The Diversity Order of the Semidefinite Relaxation Detector

Joakim Jaldén and Björn Ottersten
Signal Processing Lab, School of Electrical Engineering,
KTH, Royal Institute of Technology,
Stockholm, Sweden
[joakim.jalden,bjorn.ottersten]@ee.kth.se

Abstract

We consider the detection of binary (antipodal) signals transmitted in a spatially multiplexed fashion over a fading multiple-input multiple-output (MIMO) channel and where the detection is done by means of semidefinite relaxation (SDR). The SDR detector is an attractive alternative to maximum likelihood (ML) detection since the complexity is polynomial rather than exponential. Assuming that the channel matrix is drawn with i.i.d. real valued Gaussian entries, we study the receiver diversity and prove that the SDR detector achieves the maximum possible diversity. Thus, the error probability of the receiver tends to zero at the same rate as the optimal maximum likelihood (ML) receiver in the high signal to noise ratio (SNR) limit. This significantly strengthens previous performance guarantees available for the semidefinite relaxation detector. Additionally, it proves that full diversity detection is in certain scenarios also possible when using a non-combinatorial receiver structure.

Index Terms

Semidefinite relaxation, diversity, MIMO, detection.

Manuscript submitted June 13, 2006. The material in this paper will be presented in part at the Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, October 2006.

I. INTRODUCTION

Herein, we consider the detection of binary symbols transmitted over an n by m multiple-input multiple-output (MIMO) channel modelled according to

$$\mathbf{y} = \mathbf{H}\mathbf{s} + \mathbf{v} \tag{1}$$

where $\mathbf{s} \in \mathcal{B}^m \triangleq \{\pm 1\}^m$, $\mathbf{H} \in \mathbb{R}^{n \times m}$ and $\mathbf{v}, \mathbf{y} \in \mathbb{R}^n$. In what follows, $\mathbf{y}$ is referred to as the vector of *received signals*; $\mathbf{H}$ as the *channel matrix*; $\mathbf{s}$ as the *transmitted message*; and $\mathbf{v}$ as the additive *noise* based on their physical interpretations in the digital communications context. The additive noise is assumed to be white and Gaussian with a variance of ρ^{-1} per component. It will also be assumed that the channel matrix, $\mathbf{H}$, is known to the receiver and that all possible transmitted messages, $\mathbf{s}$, are equally likely.

The problem of detecting a vector of symbols (not necessarily binary) transmitted over a MIMO channel is of general interest as it arises frequently in digital communications. Examples include, but are not limited to, the multiuser detection problem in CDMA [1] and communications over a multiple antenna channel [2]. However, while the detection problem is the same for many areas, the structure and assumptions regarding the channel matrix, $\mathbf{H}$, will typically differ depending on the specific context. In the interest of simplicity, we will assume that the channel matrix may be modelled using i.i.d. Gaussian entries with zero mean and finite variance, an assumption motivated by the problem of wireless communication over a richly scattered fading multiple antenna channel [2]. The signal to noise ratio (SNR) of the channel is equal to ρ and we will focus on an analysis of the high SNR regime.

The maximum likelihood (ML) estimate of $\mathbf{s}$, $\hat{\mathbf{s}}_{\mathrm{ML}}$, is well known to be given by

$$\hat{\mathbf{s}}_{\mathrm{ML}} = \arg \min_{\hat{\mathbf{s}} \in \mathcal{B}^m} \|\mathbf{y} - \mathbf{H}\hat{\mathbf{s}}\|^2 \tag{2}$$

where $\|\cdot\|$ denotes the Euclidian norm, i.e. the ML detector, or receiver, selects the message, $\hat{\mathbf{s}}$, which minimizes the distance between the received signals and the hypothesized noise-free message, $\mathbf{H}\hat{\mathbf{s}}$. An error is declared whenever $\hat{\mathbf{s}}_{\mathrm{ML}} \neq \mathbf{s}$ and it well known that the ML detector is optimal in the sense that it minimizes the probability of error. However, for a general channel matrix, $\mathbf{H}$, and vector of received signals, $\mathbf{y}$, the ML detection problem in (2) has been shown to be NP-hard [3] and the full search solution has a complexity of $O(2^m)$ where m is the number of symbols jointly detected. A similar result holds for the sphere decoding algorithm which is able to provide exact solutions to (2) at an expected complexity on the order of $O(2^{\gamma m})$ for some $\gamma \in (0, 1]$ [4]. The complexity is thus, although significantly lower than the full search, still exponential.

Thus, the use of suboptimal (but computationally advantageous) alternatives to ML detection is motivated. However, when applied to a fading channel there is unfortunately often a significant loss

in performance associated with many of the suboptimal alternatives. This is illustrated in Fig. 1 where the probability of error for three different detectors is shown for the case where $\mathbf{H} \in \mathbb{R}^{4\times 4}$. By comparing the ML detector and minimum mean square error (MMSE) detector [2] it can be seen that not only is the MMSE suboptimal, but the rate at which the probability of error tends to zero with increasing SNR is significantly lower than that of the optimal ML detector. This in turn results in a large loss in performance in the high SNR regime. The rate at which the error probability vanishes, or more precisely the slope (in log-log scale) of the error probability curve in the high SNR regime, is commonly referred to as the *diversity* of the detector and it is well known that the MMSE detector has a significantly lower diversity than the ML detector [2]. However, the third curve in Fig. 1 shows the probability of error for a receiver structure known as the semidefinite relaxation (SDR) detector or receiver. The SDR detector was (in the communications literature) first proposed in [5], [6], [7] for CDMA multiuser detection but is applicable for the detection of binary signals transmitted over any MIMO channel on the form of (1). The SDR receiver is based on a convex relaxation technique where the optimization in (2) is simplified by first expanding the feasible set and then applying a rounding procedure to obtain an approximate solution to (2). Note that this statement is also true for the zero forcing (ZF) and MMSE receivers where an unconstrained least squares problem (a regularized least squares problem in the MMSE case) is initially solved and where the symbol estimates are then obtained by componentwise threshold decisions. However, the semidefinite relaxation differs from ZF and MMSE receivers in that the problem is first lifted into a higher dimensional space before the relaxation takes place. From Fig. 1 it is apparent that the SDR receiver, although suboptimal in the sense that it does not achieve the minimum probability of error, does not suffer the loss in diversity experienced by the MMSE receiver.

The main contribution of this work is the analytic proof of the observation above. Namely, if the entries of $\mathbf{H} \in \mathbb{R}^{n\times m}$ are i.i.d. zero mean Gaussian with a finite variance and $n \geq m$, then the SDR receiver achieves the maximum possible receiver diversity. The result is formally stated in Theorem 1 in Section II-B and represents a non-trivial extension of previously known performance guarantees available for the SDR detector, see e.g. [8], [6], [9].

The topic of receiver diversity has received significant attention in the digital communications literature and other low complexity receivers have been designed specifically with diversity in mind. Perhaps, most prominent among these receivers are the lattice-reduction-aided (LRA) receivers [10], [11]. In the LRA receiver one performs a change of basis under which the conditioning of $\mathbf{H}$ is improved and then applies a simple (e.g. ZF, MMSE or decision feedback) detector in the new basis. It has also recently been shown that it is possible to construct (low complexity) full diversity receivers based on these ideas [12],

again under the assumption that $n \geq m$. However, the design philosophies underlying the LRA and SDR detectors are fundamentally different. Were as the LRA is combinatorial in nature the SDR detector is based on the minimization of a continuous function over a convex set. Further, in the LRA receiver it is assumed that the transmitted message belongs to an (infinite) integer lattice which enables the change of basis while in the SDR approach explicit use is made of the binary symbol assumption.

As previously stated, we treat the SDR receiver under the assumption that the channel matrix is i.i.d. Gaussian and real valued. The main reason for this is that the SDR receiver is most easily treated in the real valued case. It should however be mentioned that the extension to the complex case is non-trivial and that numerical results suggest that a theorem, analogous to Theorem 1, may not hold in this case. However, the numerical results also indicate that the loss in diversity (with respect to the ML detector) remains small. We discuss this issue further in Section VI-B. Additionally, the underdetermined ($n < m$) case is treated in Section VI-A. In the latter case our proof of Theorem 1 provides a lower bound on the diversity achieved by the SDR receiver which shows that if $m - n$ is not too large, then the diversity of the SDR is strictly larger than that of the MMSE and ZF receivers.

In Section II we review the SDR receiver and present the main contribution of this work, namely Theorem 1. In Section III a short outline of the proof is given and the rigorous analysis is given in Section IV and Section V. Further, a short discussion of how the results may possibly be generalized to other scenarios is given in Section VI. Also, although it makes no difference for the analytical results, we will in the numerical examples normalize the channel matrix, $\mathbf{H}$, such that each component has a variance of n^{-1}, yielding unit energy symbols at the receiver.

II. Semidefinite Relaxation

The use of semidefinite relaxation for bounding the optimal value of a combinatorial optimization problem was first considered in the late seventies [13] (where it was used to bound the Shannon capacity of a graph). Theoretical work in the nineties [14] along with the introduction of practical methods for solving semidefinite programs [15], [16], [17] made the semidefinite relaxation a viable method for finding approximate solutions to many combinatorial problems. A famous example where the SDR technique can be applied is the *max cut* problem in graph theory [18]. The application of SDR to the detection problem considered herein has also been studied in the communications literature [5], [6], [7].

We will in Section II-A provide a short review of the SDR detector in the communications context. It is not the intention to give a complete treatment of the SDR detector in terms of implementation or to discuss the various improvements which have been proposed but rather to introduce notation and capture

specific assumptions made herein. The reader is instead referred to the original works [5], [6], [7] for a thorough treatment of the SDR detector in the context of digital communications. See also, apart from the above, [19] for a comprehensive collection of results regarding semidefinite programming in general and also specific results regarding the semidefinite relaxation technique.

A. The SDR Detector

In order to introduce the semidefinite relaxation technique it is useful to note that the (non-convex) optimization problem given by

$$\begin{aligned} \min_{\mathbf{X},\,\mathbf{x}} \quad & \mathrm{Tr}(\mathbf{L}\mathbf{X}) \\ \text{s.t.} \quad & \mathrm{diag}(\mathbf{X}) = \mathbf{e} \\ & \mathbf{X} = \mathbf{x}\mathbf{x}^{\mathrm{T}} \end{aligned} \tag{3}$$

where $\mathbf{e}$ is the vector of all ones and where

$$\mathbf{L} \triangleq \begin{bmatrix} \mathbf{H}^{\mathrm{T}}\mathbf{H} & -\mathbf{H}^{\mathrm{T}}\mathbf{y} \\ -\mathbf{y}^{\mathrm{T}}\mathbf{H} & \mathbf{y}^{\mathrm{T}}\mathbf{y} \end{bmatrix}, \quad \mathbf{x} \triangleq \begin{bmatrix} \hat{\mathbf{s}} \\ 1 \end{bmatrix} \tag{4}$$

is equivalent to (2) in the sense that the solution to (2) is easily obtained from the solution to (3) and vice verse [5], [6], [19]. Essentially, the formulation of (3) is obtained by lifting (2) into a higher dimension where the criterion is linear in the optimization variable. The rank one constraint on $\mathbf{X}$ along with the diagonal constraint ensure there is a one to one correspondence between the feasible sets of (2) and (3). The optimal point of (2) is related to the optimal point of (3) through $\mathbf{x}$ as shown in (4).

As (3) and (2) are equivalent they are also equally hard to solve from a complexity theoretic point of view. In particular, it follows from [3] that (5) is also NP-hard in general. However, consider now instead the optimization problem given by

$$\begin{aligned} \min_{\mathbf{X}} \quad & \mathrm{Tr}(\mathbf{L}\mathbf{X}) \\ \text{s.t.} \quad & \mathrm{diag}(\mathbf{X}) = \mathbf{e} \\ & \mathbf{X} \succeq \mathbf{0} \end{aligned} \tag{5}$$

where $\mathbf{X} \succeq \mathbf{0}$ means that $\mathbf{X}$ is symmetric and positive definite. Since $\mathbf{X} = \mathbf{x}\mathbf{x}^{\mathrm{T}}$ implies $\mathbf{X} \succeq \mathbf{0}$ it follows that (5) represents a relaxation of (3). The problem in (5) is referred to as the semidefinite relaxation of (3) (or equivalently (2)) and serves as the basis for the semidefinite relaxation detector.

It is useful to note that (5) is a *convex* problem which can be efficiently solved in polynomial time [16], [20]. In particular, there is an interior point algorithm which solves (5) to any fixed precision in $O(m^{3.5})$ time [21], see also [5] where this algorithm is presented in the digital communications context. In practice,

only a few iterations with a complexity comparable to that of inverting an m by m matrix are required in order to obtain an approximate solution to (5).

It is straightforward to see that when the optimal solution to (5) is rank one it is also an optimal solution to (3). The existence of rank one solutions to (5) is however by no means guaranteed and in general, the solution to (5) can only serve as a basis for obtaining an approximate solution to (3). In fact, it is possible to characterize exactly (in terms of $\mathbf{H}$, $\mathbf{s}$ and $\mathbf{v}$) when (5) will and will not have rank one solutions, see [22] for necessary and sufficient conditions.

When the optimal point of (5) is not rank one, some type of rounding procedure has to be used to round the optimal point of (5) to a point in the feasible set (3). There are several suggestions for this in the literature. Among the more powerful approaches are a randomization technique [18], [6] and an approximation using the dominant eigenvector [5]. Numerical evidence suggests that the randomization technique results in superior error performance. We shall however consider the very simple strategy of simply using the signs of the last column of $\mathbf{X}^\star$ where $\mathbf{X}^\star$ is an optimal point of (5). This approach was also mentioned in [5] but discarded in favor of the (superior) eigenvector approach. However, as the simpler approach already achieves the maximum diversity we shall only consider this approach in detail. It should however be noted that the proof extends to the dominant eigenvector case in a straightforward manner by simply appealing to results regarding the continuity of eigenvectors corresponding to distinct (multiplicity one) eigenvalues.

To summarize, we obtain the SDR estimate, $\hat{\mathbf{s}}_{\mathrm{SDR}}$ as follows. Let $\mathbf{X}^\star$ be the minimizer of (5). Then $\hat{\mathbf{s}}_{\mathrm{SDR}}$ is defined according to

$$[\hat{\mathbf{s}}_{\mathrm{SDR}}]_i \triangleq \mathrm{sgn}([\mathbf{X}^\star]_{i,m+1}), \quad i = 1, \ldots, m \tag{6}$$

where

$$\mathrm{sgn}(x) = \begin{cases} 1 & x > 0 \\ -1 & x \leq 0 \end{cases}$$

is the sign function, i.e. $\hat{\mathbf{s}}_{\mathrm{SDR}}$ is given by the signs of the last column of $\mathbf{X}^\star$. Note that although it is possible for (5) to have several optimal solutions it is always possible to pick some unique optimizer, $\mathbf{X}^\star$, from the optimal set. Thus, it can be assumed that $\hat{\mathbf{s}}_{\mathrm{SDR}}$ is uniquely determined by $\mathbf{y}$ and $\mathbf{H}$.

