiuser detection using semi-definite relaxation with application to synchronous CDMA,” *IEEE Trans. Signal Processing*, vol. 50, pp. 912–922, April 2002.

[13] G. H. Golub and C. F. Van Loan, *Matrix Computations.* Baltimore: Johns Hopkins University Press, 3rd ed., 1996.

[14] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory.* Englewood Cliffs (NJ): Prentice Hall, 1993.

[15] B. Hassibi, “A fast square-root implementation for BLAST,” in *Proc. 34th Asilomar Conf. Signals, Systems, Computers*, (Pacific Grove, CA), pp. 1255–1259, Nov./Dec. 2000.

[16] M. Rupp, M. Guillaud, and S. Das, “On MIMO decoding algorithms for UMTS,” in *Proc. 35th Asilomar Conf. Signals, Systems, Computers*, vol. 2, (Pacific Grove, CA), pp. 975–979, Nov. 2001.

[17] D. Seethaler, H. Artés, and F. Hlawatsch, “Dynamic nulling-and-cancelling with near-ML performance,” in *Proc. IEEE ICASSP 2004*, vol. IV, (Montreal, Canada), pp. 777–780, May 2004.

[18] W.-J. Choi, R. Negi, and J. Cioffi, “Combined ML and DFE decoding for the V-BLAST system,” in *Proc. IEEE-ICC-00*, (New Orleans, LA), pp. 18–22, June 2000.

[19] J. Benesty, Y. Huang, and J. Chen, “A fast recursive algorithm for optimum sequential signal detection in a BLAST system,” *IEEE Trans. Signal Processing*, vol. 51, pp. 1722–1730, July 2003.

[20] M. Damen, H. El Gamal, and G. Caire, “On maximum-likelihood detection and the search for the closest lattice point,” *IEEE Trans. Inf. Theory*, vol. 49, pp. 2389–2402, Oct. 2003.

[21] E. Viterbo and J. Boutros, “A universal lattice code decoder for fading channels,” *IEEE Trans. Inf. Theory*, vol. 45, pp. 1639–1642, July 1999.

[22] B. M. Hochwald and S. ten Brink, “Achieving near-capacity on a multiple-antenna channel,” *IEEE Trans. Inf. Theory*, vol. 51, pp. 389–399, March 2003.

[23] C. W. Therrien, *Discrete Random Signals and Statistical Signal Processing.* Englewood Cliffs (NJ): Prentice Hall, 1992.