Finally, it should be mentioned that extensions to the original semidefinite relaxation detectors have appeared in the literature. These include for example extensions to M-PSK constellations [23] and M-QAM constellations [24]. However, the analysis of these extensions is not treated herein.

B. SDR Performance

The extraordinary performance of the SDR technique in many areas have been a motivating reason for its study and there are results in the literature regarding the quality of the semidefinite relaxation approximation of (3) for more or less arbitrary choices of the matrix $\mathbf{L}$ (in (4)). These include the bound in [8] which is a generalization of a previous result for the max cut problem [18]. There are also some results relating the semidefinite relaxation to other relaxations available for binary quadratic programs (such as (2)) [25].

In the context of digital communications it has previously been shown that several low complexity detectors may be viewed as further relaxations of the SDR detector [6]. Notably, these low complexity detectors include both the ZF and MMSE detectors and give strong support for the SDR approach although the results in [6] relate to the objective values of the relaxations rather than directly to the quality of the estimates, $\hat{\mathbf{s}}$. Further, a probabilistic bound on the difference in optimal objective value between (5) and (3) was given in [9] for the large system limit. Also, as previously mentioned, the conditions for rank one solutions to (5) were complectly characterized in [22] where it was also established that the detector was free of an error floor under the assumption that $\mathbf{H}^{\mathrm{T}}\mathbf{H}$ is full rank. However, the result in [22] does not extend to a statement regarding the diversity. Specifically, it is possible to show (using the result of [22]) that an alternative SDR receiver which calls an error whenever (5) is not of rank one would not have the maximum diversity. In other words, the second phase of the SDR receiver where high rank solutions are used to obtain symbol estimates is crucial to the SDR performance and must be taken into account in the analysis.

The main contribution of this work is a rather strong statement regarding SDR performance when applied to a fading channel, namely that under the model in (1) with an i.i.d. Gaussian channel for which $n \geq m$ the SDR detector will have a diversity equal to that of the optimal, ML, detector. Loosely speaking, although suboptimal, the SDR detector will have an error probability which vanishes at the same rate as the ML detector in the high SNR limit and the loss due to suboptimality will be a shift in SNR and not a loss of *diversity*. We formally state this as follows.

Theorem 1: Assume that $\mathbf{H} \in \mathbb{R}^{n\times m}$ in (1) consist of i.i.d. Gaussian entries of zero mean and fixed (non-zero) variance. Assume further that $n \geq m$. Then

$$\lim_{\rho\to\infty} \frac{\ln \mathrm{P}\left(\hat{\mathbf{s}}_{\mathrm{SDR}} \neq \mathbf{s}\right)}{\ln \rho} = \lim_{\rho\to\infty} \frac{\ln \mathrm{P}\left(\hat{\mathbf{s}}_{\mathrm{ML}} \neq \mathbf{s}\right)}{\ln \rho} = -\frac{n}{2}.$$

It is important to note that the SDR (and maximum) diversity is $\frac{n}{2}$ in this case and not n. This is because we explicitly consider a real valued channel matrix (1) as opposed to the complex channel case

more frequently studied in the literature. It is straightforward to show the maximum achievable diversity in this case is $\frac{n}{2}$ by extending the proof of [26] to cover the real valued case. In the case of ZF and MMSE the diversity is $\frac{n-m+1}{2}$ which can be seen by following the argument of Section 8.5.1. in [2] with a real valued channel matrix.

Following [27] we will throughout this work make use of the symbol $\doteq$ to denote *exponential equality*, defined according to

$$f(\rho) \doteq \rho^{-d} \qquad \Leftrightarrow \qquad \lim_{\rho\to\infty} \frac{\ln f(\rho)}{\ln \rho} = -d. \tag{7}$$

Similar definitions will also apply to the symbols $\dot{\leq}$ and $\dot{\geq}$. For reference, we list the most important properties of the exponential equality in Appendix I. Using (7) generally allows for a more compact (and suggestive) notation and in this notation the statement of Theorem 1 becomes

$$\mathrm{P}\left(\hat{\mathbf{s}}_{\mathrm{SDR}} \neq \mathbf{s}\right) \doteq \mathrm{P}\left(\hat{\mathbf{s}}_{\mathrm{ML}} \neq \mathbf{s}\right) \doteq \rho^{-\frac{n}{2}}.$$

Now, most of remaining part of this work is devoted to the proof of Theorem 1. The formal proof is divided into several lemmas presented in Section IV and Section V. However, before presenting the proof in full, a short outline is given in Section III.

III. The SDR Diversity Proof, Outline

Note that due to the symmetry of the problem (and the detector) it can without loss of generality be assumed that $\mathbf{s} = \mathbf{e}$ was transmitted. This will also be done in the sequel. In the $m = 2$ case it is possible to graphically illustrate the feasible set, $\mathcal{X}$, of (5) in order to gain intuition. To this end, consider parameterizing $\mathbf{X} \in \mathcal{X}$ as in [28] or [5], i.e. according to

$$\mathbf{X} = \begin{bmatrix} 1 & x & y \\ x & 1 & z \\ y & z & 1 \end{bmatrix}.$$

The feasible set, $\mathcal{X}$, is illustrated in Fig. 2. The rank one matrix, $\mathbf{X}_{\mathbf{e}}$, that corresponds to the transmitted message, $\mathbf{s} = \mathbf{e}$, is also indicated in the figure.

Intuitively, one can characterize the error events of the SDR receiver as follows. When the optimal point of (5), $\mathbf{X}^\star$, is close to $\mathbf{X}_{\mathbf{e}}$ then the rounding procedure described in Section II will be able to recover the correct rank one matrix, namely $\mathbf{X}_{\mathbf{e}}$. It is only when the optimal point of (5) is far from $\mathbf{X}_{\mathbf{e}}$ that an error can occur.

Consider now the introduction of a hyperplane, $\mathcal{H}$, as in Fig. 2 that separates the points in $\mathcal{X}$ that are close to and far from $\mathbf{X}_{\mathbf{e}}$. Specifically, let $\mathcal{X}_+$ be the points in $\mathcal{X}$ that are on the same side of $\mathcal{H}$ as $\mathbf{X}_{\mathbf{e}}$

and let $\mathcal{X}_-$ be the points on the other side. Assume also that $\mathcal{H}$ is chosen such that points in $\mathcal{X}_+$ are rounded off to $\mathbf{X_e}$. Let us also first consider the zero noise case, i.e. when $\mathbf{v} = \mathbf{0}$. In this case $\mathbf{X_e}$ is always optimal for (5) with a criterion value equal to 0. Further, let $\tau \geq 0$ be given by

$$\tau = \min_{\mathbf{X} \in \mathcal{X} \cap \mathcal{H}} \mathrm{Tr}(\mathbf{LX}),$$

i.e. τ is the minimum objective value over the intersection of the hyperplane and the feasible set, assuming $\mathbf{v} = \mathbf{0}$. As the criterion function, $\mathrm{Tr}(\mathbf{LX})$, is linear and $\mathcal{X}$ is convex it follows that the criterion function for any $\mathbf{X} \in \mathcal{X}_-$ will also satisfy $\mathrm{Tr}(\mathbf{LX}) \geq \tau$.

Now allow for $\mathbf{v} \neq \mathbf{0}$ but assume that $\|\mathbf{v}\|$ is significantly smaller than τ. In this case, $\mathrm{Tr}(\mathbf{LX_e})$ is still small as $\mathrm{Tr}(\mathbf{LX_e})$ is continuous in $\mathbf{v}$. At the same time it is guaranteed that $\mathrm{Tr}(\mathbf{LX})$ is not significantly smaller than τ for any $\mathbf{X} \in \mathcal{X}_-$, again since $\mathrm{Tr}(\mathbf{LX})$ is continuous in $\mathbf{v}$. This implies that there is a point in $\mathcal{X}_+$ with a criterion value close to zero, while all points in $\mathcal{X}_-$ have objective values which are at least on the order of τ. In other words, the optimum over $\mathcal{X}$ must belong to $\mathcal{X}_+$ and therefore be close to $\mathbf{X_e}$. This in turn implies that no error is made by the SDR receiver. In short, it is sufficient that τ is large in comparison with the noise in order for the detector to make a correct decision. This statement is also made rigorously by Lemma 1 in Section IV.

The proof of Theorem 1 follows the heuristic argument given above and is divided into two parts. The first part, is concerned with proving that the error probability of the SDR detector is, in the high SNR regime, governed by the probability that τ is *atypically* small rather than the probability that $\mathbf{v}$ is atypically large. This statement is formalized by Lemma 2 in Section IV. Note that the technique of interpreting typical errors as caused by particularly bad channels (in our case channels which cause τ to be small) is common in the literature, see e.g. [2]. It is also similar in many respects to the analysis of coded multiple antenna systems where errors are typically caused by channels in *outage* [27].

The second part of the proof, contained in Section V, is concerned with bounding the probability that τ is atypically small. Note that in order for τ to be small there must be at least one $\mathbf{X} \in \mathcal{X} \cap \mathcal{H}$ for which $\mathrm{Tr}(\mathbf{LX})$ is small. In essence, the technique used to establish our bound on the probability of τ being small can be summarized as follows.

1) Cover $\mathcal{X} \cap \mathcal{H}$ (or more precisely a set isomorphic to $\mathcal{X} \cap \mathcal{H}$) with ϵ-balls and bound the probability that each specific ϵ-ball contains an $\mathbf{X}$ for which $\mathrm{Tr}(\mathbf{LX})$ is small.
2) Count the number of ϵ-balls required to cover $\mathcal{X} \cap \mathcal{H}$ and use the union bound to bound the probability that τ is small.

Much of the difficulty of the proof stems from that the probability that each ϵ-ball contains an $\mathbf{X}$ for

which $\mathrm{Tr}(\mathbf{LX})$ is small depends on where in $\mathcal{X} \cap \mathcal{H}$ the ϵ-ball is located. Also, the technically most challenging part of the proof relates to counting the number of ϵ-balls required to cover certain subsets of $\mathcal{X} \cap \mathcal{H}$. The analysis of each particular ϵ-ball is provided by Lemma 3 and the counting argument is captured in Lemma 4 in Section V. The proof of Theorem 1, given at the end of Section V, then follows by combining Lemma 3 and Lemma 4.

IV. The SDR Diversity Proof, Part I

The purpose of this section is to give rigorous justification of the first part of the heuristic argument given in Section III and show that the noise, $\mathbf{v}$, can effectively be removed from (or integrated out of) the analysis of the receiver diversity. To this end, we will begin by giving a proper definition of some of the concepts appearing in the heuristic argument.

First of all, the feasible set, $\mathcal{X}$, of (5) is given by

$$\mathcal{X} \triangleq \{\mathbf{X} \in \mathbb{S}^{m+1} \mid \mathrm{diag}(\mathbf{X}) = \mathbf{e},\ \mathbf{X} \succeq \mathbf{0}\} \tag{8}$$

where $\mathbb{S}^{m+1}$ denotes the set of symmetric matrices. Let $\mathcal{H}$ be the hyperplane (or affine subset of $\mathbb{S}^{m+1}$) given by

$$\mathcal{H} \triangleq \{\mathbf{X} \in \mathbb{S}^{m+1} \mid \mathrm{Tr}(\mathbf{M}\mathbf{X}\mathbf{M}^{\mathrm{T}}) = 1\} \tag{9}$$

where

$$\mathbf{M} \triangleq \begin{bmatrix} \mathbf{I} & -\mathbf{e} \end{bmatrix} \in \mathbb{R}^{m \times m+1}. \tag{10}$$

It will later be established that an $\mathcal{H}$ chosen this way is sufficient for separating point close to $\mathbf{X_e}$ from points far from $\mathbf{X_e}$. The optimal value of $\mathrm{Tr}(\mathbf{LX})$ over the intersection set $\mathcal{X} \cap \mathcal{H}$ is under the zero noise, $\mathbf{v} = \mathbf{0}$, assumption given by

$$\tau \triangleq \min_{\mathbf{X} \in \mathcal{X} \cap \mathcal{H}} \mathrm{Tr}(\mathbf{L_0 X}) \tag{11}$$

where

$$\mathbf{L_0} \triangleq \begin{bmatrix} \mathbf{Q} & -\mathbf{Q}\mathbf{e} \\ -\mathbf{e}^{\mathrm{T}}\mathbf{Q} & \mathbf{e}^{\mathrm{T}}\mathbf{Q}\mathbf{e} \end{bmatrix} = \mathbf{M}^{\mathrm{T}}\mathbf{Q}\mathbf{M}$$

and $\mathbf{Q} \triangleq \mathbf{H}^{\mathrm{T}}\mathbf{H}$. Note that $\mathbf{L_0}$ is equal to $\mathbf{L}$ in (4) when $\mathbf{v} = \mathbf{0}$. It is also straightforward to show that τ is equivalently given by

$$\tau = \inf_{\mathbf{Y} \in \mathcal{Y}} \mathrm{Tr}(\mathbf{QY}) \tag{12}$$

where

$$\mathcal{Y} \triangleq \mathbf{M}(\mathcal{X} \cap \mathcal{H})\mathbf{M}^{\mathrm{T}} = \tilde{\mathcal{Y}} \cap \{\mathbf{Y} \in \mathbb{S}^{m} \mid \mathrm{Tr}(\mathbf{Y}) = 1\} \tag{13}$$

and

$$\tilde{\mathcal{Y}} = \mathbf{M}\mathcal{X}\mathbf{M}^{\mathrm{T}}. \tag{14}$$

The set $\tilde{\mathcal{Y}}$ is a linear mapping of $\mathcal{X} \subset \mathbb{S}^{m+1}$ onto $\mathbb{S}^m$ given by $\mathbf{M}\mathbf{X}\mathbf{M}^{\mathrm{T}}$ under which the criterion $\mathrm{Tr}(\mathbf{L_0}\mathbf{X})$ and $\mathcal{H}$ have a somewhat simpler structure. Note also that $\tilde{\mathcal{Y}}$ is convex since it is a linear transformation of a convex set. The main reason for introducing (12) is that it is frequently more convenient to work with (12) rather than with (11) directly.

We are now able to pose and prove the first lemma regarding the error probability of the SDR detector. In essence, we wish to establish that a large τ is sufficient for correct detection. These statements are captured by Lemma 1 given below (note again that $\mathbf{s} = \mathbf{e}$ is assumed to be the transmitted message).

Lemma 1: Let τ be given by (11). Then

$$\tau > 4\|\mathbf{v}\|^2 \Rightarrow \hat{\mathbf{s}}_{\mathrm{SDR}} = \mathbf{e}.$$

Proof: We will first prove the lemma under the assumption that the optimal point of (5) is rank deficient and then argue that this assumption can be made without loss of generality. Thus, consider an $\mathbf{X} \in \mathcal{X}$ for which $\mathbf{X} \succcurlyeq \mathbf{0}$ ($\mathbf{X}$ is positive semidefinite but not positive definite) and partition $\mathbf{X}$ as

$$\mathbf{X} = \begin{bmatrix} \mathbf{A}^{\mathrm{T}} \\ \mathbf{a}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \mathbf{A} & \mathbf{a} \end{bmatrix} = \begin{bmatrix} \mathbf{A}^{\mathrm{T}}\mathbf{A} & \mathbf{A}^{\mathrm{T}}\mathbf{a} \\ \mathbf{a}^{\mathrm{T}}\mathbf{A} & \mathbf{a}^{\mathrm{T}}\mathbf{a} \end{bmatrix}$$

where $\mathbf{A} \in \mathbb{R}^{m\times m}$ and $\mathbf{a} \in \mathbb{R}^m$. Note that this is possible since $\mathbf{X}$ has at most rank m. Note also that $\|\mathbf{a}\| = 1$ follows from $\mathrm{diag}(\mathbf{X}) = \mathbf{e}$. Further, note that the matrix $\mathbf{L}$ defined in (4) can be written as

$$\mathbf{L} \triangleq \begin{bmatrix} \mathbf{H}^{\mathrm{T}}\mathbf{H} & -\mathbf{H}^{\mathrm{T}}\mathbf{y} \\ -\mathbf{y}^{\mathrm{T}}\mathbf{H} & \mathbf{y}^{\mathrm{T}}\mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{H}^{\mathrm{T}} \\ -\mathbf{y}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \mathbf{H} & -\mathbf{y} \end{bmatrix}.$$

Thus,

$$\begin{aligned}
\mathrm{Tr}(\mathbf{L}\mathbf{X}) =& \mathrm{Tr}\left(\begin{bmatrix} \mathbf{H}^{\mathrm{T}} \\ -\mathbf{y}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \mathbf{H} & -\mathbf{y} \end{bmatrix} \begin{bmatrix} \mathbf{A}^{\mathrm{T}} \\ \mathbf{a}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \mathbf{A} & \mathbf{a} \end{bmatrix} \right) \\
=& \mathrm{Tr}\left(\begin{bmatrix} \mathbf{H} & -\mathbf{y} \end{bmatrix} \begin{bmatrix} \mathbf{A}^{\mathrm{T}} \\ \mathbf{a}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \mathbf{A} & \mathbf{a} \end{bmatrix} \begin{bmatrix} \mathbf{H}^{\mathrm{T}} \\ -\mathbf{y}^{\mathrm{T}} \end{bmatrix} \right) \\
=& \mathrm{Tr}((\mathbf{H}\mathbf{A}^{\mathrm{T}} - \mathbf{y}\mathbf{a}^{\mathrm{T}})(\mathbf{H}\mathbf{A}^{\mathrm{T}} - \mathbf{y}\mathbf{a}^{\mathrm{T}})^{\mathrm{T}}) \\
=& \|\mathbf{H}\mathbf{A}^{\mathrm{T}} - \mathbf{y}\mathbf{a}^{\mathrm{T}}\|^2
\end{aligned}$$

where $\|\cdot\|$ above refers to the the Frobenius norm. Now, the model of (1) for $\mathbf{s} = \mathbf{e}$ yields (through $\mathbf{y}$)

$$\mathrm{Tr}(\mathbf{L}\mathbf{X}) = \|\mathbf{H}(\mathbf{A}^{\mathrm{T}} - \mathbf{e}\mathbf{a}^{\mathrm{T}}) - \mathbf{v}\mathbf{a}^{\mathrm{T}}\|^2.$$

Note that

$$\begin{aligned}&\|\mathbf{H}(\mathbf{A}-\mathbf{e}\mathbf{a}^{\mathrm{T}})-\mathbf{v}\mathbf{a}^{\mathrm{T}}\|\\ \geq&\|\mathbf{H}(\mathbf{A}-\mathbf{e}\mathbf{a}^{\mathrm{T}})\|-\|\mathbf{v}\mathbf{a}^{\mathrm{T}}\|\\ =&\|\mathbf{H}(\mathbf{A}-\mathbf{e}\mathbf{a}^{\mathrm{T}})\|-\|\mathbf{v}\|\end{aligned}$$

where the last equality follows from $\|\mathbf{a}\|=1$. Thus, whenever

$$\|\mathbf{H}(\mathbf{A}-\mathbf{e}\mathbf{a}^{\mathrm{T}})\|>2\|\mathbf{v}\|\Leftrightarrow\|\mathbf{H}(\mathbf{A}-\mathbf{e}\mathbf{a}^{\mathrm{T}})\|^2>4\|\mathbf{v}\|^2$$

it follows that

$$\mathrm{Tr}(\mathbf{L}\mathbf{X})>\|\mathbf{v}\|^2. \tag{15}$$

At the same time, for

$$\mathbf{X}_{\mathbf{e}}\triangleq\begin{bmatrix}\mathbf{e}\\1\end{bmatrix}\begin{bmatrix}\mathbf{e}^{\mathrm{T}} & 1\end{bmatrix}$$

it follows that

$$\begin{aligned}\mathrm{Tr}(\mathbf{L}\mathbf{X}_{\mathbf{e}})=&\mathrm{Tr}\left(\begin{bmatrix}\mathbf{H}^{\mathrm{T}}\\-\mathbf{y}^{\mathrm{T}}\end{bmatrix}\begin{bmatrix}\mathbf{H} & -\mathbf{y}\end{bmatrix}\begin{bmatrix}\mathbf{e}\\1\end{bmatrix}\begin{bmatrix}\mathbf{e}^{\mathrm{T}} & 1\end{bmatrix}\right)\\ =&\mathrm{Tr}\left(\begin{bmatrix}\mathbf{H} & -\mathbf{y}\end{bmatrix}\begin{bmatrix}\mathbf{e}\\1\end{bmatrix}\begin{bmatrix}\mathbf{e}^{\mathrm{T}} & 1\end{bmatrix}\begin{bmatrix}\mathbf{H}^{\mathrm{T}}\\-\mathbf{y}^{\mathrm{T}}\end{bmatrix}\right)\\ =&\mathrm{Tr}((\mathbf{H}\mathbf{e}-\mathbf{y})(\mathbf{H}\mathbf{e}-\mathbf{y})^{\mathrm{T}})\\ =&\|\mathbf{H}\mathbf{e}-\mathbf{y}\|^2=\|\mathbf{v}\|^2.\end{aligned} \tag{16}$$

Thus, by (15) and (16), it follows that

$$\|\mathbf{H}(\mathbf{A}-\mathbf{e}\mathbf{a}^{\mathrm{T}})\|^2>4\|\mathbf{v}\|^2\ \Rightarrow\ \mathrm{Tr}(\mathbf{L}\mathbf{X})>\mathrm{Tr}(\mathbf{L}\mathbf{X}_{\mathbf{e}}) \tag{17}$$

which implies that $\mathbf{X}$ can not be optimal for (5) if

$$\|\mathbf{H}(\mathbf{A}-\mathbf{e}\mathbf{a}^{\mathrm{T}})\|^2>4\|\mathbf{v}\|^2\Leftrightarrow\|\mathbf{H}(\mathbf{A}-\mathbf{e}\mathbf{a}^{\mathrm{T}})\|>2\|\mathbf{v}\|.$$

Now, note that

$$(\mathbf{A}-\mathbf{e}\mathbf{a}^{\mathrm{T}})=\mathbf{M}\begin{bmatrix}\mathbf{A}^{\mathrm{T}}\\\mathbf{a}^{\mathrm{T}}\end{bmatrix},$$

for $\mathbf{M}$ defined in (10) and

$$\begin{aligned}
&\|\mathbf{H}(\mathbf{A}-\mathbf{e}\mathbf{a}^{\mathrm{T}})\|^2 \\
=&\mathrm{Tr}\left(\mathbf{H}\mathbf{M}\begin{bmatrix}\mathbf{A}^{\mathrm{T}}\\ \mathbf{a}^{\mathrm{T}}\end{bmatrix}\begin{bmatrix}\mathbf{A} & \mathbf{a}\end{bmatrix}\mathbf{M}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\right) \\
=&\mathrm{Tr}\left(\mathbf{H}^{\mathrm{T}}\mathbf{H}\mathbf{M}\begin{bmatrix}\mathbf{A}^{\mathrm{T}}\\ \mathbf{a}^{\mathrm{T}}\end{bmatrix}\begin{bmatrix}\mathbf{A} & \mathbf{a}\end{bmatrix}\mathbf{M}^{\mathrm{T}}\right) \\
=&\mathrm{Tr}(\mathbf{H}^{\mathrm{T}}\mathbf{H}\mathbf{M}\mathbf{X}\mathbf{M}^{\mathrm{T}}). \qquad (18)
\end{aligned}$$

Let $\mathbf{X}^\star \in \mathcal{X}$ be the optimal point for (5) and let $\mathbf{Y}^\star \in \tilde{\mathcal{Y}}$ be given by $\mathbf{Y}^\star \triangleq \mathbf{M}\mathbf{X}^\star\mathbf{M}^{\mathrm{T}}$. Note that

$$\mathrm{Tr}(\mathbf{Q}\mathbf{Y}^\star) \leq 4\|\mathbf{v}\|^2$$

for $\mathbf{Q} = \mathbf{H}^{\mathrm{T}}\mathbf{H}$ as otherwise $\mathbf{X}^\star$ would not be optimal due to (17) and (18).

Assume (as in the lemma) that

$$\tau > 4\|\mathbf{v}\|^2.$$

This implies that $\mathrm{Tr}(\mathbf{Q}\mathbf{Y}) > 4\|\mathbf{v}\|^2$ for any $\mathbf{Y} \in \mathcal{Y}$. The same conclusion could also be drawn for any $\mathbf{Y} \in \tilde{\mathcal{Y}}$ which satisfies $\mathrm{Tr}(\mathbf{Y}) \geq 1$. This follows since $\tilde{\mathcal{Y}}$ is a convex set which contains $\mathbf{0}$ (since $\mathbf{0} = \mathbf{M}\mathbf{X}_{\mathbf{e}}\mathbf{M}^{\mathrm{T}}$). That is, if there were $\mathbf{Y} \in \tilde{\mathcal{Y}}$ for which $\mathrm{Tr}(\mathbf{Y}) \geq 1$ and $\mathrm{Tr}(\mathbf{Q}\mathbf{Y}) \leq 4\|\mathbf{v}\|^2$ then $\tilde{\mathbf{Y}} \triangleq \gamma\mathbf{Y} \in \mathcal{Y}$ for some $\gamma \in (0,1]$ and $\mathrm{Tr}(\mathbf{Q}\tilde{\mathbf{Y}}) \leq 4\|\mathbf{v}\|^2$ contrary to the assumption.

Thus, under the assumption of the lemma, it follows that

$$\mathrm{Tr}(\mathbf{Y}^\star) < 1$$

and $\|\mathrm{diag}(\mathbf{Y}^\star)\|_\infty < 1$ as $\mathbf{Y}^\star \succeq \mathbf{0}$ implies that $\mathbf{Y}^\star$ has positive diagonal elements. Now, partition $\mathbf{X}^\star$ as

$$\mathbf{X}^\star = \begin{bmatrix}\mathbf{B} & \mathbf{b}\\ \mathbf{b}^{\mathrm{T}} & 1\end{bmatrix}$$

where $\mathrm{diag}(\mathbf{B}) = \mathbf{e}$ due to $\mathrm{diag}(\mathbf{X}^\star) = \mathbf{e}$. Computing $\mathbf{Y}^\star$ explicitly under this partitioning yields

$$\mathbf{Y}^\star = \mathbf{M}\mathbf{X}^\star\mathbf{M}^{\mathrm{T}} = \mathbf{B} - \mathbf{e}\mathbf{b}^{\mathrm{T}} - \mathbf{b}\mathbf{e}^{\mathrm{T}} + \mathbf{e}\mathbf{e}^{\mathrm{T}}$$

which implies

$$\|\mathbf{e}-\mathbf{b}\|_\infty = \tfrac{1}{2}\|\mathrm{diag}(\mathbf{Y}^\star)\|_\infty < \tfrac{1}{2}$$

since $\mathrm{diag}(\mathbf{Y}^\star) = 2\mathbf{e} - 2\mathbf{b}$. Thus, the rounding procedure given in (6) will round the last column of $\mathbf{X}^\star$, namely $\mathbf{b}$, to $\mathbf{e}$ and it follows that $\hat{\mathbf{s}}_{\mathrm{SDR}} = \mathbf{e}$.

What remains now is to show that the optimal point of (5) must be rank deficient. By applying the result in [29] it is known that there will always be a rank deficient optimal point. A potential problem could arise if there are several optimal points, some of which are full rank. We will however show this that this is not possible.

In order for any optimal point of (5) to be full rank, all off diagonal elements of $\mathbf{L}$ in (4) must be identically zero. This follows since otherwise there would be a search direction in the nullspace of $\mathrm{diag}(\mathbf{X}) = \mathbf{e}$ for which the criterion function would decrease, contradicting the optimality of any full rank $\mathbf{X}$. Thus $\mathbf{H}^{\mathrm{T}}\mathbf{H}$ has zero off diagonal elements (as it appears in $\mathbf{L}$) and $\mathbf{H}$ has orthogonal columns. In this special case the SDR will always have rank one solutions which are unique as long as the ML problem has a unique solution [22]. However, the assumption that $\tau > 4\|\mathbf{v}\|^2$ implies that

$$\|\mathbf{y} - \mathbf{H}\mathbf{e}\|^2 < \|\mathbf{y} - \mathbf{H}\hat{\mathbf{s}}\|^2$$

for any $\hat{\mathbf{s}} \in \mathcal{B}^m$, $\hat{\mathbf{s}} \neq \mathbf{e}$, and it follows that the ML solution is unique. Therefore, there are no full rank solutions under the assumption in the lemma. This completes the proof. ■

Essentially, Lemma 1 states that for an error to occur in the high SNR regime one of two thing must happen. Either τ is atypically small or $\mathbf{v}$ is atypically large. As stated in Section III it can be argued that the probability of the former event outweighs the probability of the latter. This is formally stated by the following Lemma which concludes this section.

Lemma 2: Let τ be given by (11). Then

$$\mathrm{P}\left(\tau \leq \rho^{-1}\right) \dot{\leq} \rho^{-d} \qquad \Rightarrow \qquad \mathrm{P}\left(\hat{\mathbf{s}}_{\mathrm{SDR}} \neq \mathbf{e}\right) \dot{\leq} \rho^{-d}. \tag{19}$$

Proof: Assume (as was done in the lemma) that

$$\mathrm{P}\left(\tau \leq \rho^{-1}\right) \dot{\leq} \rho^{-d}.$$

This, combined with $\mathrm{P}\left(\tau \leq \rho^{-1}\right) \leq 1$, implies that for any arbitrarily small $\delta > 0$ there is a constant, c, for which

$$\mathrm{P}\left(\tau \leq \rho^{-1}\right) \leq c\rho^{-d+\delta}$$

for all $\rho \geq 0$. Now, by Lemma 1,

$$p_e \triangleq \mathrm{P}\left(\hat{\mathbf{s}} \neq \mathbf{e}\right) \leq \mathrm{P}\left(\tau \leq 4\|\mathbf{v}\|^2\right).$$

Introduce a Gaussian vector, $\mathbf{w} \in \mathbb{R}^n$, with i.i.d. zero mean elements of variance one and note that $\rho^{-1}\|\mathbf{w}\|^2$ has the same distribution as $\|\mathbf{v}\|^2$. Let $f_{\|\mathbf{w}\|^2}(\gamma)$ denote the probability density function of

$\gamma = \|\mathbf{w}\|^2$. As τ is independent of $\mathbf{v}$ (and $\mathbf{w}$) it follows that,

$$\begin{aligned} p_e &\leq \mathrm{P}\left(\tau \leq 4\rho^{-1}\|\mathbf{w}\|^2\right) \\ &= \int_0^\infty \mathrm{P}\left(\tau \leq 4\rho^{-1}\|\mathbf{w}\|^2 \mid \|\mathbf{w}\|^2 = \gamma\right) f_{\|\mathbf{w}\|^2}(\gamma) d\gamma \\ &= \int_0^\infty \mathrm{P}\left(\tau \leq 4\rho^{-1}\gamma\right) f_{\|\mathbf{w}\|^2}(\gamma) d\gamma \\ &\leq c 4^{d-\delta}\rho^{-d+\delta} \int_0^\infty \gamma^{d-\delta} f_{\|\mathbf{w}\|^2}(\gamma) d\gamma \\ &= c 4^{d-\delta}\rho^{-d+\delta} \mathrm{E}\left\{\|\mathbf{w}\|^{2(d-\delta)}\right\} = c'\rho^{-d+\delta} \end{aligned}$$

for some c' independent of ρ. Note that $c' < \infty$ follows since $\|\mathbf{w}\|$ has finite moments. Thus,

$$p_e \dot{\leq} \rho^{-d+\delta}.$$

However, as the relation holds for arbitrary small $\delta > 0$ it follows that

$$p_e \dot{\leq} \rho^{-d}$$

which concludes the proof. ■

V. The SDR Diversity Proof, Part II

Let τ be given by (11) or equivalently (12). In light of Lemma 2 all that remains to be done in order to prove Theorem 1 is to provide a bound on

$$\mathrm{P}\left(\tau \leq \rho^{-1}\right)$$

in the high SNR limit. Note however that at this point the variable ρ^{-1} is just a dummy variable and we can, and will, replace ρ^{-1} by ϵ and study the probability that $\tau \leq \epsilon$ for small $\epsilon > 0$. Thus, what remains to be done is to bound $\mathrm{P}\left(\tau \leq \epsilon\right)$ around $\epsilon = 0$. We will also in the remaining part of this work focus on the optimization problem given in (12) rather than the equivalent problem in (11).

The probability that $\mathrm{Tr}(\mathbf{QY}) \leq \epsilon$ for some particular $\mathbf{Y} \in \mathcal{Y}$ will generally depend on the specific $\mathbf{Y}$ considered (as mentioned in Section III). In order to deal with this we shall first partition $\mathcal{Y}$ into a finite number of subsets $\{\mathcal{Y}_i\}$,

$$\mathcal{Y} \subset \bigcup_i \mathcal{Y}_i,$$

such that $\mathrm{P}\left(\mathrm{Tr}(\mathbf{QY}) \leq \epsilon\right)$ is more or less constant for all $\mathbf{Y}$ within one such subset. Then, the probability that $\tau \leq \epsilon$ will be bounded by applying the union bound according to

$$\mathrm{P}\left(\tau \leq \epsilon\right) \leq \sum_i \mathrm{P}\left(\tau_i \leq \epsilon\right) \tag{20}$$

where

$$\tau_i \triangleq \inf_{\mathbf{Y}\in\mathcal{Y}_i} \mathrm{Tr}(\mathbf{QY})$$

and where by property (37b) in Appendix I it is known that the sum in (20) will in the exponential equality sense be given (or completely dominated) by its maximal term.

It is interesting to note that this corresponds to the identification of *typical* error events (or classes of error events), which is closely related to the analysis of typical *outage* events in [27]. However, in [27] typical events where identified by classifying particularly bad channels, $\mathbf{H}$, while here, we shall use the concept to identify particularly troublesome subsets of $\mathcal{Y}$. In essence, we shall partition $\mathcal{Y}$ based on the eigenvalues of $\mathbf{Y} \in \mathcal{Y}$ (or how close to singular $\mathbf{Y}$ is). Then the subset which dominates (20) will be found by optimizing over the possible eigenvalue combinations. Note also that these subsets will generally depend on ϵ but that we will adopt a somewhat casual terminology and refer to them simply as subsets rather than by the technically more correct term "*sequence* of subsets". However, before considering the general partitioning of $\mathcal{Y}$ into such subsets we will treat two motivating, and relatively simple, special cases to gain intuition.

A. Special cases

1) Rank one matrices: First, let us consider the set of rank one matrices $\mathbf{Y} \in \mathcal{Y}$, i.e. the set given by

$$\mathcal{Y}_{\mathrm{R1}} \triangleq \mathcal{Y} \cap \{\mathbf{Y} \mid \mathrm{Rank}(\mathbf{Y}) = 1\}.$$

For any particular $\mathbf{Y}$ in this set, with an eigenvalue decomposition given by $\mathbf{Y} = \sigma\mathbf{u}\mathbf{u}^{\mathrm{T}}$ where $\|\mathbf{u}\| = 1$, we have

$$\mathrm{Tr}(\mathbf{QY}) = \sigma\mathbf{u}^{\mathrm{T}}\mathbf{Q}\mathbf{u}. \tag{21}$$

As $\sigma = 1$ due to the constraint $\mathrm{Tr}(\mathbf{Y}) = 1$ it follows that

$$\mathrm{P}\left(\mathrm{Tr}(\mathbf{QY}) \leq \epsilon\right) = \mathrm{P}\left(\|\mathbf{Hu}\|^2 \leq \epsilon\right) \doteq \epsilon^{\frac{n}{2}}$$

for this particular $\mathbf{Y} \in \mathcal{Y}_{\mathrm{R1}}$. It can also be shown that there are exactly $2^m - 1$ distinct $\mathbf{Y} \in \mathcal{Y}_{\mathrm{R1}}$. In essence, each such $\mathbf{Y}$ corresponds to the point at which line (in $\mathcal{X}$) connecting

$$\mathbf{X}_{\hat{\mathbf{s}}} \triangleq \begin{bmatrix} \hat{\mathbf{s}} \\ 1 \end{bmatrix} \begin{bmatrix} \hat{\mathbf{s}}^{\mathrm{T}} & 1 \end{bmatrix}$$

and $\mathbf{X}_{\mathbf{e}}$ intersects the hyperplane $\mathcal{H}$, given in (9). Therefore, by applying the union bound to the finite number of rank one $\mathbf{Y} \in \mathcal{Y}_{\mathrm{R1}}$ it follows that

$$\mathrm{P}\left(\tau_{\mathrm{R1}} \leq \epsilon\right) \doteq \epsilon^{\frac{n}{2}}$$

where

$$\tau_{\mathrm{R1}} = \inf_{\mathbf{Y}\in\mathcal{Y}_{\mathrm{R1}}} \mathrm{Tr}(\mathbf{QY}).$$

Note also that there is a one-to-one correspondence between the rank one matrices and all possible messages (not equal to the transmitted message), $\hat{\mathbf{s}} \in \mathcal{B}^m \backslash \mathbf{e}$, that are searched over by the ML detector. This is also the reason why

$$\mathrm{P}\left(\tau_{\mathrm{R1}} \leq \epsilon\right) \doteq \mathrm{P}\left(\hat{\mathbf{s}}_{\mathrm{ML}} = \mathbf{e}\right).$$

2) Full rank matrices: Next, consider the set of full rank (or more precisely *well conditioned*) $\mathbf{Y} \in \mathcal{Y}$ given by

$$\mathcal{Y}_{\mathrm{FR}} \triangleq \mathcal{Y} \cap \{\mathbf{Y} \mid \mathbf{Y} \succeq c\mathbf{I}\}$$

for some constant $c > 0$, and let

$$\tau_{\mathrm{FR}} \triangleq \inf_{\mathbf{Y}\in\mathcal{Y}_{\mathrm{FR}}} \mathrm{Tr}(\mathbf{QY}).$$

As the criterion function, $\mathrm{Tr}(\mathbf{QY})$, may be bounded as

$$\mathrm{Tr}(\mathbf{QY}) \geq c\mathrm{Tr}(\mathbf{Q}) = c\|\mathbf{H}\|^2$$

for any $\mathbf{Y} \in \mathcal{Y}_{\mathrm{FR}}$ it follows directly that

$$\mathrm{P}\left(\tau_{\mathrm{FR}} \leq \epsilon\right) \dot{\leq} \epsilon^{\frac{mn}{2}}$$

by applying property (37d) in Appendix I. This result can also be strengthened to show that

$$\mathrm{P}\left(\tau_{\mathrm{FR}} \leq \epsilon\right) \doteq \epsilon^{\frac{mn}{2}}.$$

3) Discussion: The implication of the result in Sections V-A.1 and V-A.2 is that the event that $\tau \leq \epsilon$ is (in the limit) much less likely to be caused by one of the matrices in $\mathcal{Y}_{\mathrm{FR}}$ than one of the matrices in $\mathcal{Y}_{\mathrm{R1}}$. The probability of the former is on the order of $\epsilon^{\frac{mn}{2}}$ while the later is only $\epsilon^{\frac{n}{2}}$ and $\epsilon^{\frac{mn}{2}} \ll \epsilon^{\frac{n}{2}}$ when ϵ is small (provided $m > 1$). Thus, (in a very loose sense) the reason for the high diversity of the SDR detector is that the elements added in the relaxation (the ones in $\mathcal{Y}_{\mathrm{FR}}$) are less likely to cause errors than the elements already present in the feasible set of the ML detection problem (the ones in $\mathcal{Y}_{\mathrm{R1}}$).

The question which however remains to be answered is if there is some other set of $\mathbf{Y}$, somewhere between the full rank and rank one matrices, which can cause $\tau \leq \epsilon$ to occur with a probability substantially larger than $\epsilon^{\frac{n}{2}}$. The answer to this question is somewhat surprisingly *no* provided that $n \geq m$ (but *yes* in some $n < m$ cases). In fact, most of the remaining part of the paper is concerned with the formal proof of this statement.

B. The General Case

In the general case we consider sets on the form given by

$$\mathcal{Y}(\mathbf{a},\mathbf{b}) \triangleq \mathcal{Y} \cap \{\mathbf{Y} \mid \epsilon^{a_k} \leq \sigma_k(\mathbf{Y}) \leq \epsilon^{b_k}\} \tag{22}$$

where $\mathbf{a} = (a_1, \ldots, a_m)$, $\mathbf{b} = (b_1, \ldots, b_m)$ and $\sigma_k(\mathbf{Y})$ denotes the kth eigenvalue of $\mathbf{Y}$. For notational convenience we will also in (22) interpret ϵ^{a_k} as 0 for $a_k = \infty$ in order to allow one or more eigenvalues to be identically equal to zero. We can without loss of generality assume that the eigenvalues are ordered and that $0 \leq a_1 \leq \ldots \leq a_m$, $0 = b_1 \leq \ldots \leq b_m$ and $b_k \leq a_k$ for $k = 1, \ldots, m$. Note that the assumption that $b_1 = 0$ can be made since (22) would, due to the $\mathrm{Tr}(\mathbf{Y}) = 1$ constraint of $\mathcal{Y}$ in (13), be empty otherwise. Similarly to before we define

$$\tau(\mathbf{a},\mathbf{b}) \triangleq \inf_{\mathbf{Y} \in \mathcal{Y}(\mathbf{a},\mathbf{b})} \mathrm{Tr}(\mathbf{Q}\mathbf{Y}). \tag{23}$$

In what follows, a bound on the probability of $\tau(\mathbf{a},\mathbf{b}) \leq \epsilon$ is obtained by first partitioning $\mathcal{Y}(\mathbf{a},\mathbf{b})$ into even smaller sets (essentially ϵ-balls) and then using the union bound to bound $\mathrm{P}\left(\tau(\mathbf{a},\mathbf{b}) \leq \epsilon\right)$. It will be more convenient to work with a square root factorization of $\mathbf{Y} \in \mathcal{Y}$ instead of with $\mathbf{Y}$ directly. Thus, we define a function,

$$\varphi : \mathbb{S}_+^m \mapsto \mathbb{R}^{m \times m} \tag{24}$$

(where $\mathbb{S}_+^m$ denotes the set of symmetric, positive semidefinite matrices) for which $\mathbf{A} = \varphi(\mathbf{Y})$ satisfies $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}^{\frac{1}{2}}$ and where $\mathbf{U}\mathbf{\Sigma}\mathbf{U}^{\mathrm{T}} = \mathbf{Y}$ is the eigenvalue decomposition of $\mathbf{Y}$. That is, φ provides square root factors of $\mathbf{Y}$ which have orthogonal columns with norms equal to $\sqrt{\sigma_i}$. Let $\mathcal{A}(\mathbf{a},\mathbf{b})$ be given by

$$\mathcal{A}(\mathbf{a},\mathbf{b}) \triangleq \varphi(\mathcal{Y}(\mathbf{a},\mathbf{b})), \tag{25}$$

i.e. $\mathcal{A}(\mathbf{a},\mathbf{b})$ is the set of square root factors which can be obtained from $\mathbf{Y} \in \mathcal{Y}(\mathbf{a},\mathbf{b})$. Note that $\mathrm{Tr}(\mathbf{Q}\mathbf{Y}) = \|\mathbf{H}\mathbf{A}\|^2$ since $\mathbf{Q} = \mathbf{H}^{\mathrm{T}}\mathbf{H}$ and $\mathbf{A} = \varphi(\mathbf{Y})$. The random variable $\tau(\mathbf{a},\mathbf{b})$, defined in (23), can thus be equivalently defined by

$$\tau(\mathbf{a},\mathbf{b}) = \inf_{\mathbf{A} \in \mathcal{A}(\mathbf{a},\mathbf{b})} \|\mathbf{H}\mathbf{A}\|^2. \tag{26}$$

We are now ready to provide the first lemma regarding the probability that $\|\mathbf{H}\tilde{\mathbf{A}}\|^2 \leq \epsilon$ for any $\tilde{\mathbf{A}}$ in an $\epsilon^{\frac{1}{2}}$-ball around a given center point $\mathbf{A} \in \mathcal{A}(\mathbf{a},\mathbf{b})$.

Lemma 3: Consider $\mathbf{A} \in \mathcal{A}(\mathbf{a},\mathbf{b})$ and define

$$\mathcal{A}_\epsilon(\mathbf{A}) \triangleq \{\tilde{\mathbf{A}} \mid \|\tilde{\mathbf{A}} - \mathbf{A}\| \leq \epsilon^{\frac{1}{2}}\}. \tag{27}$$

Further, let

$$\tau(\mathbf{A}) \triangleq \inf_{\tilde{\mathbf{A}} \in \mathcal{A}_\epsilon(\mathbf{A})} \|\mathbf{H}\tilde{\mathbf{A}}\|^2. \tag{28}$$

Then,

$$\mathrm{P}\left(\tau(\mathbf{A}) \leq \epsilon\right) \dot{\leq} \epsilon^{\nu} \qquad \text{where} \qquad \nu \triangleq \sum_{k=1}^{m} \frac{n(1-a_k)^+}{2}.$$

and where $(\cdot)^+ = \max(0, \cdot)$.

Proof: Note that, due to the rotational symmetry of the distribution of $\mathbf{H}$, it can without loss of generality be assumed that $\mathbf{A}$ is diagonal (and equal to $\mathbf{\Sigma}^{\frac{1}{2}}$ where $\mathbf{\Sigma}$ is a diagonal matrix containing the eigenvalues of $\mathbf{Y} \in \mathcal{Y}$ for which $\mathbf{A} = \varphi(\mathbf{Y})$).

Pick some $\delta > 0$ and consider the event that

$$\|\mathbf{H}\| \leq \epsilon^{-\delta} \tag{29}$$

and where at least one column of $\mathbf{H}$, $\mathbf{h}_k$, satisfies

$$\|\mathbf{h}_k\| \geq 2\epsilon^{\frac{1-a_k}{2}-\delta}. \tag{30}$$

We will first show that this event implies that $\tau(\mathbf{A}) > \epsilon$ and next that the event fails to occur with a probability which is no larger (in the $\dot{\leq}$ sense) than $\epsilon^{\nu - nm\delta}$. Hence

$$\begin{aligned}\mathrm{P}\left(\tau(\mathbf{A}) \leq \epsilon\right) &\leq \mathrm{P}\left(\|\mathbf{H}\| \geq \epsilon^{-\delta} \cup \|\mathbf{h}_k\| < 2\epsilon^{\frac{1-a_k}{2}-\delta}\ \forall k\right) \\ &\dot{\leq} \epsilon^{\nu - nm\delta}.\end{aligned}$$

Note first that (30) implies

$$\|\mathbf{h}_k \sigma_k^{\frac{1}{2}}\| \geq 2\epsilon^{\frac{1}{2}-\delta}$$

for at least one k since $\sigma_k \geq \epsilon^{a_k}$. Note also that this implies

$$\|\mathbf{H}\mathbf{A}\| = \|\mathbf{H}\mathbf{\Sigma}^{\frac{1}{2}}\| \geq 2\epsilon^{\frac{1}{2}-\delta}.$$

Now, consider $\|\mathbf{H}\tilde{\mathbf{A}}\|$ for any $\tilde{\mathbf{A}}$ satisfying $\|\tilde{\mathbf{A}} - \mathbf{A}\| \leq \epsilon^{\frac{1}{2}}$. Under the additional assumption of (29) it follows that

$$\begin{aligned}\|\mathbf{H}\tilde{\mathbf{A}}\| &= \|\mathbf{H}\mathbf{A} - \mathbf{H}(\mathbf{A} - \tilde{\mathbf{A}})\| \\ &\geq \|\mathbf{H}\mathbf{A}\| - \|\mathbf{H}(\mathbf{A} - \tilde{\mathbf{A}})\| \\ &\geq 2\epsilon^{\frac{1}{2}-\delta} - \epsilon^{\frac{1}{2}-\delta} \\ &= \epsilon^{\frac{1}{2}-\delta} > \epsilon^{\frac{1}{2}}\end{aligned}$$

where the last inequality holds whenever $\epsilon \leq 1$. Note also that $\|\mathbf{H}\tilde{\mathbf{A}}\| > \epsilon^{\frac{1}{2}}$ implies $\|\mathbf{H}\tilde{\mathbf{A}}\|^2 > \epsilon$. Therefore, (29) and (30) implies that $\tau(\mathbf{A}) > \epsilon$.

Now, consider the probability that (30) fails to hold, e.g. that

$$\|\mathbf{h}_k\| < 2\epsilon^{\frac{1-a_k}{2}-\delta}$$

for all $k = 1, \ldots, m$. As the columns of $\mathbf{H}$ are independent this probability can be upper bounded as

$$\begin{aligned}
&\mathrm{P}\left(\|\mathbf{h}_k\| < 2\epsilon^{\frac{1-a_k}{2}-\delta}\ \forall k\right) \\
&= \prod_{k=1}^{m} \mathrm{P}\left(\|\mathbf{h}_k\| < 2\epsilon^{\frac{1-a_i}{2}-\delta}\right) \\
&\dot{\leq} \prod_{k=1}^{m} \epsilon^{\frac{n(1-a_k-2\delta)^+}{2}} \leq \epsilon^{\nu-nm\delta}
\end{aligned}$$

where we have used

$$\mathrm{P}\left(\|\mathbf{h}\| \leq \epsilon^{\frac{c}{2}}\right) = \mathrm{P}\left(\|\mathbf{h}\|^2 \leq \epsilon^{c}\right) \dot{\leq} \epsilon^{\frac{nc^+}{2}}$$

according to (37d) in Appendix I with $\epsilon = \rho^{-1}$. The probability that (29) fails to hold can be upper bounded as

$$\mathrm{P}\left(\|\mathbf{H}\| > \epsilon^{-\delta}\right) \dot{\leq} \epsilon^{\infty}$$

according to (37e) in Appendix I. Therefore, by applying the union bound,

$$\begin{aligned}
\mathrm{P}\left(\tau(\mathbf{A}) \leq \epsilon\right) &\leq \mathrm{P}\left(\|\mathbf{H}\| \geq \epsilon^{-\delta} \cup \|\mathbf{h}_k\| < 2\epsilon^{\frac{1-a_k}{2}-\delta}\ \forall k\right) \\
&\dot{\leq} \epsilon^{\nu-nm\delta} + \epsilon^{\infty} \dot{\leq} \epsilon^{\nu-nm\delta}.
\end{aligned}$$

However, as $\delta > 0$ was arbitrary it follows that

$$\mathrm{P}\left(\tau(\mathbf{A}) \leq \epsilon\right) \dot{\leq} \epsilon^{\nu}$$

which concludes the proof. ■

The next lemma provides a bound on the number of $\epsilon^{\frac{1}{2}}$-balls (defined as in (27)) which are required to completely cover the set $\mathcal{A}(\mathbf{a}, \mathbf{b})$. Lemma 4 is the technically most difficult result of this work and we discuss this lemma below but save the the stringent proof for Appendix II.

Lemma 4: Let $\mathcal{A}(\mathbf{a}, \mathbf{b})$ and $\mathcal{A}_\epsilon(\mathbf{A})$ be defined as in (25) and (27), respectively. Then there is a collection of points, $\mathbf{A} = \{\mathbf{A}_i\}$, for which

$$\mathcal{A}(\mathbf{a}, \mathbf{b}) \subset \bigcup_{\mathbf{A}_i \in \mathbf{A}} \mathcal{A}_\epsilon(\mathbf{A}_i)$$

and

$$|\mathbf{A}| \dot{\leq} \epsilon^{-\mu}$$

where $|\mathbf{A}|$ denotes the number of elements of $\mathbf{A}$ and where

$$\mu \triangleq \sum_{k=2}^{m} \frac{(m-k+2)(1-b_k)^+}{2}. \tag{31}$$

Proof: Given in Appendix II. ■

Essentially, the proof of Lemma 4 relies on a geometric argument based on the dimensionality of low rank subsets of $\mathcal{A}$. Specifically, as part of the proof of Lemma 4 it is shown that the set of rank r matrices $\mathbf{A} \in \mathcal{A}$, i.e.

$$\mathcal{A}_{\mathrm{R}r} \triangleq \mathcal{A} \cap \{\mathbf{A} \mid \mathrm{Rank}(\mathbf{A}) = r\},$$

is part of a d_r-dimensional (smooth) manifold where

$$d_r \triangleq \sum_{k=2}^{r}(m-k+2), \quad r = 2, \ldots, m$$

and $d_1 \triangleq 0$. The manifold containing $\mathcal{A}_{\mathrm{R}r}$ is locally diffeomorphic (having a one-to-one differentiable relation) with the d_r-dimensional unit cube in $\mathbb{R}^{d_r}$ (this is a property of any smooth d_r-dimensional manifold [30] and not specific to $\mathcal{A}_{\mathrm{R}r}$). The volume, V, covered by one d_r-dimensional $\epsilon^{\frac{1}{2}}$-ball is on the order of

$$V \doteq (\epsilon^{\frac{1}{2}})^{d_r} = \epsilon^{\frac{d_r}{2}}$$

and therefore one needs on the order of

$$N \doteq \frac{1}{V} \doteq \epsilon^{\frac{-d_r}{2}} \tag{32}$$

such $\epsilon^{\frac{1}{2}}$-balls to cover the unit cube in $\mathbb{R}^{d_r}$. By exploiting that there is a differentiable (and therefore continuous) map between the unit cube and the manifold this result carries over to a covering of $\mathcal{A}_{\mathrm{R}r}$.

Thus, the set of rank r matrices, $\mathcal{A}_{\mathrm{R}r}$, can be covered by a collection of points, $\mathbf{A}_r$, satisfying

$$|\mathbf{A}_r| \dot{\leq} \epsilon^{-\mu_r}$$

where

$$\mu_r = \frac{d_r}{2} = \sum_{k=2}^{r} \frac{(m-k+2)}{2}.$$

Extending this line of reasoning from rank r dimensional subsets, $\mathcal{A}_{\mathrm{R}r}$, to subsets which are close to being low rank in the sense that the singular values of $\mathbf{A}$ are bounded by powers of ϵ yields the result stated in Lemma 4. Note also that this is similar to the discussion following Theorem 4 in [27].

Now, Lemma 3 and Lemma 4 can be combined in order to bound the probability that $\mathcal{A}(\mathbf{a},\mathbf{b})$ contains an $\mathbf{A}$ for which $\|\mathbf{HA}\|^2 \leq \epsilon$. Then, by optimizing over $\mathbf{a}$ and $\mathbf{b}$, one can find the set of the form of $\mathcal{A}(\mathbf{a},\mathbf{b})$ most likely to contain such an $\mathbf{A}$. It can also be argued that this set will dominate the probability of error in the high SNR regime. These ideas are captured by the following lemma.

Lemma 5: Let τ be defined as in (11). Then

$$\mathrm{P}\left(\tau \leq \epsilon\right) \dot{\leq} \epsilon^{\zeta}$$

where

$$\zeta \triangleq \inf_{1 \geq c_2 \geq \ldots \geq c_m \geq 0} \frac{n}{2} + \sum_{k=2}^{m} \frac{(n-m+k-2)c_k}{2}. \tag{33}$$

Proof: Consider picking some $\mathbf{b} = (b_1, \ldots, b_m)$ for which $b_1 = 0$ and $b_1 \leq b_2 \leq \ldots \leq b_m \leq 1$ and choose a $\delta > 0$. Let $\mathbf{a} = (a_1, \ldots, a_m)$ be given such that $a_1 = \delta$ and $a_k = b_k + \delta$ if $b_k + \delta \leq 1$ or $a_k = \infty$ otherwise for $k = 2, \ldots, m$.

The probability that $\tau(\mathbf{a},\mathbf{b}) \leq \epsilon$ where $\tau(\mathbf{a},\mathbf{b})$ is defined in (23) can be bounded, using the union bound according as

$$\mathrm{P}\left(\tau(\mathbf{a},\mathbf{b}) \leq \epsilon\right) \leq \sum_{\mathbf{A}_i \in \mathbf{A}} \mathrm{P}\left(\tau(\mathbf{A}_i) \leq \epsilon\right)$$

where $\mathbf{A}$ is chosen according to Lemma 4 and where $\tau(\mathbf{A}_i)$ is given by (28). Each term in the sum is upper bounded by

$$\mathrm{P}\left(\tau(\mathbf{A}_i) \leq \epsilon\right) \dot{\leq} \epsilon^{\nu}$$

where ν is given in Lemma 3. The number of terms in the sum is upper bounded by

$$|\mathbf{A}| \dot{\leq} \epsilon^{-\mu}$$

where μ is given by (31). Thus, the probability that $\tau(\mathbf{a},\mathbf{b}) \leq \epsilon$ is bounded as

$$\mathrm{P}\left(\tau(\mathbf{a},\mathbf{b}) \leq \epsilon\right) \dot{\leq} \epsilon^{\nu-\mu}$$

where

$$\begin{aligned}
\nu - \mu =& \sum_{k=1}^{m} \frac{n(1-a_k)^+}{2} - \sum_{k=2}^{m} \frac{(m-k+2)(1-b_k)^+}{2} \\
\geq& \frac{n}{2} + \sum_{k=2}^{m} \frac{(n-m+k-2)(1-b_k)^+}{2} - \frac{mn\delta}{2} \\
\geq& \zeta - \frac{mn\delta}{2}
\end{aligned}$$

and where the property

$$(1-a_k)^+ \geq (1-b_k)^+ - \delta$$

(for a_k chosen as above) was used to establish the first inequality. The second inequality follows by the definition of ζ in (33) along with $b_k \geq 0$.

Now, let

$$\mathcal{A} \triangleq \varphi(\mathcal{Y})$$

where φ is given by (24). Note that we can pick a finite set of $\mathbf{b} \in [0,1]^m$, $\mathbf{B} = \{\mathbf{b}_i\}$, such that

$$\mathcal{A} \subset \bigcup_{\mathbf{b}\in\mathbf{B}} \mathcal{A}(\mathbf{a},\mathbf{b}) \tag{34}$$

where $\mathbf{a} = \mathbf{a}(\mathbf{b})$ according to the above. This follows since by specifying $\mathbf{b} = (b_1,\ldots,b_m)$ we include the matrices $\mathbf{Y} \in \mathcal{Y}$ for which the kth eigenvalue satisfies $\epsilon^{b_k+\delta} \leq \sigma_k \leq \epsilon^{b_k}$ if $b_k < 1$ and $\sigma_k \leq \epsilon$ if $b_k = 1$. Thus we can cover the entire range of $\sigma_k \in [0,1]$ with a finite number of $b_k \in [0,1]$. For the special case of $k = 1$ we know that σ_1 is bounded away from 0 due to $\mathrm{Tr}(\mathbf{Y}) = 1$ which implies that $\sigma_1 \in [\epsilon^\delta, 1]$ for sufficiently small ϵ given that $\delta > 0$ which is why $b_1 = 0$ can be assumed without loss of generality.

Using the union bound it follows that

$$\begin{aligned}\mathrm{P}\left(\tau \leq \epsilon\right) &\leq \sum_{\mathbf{b}\in\mathbf{B}} \mathrm{P}\left(\tau(\mathbf{a},\mathbf{b}) \leq \epsilon\right)\\ &\dot{\leq} \epsilon^{\zeta - \frac{mn\delta}{2}}\end{aligned}$$

since each term in the sum satisfies

$$\mathrm{P}\left(\tau(\mathbf{a},\mathbf{b}) \leq \epsilon\right) \dot{\leq} \epsilon^{\zeta - \frac{mn\delta}{2}}$$

and the number of terms is finite. However, as $\delta > 0$ was arbitrary it follows that

$$\mathrm{P}\left(\tau(\mathbf{a},\mathbf{b}) \leq \epsilon\right) \dot{\leq} \epsilon^{\zeta}$$

which concludes the proof. ■

In light of Lemma 5 the proof of Theorem 1 is now almost trivial. All that remains is to compute ζ in (33) and apply Lemma 2. We give the proof below.

Proof (of Theorem 1): For the case where $n \geq m$ all terms in the sum appearing in (33) are non negative. Thus, the minimum in (33) is achieved for $c_2 = \ldots = c_m = 0$ and it follows that

$$\zeta = \frac{n}{2}.$$

This, combined with Lemma 2, proves that

$$\mathrm{P}\left(\hat{\mathbf{s}}_{\mathrm{SDR}} \neq \mathbf{e}\right) \dot{\leq} \rho^{-\frac{n}{2}}.$$

Next, note that the error probability of the SDR receiver is lower bounded by

$$\mathrm{P}\left(\hat{\mathbf{s}}_{\mathrm{SDR}} \neq \mathbf{e}\right) \geq \mathrm{P}\left(\hat{\mathbf{s}}_{\mathrm{ML}} \neq \mathbf{e}\right) \doteq \rho^{-\frac{n}{2}}$$

since the ML detector achieves the minimum probability of error. It therefore follows that

$$\mathrm{P}\left(\hat{\mathbf{s}}_{\mathrm{SDR}} \neq \mathbf{e}\right) \doteq \mathrm{P}\left(\hat{\mathbf{s}}_{\mathrm{ML}} \neq \mathbf{e}\right) \doteq \rho^{-\frac{n}{2}}.$$

By noting again that $\mathbf{s} = \mathbf{e}$ can be assumed without loss of generality the statement of Theorem 1 follows. ■

VI. Extensions

At this stage, only the case of real valued systems on the form of (1) have been considered. Also, for the proof of Theorem 1 it was assumed that $n \geq m$. In this section, we discuss the extensions which would follow by relaxing these constraints and some illustrative numerical examples are given.

A. The $n < m$ case

As stated above, full diversity has so far been shown under the condition that $n \geq m$. However, a careful inspection of the proofs show that the only part which explicitly relies on this assumption is when it is argued that $c_2 = \ldots = c_m = 0$ is an optimal point for (33) in the $n \geq m$ case. However, nontrivial bounds on the diversity will follow whenever ζ in (33) is strictly positive. The following theorem provides a lower bound on the diversity for the case when $n < m$.

Theorem 2: Given the assumptions of Theorem 1 but for $r \triangleq m - n > 0$, it holds that

$$\lim_{\rho \to \infty} \frac{\ln \mathrm{P}\left(\hat{\mathbf{s}}_{\mathrm{SDR}} \neq \mathbf{s}\right)}{\ln \rho} \leq -d$$

where

$$d = \frac{1}{2}\left(m - \frac{r(r+3)}{2}\right) \tag{35}$$

Proof: All that needs to be done in this case is to find the optimum in (33) and apply Lemma 2. To this end, note that the optimum of (33) is achieved for $c_k = 1$ for all k satisfying

$$n - m + k - 2 < 0 \Leftrightarrow k \leq m - n + 1$$

and $c_k = 0$ for k satisfying

$$n - m + k - 2 \geq 0 \Leftrightarrow k \geq m - n + 2.$$

The value of ζ in (33) is thus given as

$$\zeta = \frac{n}{2} + \sum_{k=2}^{m-n+1} \frac{n-m+k-2}{2} = \frac{1}{2}\left(m - \frac{r(r+3)}{2}\right)$$

This completes the proof. ■

Note that this result is only nontrivial if

$$m > \frac{r(r+3)}{2}$$

as otherwise Theorem 2 would simply state that the probability of error is less than one. Further, we have no specific reason to believe that the bound is tight (in the sense that $\dot{\leq}$ could be replaced by $\doteq$) in the $n < m$ case, even in the cases where the bound is non-trivial. An indication of this is given in Fig. 3 where the diversity of the SDR detector seems to be larger than 2 which is predicted by (35). It is however also unreasonable to expect the bound to be very loose in the sense that the SDR detector would maintain the same diversity as the ML detector in the general case where $n < m$. This is indicated by Fig. 4 where the error probability of the SDR is significantly larger than that of the ML detector. Intuitively, in the $n < m$ case, it can become likely that a matrix with higher rank than one achieves the minimum in (12). Therefore, the typical error events of the SDR detector no longer coincide with the error events of the ML detector and the SDR detector can experience a loss in diversity. We do not however, as pointed out above, expect the loss to be as large as what is indicated by (35).

A possible way to strengthen the analysis in the $n < m$ case can actually be seen by turning back to Fig. 2. Essentially, as part of proving Theorem 1 (and Theorem 2) the intersection of $\mathcal{X}$ and $\mathcal{H}$ is covered with ϵ-balls. However, due to the linearity of the objective function it is already known that the minimum objective value over the intersection set must be achieved by one of the boundary points of $\mathcal{X}$. Therefore, it would suffice to cover the intersection of $\mathcal{H}$ with the *boundary* of $\mathcal{X}$. This would in turn strengthen the bound on $|\mathbf{A}|$ in Lemma 4 but would also require a framework for parameterizing the boundary set. It may also be possible to use the structure of the problem in other ways. One such way could be to make use of the results in [29] (where bounds on the rank of extremal matrices for semidefinite programs are provided) to further limit the part of the feasible set that needs to be covered.

B. Complex channel matrices

It is well known that the SDR receiver is also applicable to the case where 4-QAM symbols are transmitted over a complex valued MIMO channel, see e.g. [5]. The most direct strategy is to rewrite the problem in an equivalent real valued form according to

$$\begin{bmatrix} \Re(\mathbf{y}_c) \\ \Im(\mathbf{y}_c) \end{bmatrix} = \begin{bmatrix} \Re(\mathbf{H}_c) & -\Im(\mathbf{H}_c) \\ \Im(\mathbf{H}_c) & \Re(\mathbf{H}_c) \end{bmatrix} \begin{bmatrix} \Re(\mathbf{s}_c) \\ \Im(\mathbf{s}_c) \end{bmatrix} + \begin{bmatrix} \Re(\mathbf{v}_c) \\ \Im(\mathbf{v}_c) \end{bmatrix} \tag{36}$$

where $\mathbf{y}_c \in \mathbb{C}^N$, $\mathbf{H}_c \in \mathbb{C}^{N\times M}$, $\mathbf{s}_c \in \mathbb{C}^M$ and $\mathbf{v}_c \in \mathbb{C}^N$ are the (to (1)) corresponding complex valued quantities and where $\Re(\cdot)$ and $\Im(\cdot)$ denote the real and imaginary parts.

However, the proof of Theorem 1 does unfortunately not extend to cover this case. The specific reason is found in Lemma 3 where the rotational symmetry of $\mathbf{H}$ is explicitly used. This symmetry is lost in the formulation given in (36), even in the case where $\mathbf{H}_c$ is i.i.d. compex, circularly symmetric, zero mean Gaussian. More importantly, numerical simulations suggest that the extension of Theorem 1 to this case may not even be true. An indication of this can be seen in Fig. 5 where it is plausible to believe that the SDR receiver does experience a loss of diversity. However, it should also be pointed out that we do not expect the loss (if any) to be very large in general. This belief is based on extensive simulations, such as the one shown in Fig. 6, that indicates a high SDR diversity in the complex case.

At first sight, what would be required in order to cover the complex case would be to update Lemma 3 for the structure of the effective channel matrix, $\mathbf{H}$, in (36). It is however also likely that Lemma 4 would need to be strengthened (as discussed in Section VI-A) in order to obtain a tight bound on the diversity. However, these steps remain a challenge. Also, note that if the SDR detector does not achieve full diversity, the issue of providing a lower bound on the error probability (or equivalently an upper bound on diversity) will also become more challenging.

VII. CONCLUSIONS

In this paper we have shown that when applied to a fading channel, modelled by a real valued matrix with i.i.d. Gaussian entries of zero mean and finite variance, the semidefinite relaxation detector achieves the maximum possible diversity. This provides a strong performance guarantee for the SDR approach, when applied in the communications context. Based on the discussions in Section VI it does not seem reasonable to expect such a strong statement to hold for an arbitrary system. Nonetheless, it is still reasonable to assume that the SDR detector will be superior to the class of linear detector and other relaxation techniques.

APPENDIX I
EXPONENTIAL EQUALITY

For the readers convenience, we list the (for this work) most important properties associated with the definition of *exponential equality* in (7). These properties are easily derived from the definition in (7) and can also be found (often implicitly) in many texts, see e.g. [27], [2]. Thus, we state the properties without proof.

1) *Scaling property:* For any $a \in [-\infty, \infty]$ and $c \in (-\infty, \infty)$ it holds that

$$f(\rho) \doteq \rho^{-a} \Rightarrow cf(\rho) \doteq \rho^{-a}. \tag{37a}$$

2) *Summation property:* For any $a, b \in [-\infty, \infty]$ it holds that

$$f(\rho) \doteq \rho^{-a},\ g(\rho) \doteq \rho^{-b} \Rightarrow f(\rho) + g(\rho) \doteq \rho^{-\min(a,b)} \tag{37b}$$

This property extends in the obvious way to the sum of finitely many terms.

3) *Multiplication property:* For any $a, b \in [-\infty, \infty]$ it holds that

$$f(\rho) \doteq \rho^{-a},\ g(\rho) \doteq \rho^{-b} \Rightarrow f(\rho)g(\rho) \doteq \rho^{-(a+b)} \tag{37c}$$

if the cases where $a + b$ is not well defined are excluded.

4) *Extremal realizations of Gaussian vectors:* Let $\mathbf{h} \in \mathbb{R}^d$ be a vector of i.i.d. Gaussian elements of finite non-zero variance. Then

$$\mathrm{P}\left(\|\mathbf{h}\|^2 \leq \rho^{-c}\right) \doteq \rho^{-\frac{dc^+}{2}} \tag{37d}$$

for $c \in (-\infty, \infty)$, where $c^+ \triangleq \max(c, 0)$ and

$$\mathrm{P}\left(\|\mathbf{h}\|^2 \geq \rho^{c}\right) \doteq \rho^{-\infty} \tag{37e}$$

for $c > 0$. These properties follow by noting that $\|\mathbf{h}\|^2$ is χ^2 distributed with d degrees of freedom, see e.g. [2, Section 5.4.2].

It should also be noted that the properties given in (37a), (37b) and (37c) also hold with $\dot{\leq}$ or $\dot{\geq}$ in place of $\doteq$.

APPENDIX II
PROOF OF LEMMA 4

Before proving Lemma 4 we establish the following technical result regarding the feasible set of (12).

Lemma 6: The set $\mathcal{Y}$ defined in (13) satisfies

$$\mathcal{Y} = \{\mathbf{Y} \in \mathbb{S}^m \mid \mathrm{Tr}(\mathbf{Y}) = 1, \mathbf{Y} \succeq \tfrac{1}{4}\mathbf{d}\mathbf{d}^{\mathrm{T}}, \mathbf{d} = \mathrm{diag}(\mathbf{Y})\}. \tag{38}$$

Proof: Consider the transformation given by

$$\underbrace{\begin{bmatrix} \mathbf{Y} & \mathbf{a} \\ \mathbf{a}^{\mathrm{T}} & c \end{bmatrix}}_{\mathbf{P}} = \underbrace{\begin{bmatrix} \mathbf{I} & -\mathbf{e} \\ \mathbf{0}^{\mathrm{T}} & 1 \end{bmatrix}}_{\mathbf{T}} \mathbf{X} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{e}^{\mathrm{T}} & 1 \end{bmatrix} \tag{39}$$

or inversely,

$$\mathbf{X} = \underbrace{\begin{bmatrix} \mathbf{I} & \mathbf{e} \\ \mathbf{0}^{\mathrm{T}} & 1 \end{bmatrix}}_{\mathbf{R}} \begin{bmatrix} \mathbf{Y} & \mathbf{a} \\ \mathbf{a}^{\mathrm{T}} & c \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{e}^{\mathrm{T}} & 1 \end{bmatrix} \tag{40}$$

since $\mathbf{T}^{-1} = \mathbf{R}$. Note also that $\mathbf{Y}$ is given by $\mathbf{Y} = \mathbf{M}\mathbf{X}\mathbf{M}^{\mathrm{T}}$ as $\mathbf{M} = \begin{bmatrix} \mathbf{I} & -\mathbf{e} \end{bmatrix}$ by (10). Expanding $\mathbf{X}$ from (40) yields

$$\mathbf{X} = \begin{bmatrix} \mathbf{Y} + \mathbf{a}\mathbf{e}^{\mathrm{T}} + \mathbf{e}\mathbf{a}^{\mathrm{T}} + \mathbf{e}c\mathbf{e}^{\mathrm{T}} & \mathbf{a} + \mathbf{e}c \\ \mathbf{a}^{\mathrm{T}} + c\mathbf{e}^{\mathrm{T}} & c \end{bmatrix}.$$

Thus, the constraint $\mathrm{diag}(\mathbf{X}) = \mathbf{e}$ for $\mathbf{X} \in \mathcal{X}$ implies that $c = 1$ for $\mathbf{Y} \in \mathcal{Y}$ since $\mathcal{Y} \subset \tilde{\mathcal{Y}} = \mathbf{M}\mathcal{X}\mathbf{M}^{\mathrm{T}}$ for $\mathcal{Y}$ given in (13) and where $\tilde{\mathcal{Y}}$ is given in (14). Further, for $c = 1$

$$\mathrm{diag}(\mathbf{Y} + \mathbf{a}\mathbf{e}^{\mathrm{T}} + \mathbf{e}\mathbf{a}^{\mathrm{T}} + \mathbf{e}\mathbf{e}^{\mathrm{T}}) = \mathrm{diag}(\mathbf{Y}) + 2\mathbf{a} + \mathbf{e} = \mathbf{e}$$

which implies that

$$\mathbf{a} = -\tfrac{1}{2}\mathrm{diag}(\mathbf{Y}). \tag{41}$$

Thus, given a matrix $\mathbf{Y} \in \tilde{\mathcal{Y}}$ there is actually a unique $\mathbf{X} \in \mathcal{X}$ for which $\mathbf{Y} = \mathbf{M}\mathbf{X}\mathbf{M}^{\mathrm{T}}$. In other words, the mapping from $\mathcal{X}$ to $\tilde{\mathcal{Y}}$ is one-to-one.

Since $\mathbf{T}$ (and $\mathbf{R}$) are invertible the constraint $\mathbf{X} \succeq \mathbf{0}$ is equivalent to $\mathbf{P} \succeq \mathbf{0}$. However, $\mathbf{P} \succeq \mathbf{0}$ if and only if its Schur complement [20] is positive semidefinite, i.e. if

$$\mathbf{Y} - c^{-1}\mathbf{a}\mathbf{a}^{\mathrm{T}} \succeq \mathbf{0}.$$

Thus, by combining (41) with $c = 1$ and identifying $\mathbf{d} = -2\mathbf{a}$ the equalities of (13) and (38) are established. ■

We are now in a position to prove the statement given by Lemma 4. For convenience the lemma is restated below.

Lemma 4: Let $\mathcal{A}(\mathbf{a},\mathbf{b})$ and $\mathcal{A}_\epsilon(\mathbf{A})$ be defined as in (25) and (27) respectively. Then there is a collection of points, $\mathbf{A}=\{\mathbf{A}_i\}$, for which

$$\mathcal{A}(\mathbf{a},\mathbf{b}) \subset \bigcup_{\mathbf{A}_i\in\mathbf{A}} \mathcal{A}_\epsilon(\mathbf{A}_i)$$

and

$$|\mathbf{A}| \dot{\leq} \epsilon^{-\mu}$$

where

$$\mu \triangleq \sum_{k=2}^{m} \frac{(m-k+2)(1-b_k)^+}{2}.$$

Proof: Consider the triplet $(\mathbf{U},\boldsymbol{\lambda},\mathbf{z}) \in \mathbb{R}^{m\times m}\times\mathbb{R}^m\times\mathbb{R}^m$ and the system of equations given by

$$\mathrm{Tr}(\boldsymbol{\Lambda}^2) = 1 \tag{42a}$$

$$\mathrm{diag}(\mathbf{U}\boldsymbol{\Lambda}^2\mathbf{U}^{\mathrm{T}}) = \mathbf{U}\mathbf{z} \tag{42b}$$

$$\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I} \tag{42c}$$

$$\boldsymbol{\Lambda}^2 - \tfrac{1}{4}\mathbf{z}\mathbf{z}^{\mathrm{T}} \succeq \mathbf{0} \tag{42d}$$

where $\boldsymbol{\Lambda} \triangleq \mathrm{diag}(\boldsymbol{\lambda})$. The set of solutions to (42) will in what follows be denoted by $\mathcal{M}$. The set of solutions to (42a), (42b) and (42c) but not necessarily (42d) is denoted by $\mathcal{N}$ and it follows that $\mathcal{M}\subset\mathcal{N}$. From (42a) and (42c) it follows that $\boldsymbol{\lambda}$ and $\mathbf{U}$ in the solution set are bounded. However, as $\mathbf{U}$ is full rank due to (42c) it follows through (42b) that $\mathbf{z}$ is also bounded. Therefore, both $\mathcal{N}$ and $\mathcal{M}$ are compact (closed and bounded) sets.

The constraints of (42) are such that any solution, $(\mathbf{U},\boldsymbol{\lambda},\mathbf{z})$, of (42) satisfies $\mathbf{U}\boldsymbol{\Lambda}^2\mathbf{U}^{\mathrm{T}}\in\mathcal{Y}$ and any eigenvalue decomposition, $\mathbf{Y}=\mathbf{U}\boldsymbol{\Sigma}\mathbf{U}^{\mathrm{T}}$, of $\mathbf{Y}\in\mathcal{Y}$ solves (42) for $\boldsymbol{\Lambda}=\boldsymbol{\Sigma}^{\frac{1}{2}}$ and some (unique) $\mathbf{z}$. To see this, consider the eigenvalue decomposition, $\mathbf{Y}=\mathbf{U}\boldsymbol{\Sigma}\mathbf{U}^{\mathrm{T}}$, of some $\mathbf{Y}\in\mathcal{Y}$ where $\mathcal{Y}$ is given by (13). Note also that $\mathbf{Y}$ belongs to $\mathcal{Y}$ if and only if it satisfies the constraints of (38) as proven in Lemma 6. The orthogonality of $\mathbf{U}\in\mathbb{R}^{m\times m}$ is a property of the eigenvalue decomposition and therefore (42c) is satisfied. For $\boldsymbol{\Lambda}=\boldsymbol{\Sigma}^{\frac{1}{2}}$ and $\mathbf{z}=\mathbf{U}^{\mathrm{T}}\mathrm{diag}(\mathbf{U}\boldsymbol{\Lambda}^2\mathbf{U}^{\mathrm{T}})$ the constraint of (42b) is satisfied. As $\mathbf{Y}\in\mathcal{Y}$ it follows that $\mathbf{Y}-\frac{1}{4}\mathbf{d}\mathbf{d}^{\mathrm{T}}\succeq\mathbf{0}$ where $\mathbf{d}=\mathrm{diag}(\mathbf{Y})$. Therefore, $\mathrm{diag}(\mathbf{Y})=\mathrm{diag}(\mathbf{U}\boldsymbol{\Lambda}^2\mathbf{U}^{\mathrm{T}})=\mathbf{U}\mathbf{z}$ implies

$$\mathbf{U}\boldsymbol{\Lambda}^2\mathbf{U}^{\mathrm{T}} - \tfrac{1}{4}\mathbf{U}\mathbf{z}\mathbf{z}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}} \succeq \mathbf{0} \Leftrightarrow \boldsymbol{\Lambda}^2 - \tfrac{1}{4}\mathbf{z}\mathbf{z}^{\mathrm{T}} \succeq \mathbf{0}$$

which means that (42d) is satisfied. Finally, the constraint $\mathrm{Tr}(\mathbf{Y})=1$ in (38) implies $\mathrm{Tr}(\boldsymbol{\Lambda}^2)=1$ and (42a) is satisfied. Reversing the reasoning and applying Lemma 6 show that any solution to (42) must also have the property that $\mathbf{U}\boldsymbol{\Lambda}^2\mathbf{U}^{\mathrm{T}}\in\mathcal{Y}$.

The value of introducing (42) is that it will, through the implicit function theorem [31], provide a means of parameterizing the eigenvalues and vectors of $\mathbf{Y} \in \mathcal{Y}$. To this end, let

$$p \triangleq m + \frac{m(m+1)}{2} + 1,$$
$$q \triangleq m^2 + 2m,$$

and $\boldsymbol{\omega} \in \mathbb{R}^q$ be given by

$$\boldsymbol{\omega} \triangleq (\mathbf{U}, \boldsymbol{\lambda}, \mathbf{z}).$$

Define

$$H : \mathbb{R}^q \mapsto \mathbb{R}^p$$

according to

$$H(\boldsymbol{\omega}) \triangleq \begin{bmatrix} \mathrm{Tr}(\boldsymbol{\Lambda}^2) - 1 \\ \mathrm{diag}(\mathbf{U}\boldsymbol{\Lambda}^2\mathbf{U}^{\mathrm{T}}) - \mathbf{U}\mathbf{z} \\ \mathrm{svec}(\mathbf{U}^{\mathrm{T}}\mathbf{U} - \mathbf{I}) \end{bmatrix}$$

and note that $H(\boldsymbol{\omega}) = \mathbf{0}$ corresponds to (42a), (42b) and (42c). In the above, $\mathrm{svec}(\cdot)$ referrers to the vector obtained by stacking the upper triangular part of a symmetric matrix into a vector. Let

$$\bar{\boldsymbol{\omega}} \triangleq (\bar{\mathbf{U}}, \bar{\boldsymbol{\lambda}}, \bar{\mathbf{z}})$$

be a solution of (42) and $\mathcal{I}$ be an index set satisfying

$$\mathcal{I} \subset \{1, \ldots, q\} \tag{43}$$

and

$$|\mathcal{I}| = p. \tag{44}$$

Denote by $\boldsymbol{\omega}_{\mathcal{I}} \in \mathbb{R}^p$ the vector of components in $\boldsymbol{\omega}$ indexed by $\mathcal{I}$ and let $\boldsymbol{\omega}_{\mathcal{I}^c} \in \mathbb{R}^{q-p}$ be the vector consisting of the remaining components. The implicit function theorem [31] states that if

$$\left| \frac{\partial H(\boldsymbol{\omega})}{\partial \boldsymbol{\omega}_{\mathcal{I}}} \right|_{\boldsymbol{\omega}=\bar{\boldsymbol{\omega}}} \neq 0, \tag{45}$$

then there is a neighborhood, $\mathcal{U} \subset \mathbb{R}^q$, containing $\bar{\boldsymbol{\omega}}$ and a differentiable mapping

$$g : \mathbb{R}^{q-p} \mapsto \mathbb{R}^p$$

satisfying $\boldsymbol{\omega}_{\mathcal{I}} = g(\boldsymbol{\omega}_{\mathcal{I}^c})$ for any $\boldsymbol{\omega} \in \mathcal{U} \cap H^{-1}(\{\mathbf{0}\})$.

Further (45) implies the existence of a differentiable mapping

$$\psi : \mathcal{D} \mapsto \mathcal{R}$$

for which $\boldsymbol{\omega} = \psi(\boldsymbol{\xi})$, where $\boldsymbol{\xi} \triangleq \boldsymbol{\omega}_{\mathcal{I}^c} - \bar{\boldsymbol{\omega}}_{\mathcal{I}^c} \in \mathbb{R}^{q-p}$, where $\mathcal{D}$ is an open subset of $\mathbb{R}^{q-p}$ containing $\mathbf{0}$ and where $\mathcal{R} \triangleq \psi(\mathcal{D}) \subset \mathbb{R}^q$. This mapping is easily obtained from g by including the components in $\boldsymbol{\omega}_{\mathcal{I}^c}$ and performing a translation to a neighborhood of $\mathbf{0}$. Thus, assuming that (45) is satisfied, the solution set of (42) is locally parameterized by $q-p$ scalar parameters. It will in fact later be shown that given *any* solution, $\bar{\boldsymbol{\omega}}$, to (42) there will be some index set, $\mathcal{I}$, satisfying (43) and (44) for which (45) is satisfied. This implies that $\mathcal{N}$ is a $q-p$ dimensional (smooth) manifold embedded in $\mathbb{R}^q$ [32]. Note however that the specific index set, $\mathcal{I}$, required to satisfy (45) will generally depend on the particular $\bar{\boldsymbol{\omega}}$ chosen. This is analogous to the problem of parameterizing the unit circle based on solving $x^2+y^2=1$ where the choice of x or y as the *free* parameter depends on if the parametrization neighborhood should include $x=0$ or $y=0$.

Note that it can without loss of generality be assumed that the domain of ψ, is given by

$$\mathcal{D} = (-\kappa, \kappa)^{q-p}, \tag{46}$$

i.e. that $\mathcal{D}$ is an open hypercube for some $\kappa > 0$ [32]. Further, since $\mathcal{N}$ is compact it can be assumed that κ is independent of $\bar{\boldsymbol{\omega}}$. It can also, without loss of generality, be assumed that ψ is Lipschitz continuous [33] on $\mathcal{D}$. This follows since the inverse function theorem guarantees that ψ has continuous derivatives on the closure of $\mathcal{D}$, $\bar{\mathcal{D}}$ (actually, in its standard form the inverse function theorem guarantees continuous derivatives on $\mathcal{D}$ but by reducing κ if necessary the continuity can be extended to the closure of $\mathcal{D}$). Further, again due to the compactness of $\mathcal{N}$, it can be assumed that the Lipschitz constant of ψ is independent of $\bar{\boldsymbol{\omega}}$.

In order to prove the *existence* of an index set, $\mathcal{I}$, for which (45) is satisfied it is sufficient to prove that the Jacobian matrix $\mathbf{D}$,

$$\mathbf{D} \triangleq \left.\frac{\partial H(\boldsymbol{\omega})}{\partial \boldsymbol{\omega}}\right|_{\boldsymbol{\omega}=\bar{\boldsymbol{\omega}}} \in \mathbb{R}^{p\times q}, \tag{47}$$

is full rank. In this event, the index set, $\mathcal{I}$, can be taken as the indexes of any p linearly independent columns of $\mathbf{D}$. For our purposes however, we shall need to be a bit more specific about how $\mathcal{I}$ is chosen. Therefore, note again that it will be of particular interest to study parameterizations of $\mathcal{M}$ (and $\mathcal{N}$) around solutions $\bar{\boldsymbol{\omega}}$ corresponding to rank deficient $\mathbf{Y} \in \mathcal{Y}$ (see the discussion in Section V-A.3). To this end, consider some $\bar{\boldsymbol{\omega}} \in \mathcal{M}$ for which $\lambda_{r+1} = \ldots = \lambda_m = 0$, i.e. $\bar{\boldsymbol{\omega}}$ corresponds to a rank r matrix $\bar{\mathbf{Y}} \in \mathcal{Y}$. Here, and in what follows, λ_k and z_k refer to the kth component of $\boldsymbol{\lambda}$ and $\mathbf{z}$ respectively. For any $\bar{\boldsymbol{\omega}} \in \mathcal{M}$ it follows by (42d) that $|z_k| \leq 2|\lambda_k|$ for $k = 1, \ldots, m$ and in particular it follows that $z_k = 0$ whenever $\lambda_k = 0$. We will in what follows refer to any $\bar{\boldsymbol{\omega}} \in \mathcal{N}$ which satisfies both $\lambda_{r+1} = \ldots = \lambda_m = 0$ and $z_{r+1} = \ldots = z_m = 0$ as a rank r point, even in the case that $\bar{\boldsymbol{\omega}} \neq \mathcal{M}$. The reason for using

this terminology is that it is often difficult to verify that (42d) is satisfied but sufficient to provide a parametrization around rank r points, $\bar{\boldsymbol{\omega}} \in \mathcal{N}$.

Let

$$p_r \triangleq m + \frac{r(r+1)}{2} + 1$$

and

$$q_r \triangleq r(m+2)$$

and note that $p = p_m$ and $q = q_m$. Further, let $\mathbf{u}_k$ denote the kth column of $\mathbf{U}$. It will in what follows be shown that $\boldsymbol{\omega}$, in a neighborhood of a rank r point, $\bar{\boldsymbol{\omega}}$, can be parameterized by specifying λ_k and z_k for $k = r+1, \ldots, m$, a subset of $m-k$ parameters from $\mathbf{u}_k$ for $k = r+1, \ldots, m$, and a subset of $q_r - p_r$ parameters from

$$\boldsymbol{\omega}_r \triangleq (\mathbf{u}_1, \ldots, \mathbf{u}_r, \lambda_1, \ldots, \lambda_r, z_1, \ldots, z_r).$$

It is straightforward to verify that this amounts to a total of $q-p$ parameters. The specific parameters chosen from $\mathbf{u}_k$ for $k = r+1, \ldots, m$ and from $\boldsymbol{\omega}_r$ will remain unspecified. In line with the previous discussion these must ultimately depend on the specific $\bar{\boldsymbol{\omega}}$ around which $\mathcal{M}$ or $\mathcal{N}$ is parameterized.

Before proving the preceding statement consider first the slightly more general system of equations given by

$$\operatorname{Tr}(\boldsymbol{\Lambda}_r) + \eta = 1 \tag{48a}$$

$$\operatorname{diag}(\mathbf{U}_r \boldsymbol{\Lambda}_r \mathbf{U}_r) + \boldsymbol{\gamma} = \mathbf{U}_r \mathbf{z}_r \tag{48b}$$

$$\mathbf{U}_r^{\mathrm{T}} \mathbf{U}_r = \mathbf{I} \tag{48c}$$

where $(\mathbf{U}_r, \boldsymbol{\lambda}_r, \mathbf{z}_r, \boldsymbol{\gamma}, \eta) \in \mathbb{R}^{m \times r} \times \mathbb{R}^r \times \mathbb{R}^r \times \mathbb{R}^m \times \mathbb{R}^1$ for some r, $1 \leq r \leq m$. For now, it is sufficient to view the addition of $\boldsymbol{\gamma}$ and η as (small) perturbations of the constraints in (48). These will later be used to develop a perturbation analysis of the solutions to (42) around the rank r points.

Let

$$\boldsymbol{\omega}_r \triangleq (\mathbf{U}_r, \boldsymbol{\lambda}_r, \mathbf{z}_r)$$

and define $\bar{\boldsymbol{\omega}}_r$ analogously. Define

$$H_r : \mathbb{R}^{q_r + m + 1} \mapsto \mathbb{R}^{p_r}$$

according to

$$H_r(\boldsymbol{\omega}_r, \boldsymbol{\gamma}, \eta) \triangleq \begin{bmatrix} \operatorname{Tr}(\boldsymbol{\Lambda}_r^2) + \eta - 1 \\ \operatorname{diag}(\mathbf{U}_r \boldsymbol{\Lambda}_r^2 \mathbf{U}_r^{\mathrm{T}}) + \boldsymbol{\gamma} - \mathbf{U}_r \mathbf{z}_r \\ \operatorname{svec}(\mathbf{U}_r^{\mathrm{T}} \mathbf{U}_r - \mathbf{I}) \end{bmatrix}$$

and note that $H_r(\boldsymbol{\omega}_r, \boldsymbol{\gamma}, \eta) = \mathbf{0}$ is equivalent to (48). In order to establish that the solution set of (48) can (locally around a particular solution ($\bar{\boldsymbol{\omega}}_r$,$\mathbf{0}$,0)) be parameterized by $q_r - p_r + m + 1$ parameters it is sufficient to establish that the Jacobian

$$\mathbf{D}_r = \left.\frac{\partial H_r(\bar{\boldsymbol{\omega}}_r)}{\partial \bar{\boldsymbol{\omega}}_r}\right|_{\boldsymbol{\omega}=\bar{\boldsymbol{\omega}}} \in \mathbb{R}^{p_r \times q_r} \tag{49}$$

is full rank when evaluated at $\bar{\boldsymbol{\omega}}_r$ satisfying $H_r(\bar{\boldsymbol{\omega}}_r, \mathbf{0}, 0) = \mathbf{0}$.

Note that, similarly to before, if $\mathbf{D}_r$ in (49) is full rank then this implies the existence of a Lipschitz continuous function

$$\psi_r : \mathcal{D}_r \mapsto \mathcal{R}_r \tag{50}$$

where $(\mathbf{U}_r, \boldsymbol{\lambda}_r, \mathbf{z}_r) = \psi_r(\boldsymbol{\xi}_r, \boldsymbol{\gamma}, \eta)$ for $\boldsymbol{\xi}_r \in \mathbb{R}^{q_r - p_r}$, where $\mathcal{D}_r \in \mathbb{R}^{q_r - p_r + m + 1}$ is an open neighborhood of $\mathbf{0}$, and where $\mathcal{R}_r = \varphi_r(\mathcal{D}_r)$. Also, without loss of generality it can be assumed that

$$\mathcal{D}_r = (-\kappa, \kappa)^{q_r - p_r + m + 1}.$$

In order to establish the full rank property of $\mathbf{D}_r$ consider the matrix

$$\tilde{\mathbf{D}}_r \triangleq \left.\frac{\partial H_r(\bar{\boldsymbol{\omega}}_r)}{\partial(\mathbf{g}_1^{\mathrm{T}}, \ldots, \mathbf{g}_m^{\mathrm{T}}, \mathbf{z}_r^{\mathrm{T}}, \boldsymbol{\lambda}_r^{\mathrm{T}})}\right|_{\boldsymbol{\omega}=\bar{\boldsymbol{\omega}}}$$

where $\mathbf{g}_k$ is the kth *row* of $\mathbf{U}_r$, i.e.

$$\mathbf{U}_r = \begin{bmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_r \end{bmatrix} = \begin{bmatrix} \mathbf{g}_1 & \cdots & \mathbf{g}_m \end{bmatrix}^{\mathrm{T}}.$$

Note that $\tilde{\mathbf{D}}_r$ is related to $\mathbf{D}_r$ by a permutation of the columns (due to a changed order of differentiation) and that $\tilde{\mathbf{D}}_r$ is full rank if and only if $\mathbf{D}_r$ is full rank. Computing $\tilde{\mathbf{D}}_r$ (semi) explicitly yields

$$\tilde{\mathbf{D}}_r = \begin{bmatrix} \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \bar{\boldsymbol{\lambda}}_r^{\mathrm{T}} \\ 2\bar{\mathbf{g}}_1^{\mathrm{T}}\bar{\boldsymbol{\Lambda}}_r^2 - \bar{\mathbf{z}}_r^{\mathrm{T}} & \cdots & \mathbf{0}^{\mathrm{T}} & \bar{\mathbf{g}}_1^{\mathrm{T}} & 2\bar{\boldsymbol{\Lambda}}_r\bar{\mathbf{g}}_1^2 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0}^{\mathrm{T}} & \cdots & 2\mathbf{g}_m^{\mathrm{T}}\bar{\boldsymbol{\Lambda}}_r^2 - \bar{\mathbf{z}}_r^{\mathrm{T}} & \bar{\mathbf{g}}_m^{\mathrm{T}} & 2\bar{\boldsymbol{\Lambda}}_r\bar{\mathbf{g}}_m^2 \\ \bar{\mathbf{G}}_1 & \cdots & \bar{\mathbf{G}}_m & \mathbf{0} & \mathbf{0} \end{bmatrix}$$

where

$$\bar{\mathbf{G}}_k \triangleq \left.\frac{\partial G_r(\mathbf{U}_r)}{\partial \mathbf{g}_k}\right|_{\boldsymbol{\omega}_r=\bar{\boldsymbol{\omega}}_r} \quad \text{for} \quad G_r(\mathbf{U}_r) \triangleq \mathrm{svec}(\mathbf{U}_r^{\mathrm{T}}\mathbf{U}_r - \mathbf{I})$$

and where $\bar{\mathbf{g}}_i^2$ denotes element wise squaring of $\bar{\mathbf{g}}_i$. Assume first that $2\bar{\mathbf{g}}_i^{\mathrm{T}}\bar{\boldsymbol{\Lambda}}_r^2 - \bar{\mathbf{z}}_r^{\mathrm{T}} = \mathbf{0}$ for some i, $1 \le i \le m$. This implies through (48b) (and $\boldsymbol{\gamma} = \mathbf{0}$) that

$$\bar{\mathbf{g}}_i^{\mathrm{T}}\bar{\boldsymbol{\Lambda}}_r^2\bar{\mathbf{g}}_i = 2\bar{\mathbf{g}}_i^{\mathrm{T}}\bar{\boldsymbol{\Lambda}}_r^2\bar{\mathbf{g}}_i$$

and in turn $\bar{\mathbf{\Lambda}}_r^2\bar{\mathbf{g}}_i = \mathbf{0}$ as $\mathbf{\Lambda}_r^2 \succeq \mathbf{0}$. Further, it follows that $\bar{\mathbf{z}}_r = \mathbf{0}$ and that $\bar{\mathbf{\Lambda}}_r = \mathbf{0}$ by inserting $\bar{\mathbf{z}}_r = \mathbf{0}$ into (48b). This however violates (48a) and contradicts that $\bar{\boldsymbol{\omega}}_r$ is a solution to (48). Thus, it can be assumed that $2\bar{\mathbf{g}}_i^{\mathrm{T}}\bar{\mathbf{\Lambda}}_r^2 - \bar{\mathbf{z}}_r^{\mathrm{T}} \neq \mathbf{0}$ for all $i = 1, \ldots, m$ which implies that the first $m+1$ rows of $\tilde{\mathbf{D}}_r$ are linearly independent.

Establishing that the last $r(r+1)/2$ rows of $\tilde{\mathbf{D}}_r$ are linearly independent is a standard exercise in proving that the (m, r)-Stiefel manifold (the set of m by r unitary matrices) has dimension

$$mr - \frac{r(r+1)}{2}$$

which is a well known result [32]. We will for this reason not provide an explicit proof of this. In fact, the last $r(r+1)/2$ rows of $\tilde{\mathbf{D}}_r$ are not only linearly independent but also orthogonal.

What now remains to be done, in order to show that $\tilde{\mathbf{D}}_r$ is full rank, is to prove that none of the first $m+1$ rows can be written as a linear combination of the remaining $r(r+1)/2$ rows. For the first row, this is obvious due to the structure of $\tilde{\mathbf{D}}_r$ together with $\bar{\boldsymbol{\lambda}}_r \neq \mathbf{0}$. For the next m rows the only potential problem would be if $\mathbf{g}_i = \mathbf{0}$ for some i. However, as

$$G_r(\mathbf{U}_r) = \mathrm{svec}(\mathbf{U}_r^{\mathrm{T}}\mathbf{U}_r - \mathbf{I}) = \sum_{i=1}^{m} \mathrm{svec}(\mathbf{g}_i\mathbf{g}_i^{\mathrm{T}}) - \mathrm{svec}(\mathbf{I})$$

it follows that $\bar{\mathbf{G}}_i$ is linear in $\bar{\mathbf{g}}_i$ and equal to zero whenever $\bar{\mathbf{g}}_i = \mathbf{0}$. Together with the property that $2\bar{\mathbf{g}}_i^{\mathrm{T}}\bar{\mathbf{\Lambda}}_r^2 - \bar{\mathbf{z}}_r^{\mathrm{T}} \neq \mathbf{0}$ it follows that none of the first $m+1$ rows can be formed as a linear combination of the remaining $r(r+1)/2$ rows. This establishes that $\tilde{\mathbf{D}}_r$, and $\mathbf{D}_r$, are full rank. Note that as

$$\mathbf{D} = \mathbf{D}_m$$

it also follows that the assertion of (45) has been proven.

Consider again the parametrization of $\mathcal{N}$ around some rank r $\bar{\boldsymbol{\omega}} \in \mathcal{N}$ and consider the matrix

$$\mathbf{P} = \left.\frac{\partial H(\boldsymbol{\omega})}{\partial(\boldsymbol{\omega}_r, \mathbf{u}_{r+1}, \ldots, \mathbf{u}_m)}\right|_{\boldsymbol{\omega}=\bar{\boldsymbol{\omega}}}.$$

Note that $\mathbf{P}$ is nothing more than $\mathbf{D}$ with the columns corresponding to λ_k and z_k for $k = r+1, \ldots, m$ removed. It is straightforward to verify that $\mathbf{P}$ is structured as

$$\mathbf{P} = \begin{bmatrix} \mathbf{D}_r & \mathbf{0} & \cdots & \mathbf{0} \\ \times & \bar{\mathbf{F}}_{r+1}^{\mathrm{T}} & \cdots & \mathbf{0} \\ \times & \times & \ddots & \vdots \\ \times & \times & \times & \bar{\mathbf{F}}_m^{\mathrm{T}} \end{bmatrix} \tag{51}$$

where

$$\bar{\mathbf{F}}_k = \begin{bmatrix} \bar{\mathbf{u}}_1 & \cdots & \bar{\mathbf{u}}_{k-1} & 2\bar{\mathbf{u}}_k \end{bmatrix} \tag{52}$$

and where $\bar{\mathbf{u}}_i$ is the ith column of $\bar{\mathbf{U}}$ in $(\bar{\mathbf{U}}, \bar{\boldsymbol{\lambda}}, \bar{\mathbf{z}}) = \bar{\boldsymbol{\omega}}$. The structure of (52) follows by differentiating $\mathrm{svec}(\mathbf{U}_r^{\mathrm{T}}\mathbf{U}_r - \mathbf{I})$ with respect to the kth column of $\mathbf{U}_r$ (remember that svec forms a vector of the *upper* triangular part of its matrix argument). Note that $\mathbf{F}_k^{\mathrm{T}} \in \mathbb{R}^{k\times m}$ is full rank for any k, $1 \leq k \leq m$, (as the rows are orthogonal) and that $\mathbf{D}_r \in \mathbb{R}^{p_r \times q_r}$ is full rank as proven earlier. By considering the structure of $\mathbf{P}$ it follows that a linearly independent set of columns can be selected by choosing p_r columns form the set of columns containing $\mathbf{D}_r$ and k columns from each set containing $\mathbf{F}_k$ for $k = r+1, \ldots, m$. This, as elaborated on earlier, is however equivalent to the statement that the set of solutions to (42) can locally around $\bar{\boldsymbol{\omega}}$ be parameterized by specifying $q_r - p_r$ parameters from $\boldsymbol{\omega}_r$, $m-k$ parameters from $\mathbf{u}_k$ along with λ_k and z_k for $k = r+1, \ldots, m$.

Now, turn attention to the original problem posed by Lemma 4, that is, the problem of obtaining a covering of $\mathcal{A}(\mathbf{a}, \mathbf{b})$ defined in (25) and where $\mathbf{a} = (a_1, \cdots, a_m)$, $\mathbf{b} = (b_1, \cdots, b_m)$ and $0 \leq b_1 \leq \ldots \leq b_m$. Let r be the maximum integer for which

$$0 = b_1 = \ldots = b_r < b_{r+1} \leq \ldots \leq b_m.$$

As stated earlier, if $b_1 > 0$ then $\mathcal{A}(\mathbf{a}, \mathbf{b})$ will be empty for sufficiently small ϵ. It is thus safe to assume that $b_1 = 0$ and $r \geq 1$. Further, it can without loss of generality be assumed that ϵ is arbitrary small. In particular, it can be assumed that

$$\epsilon^{\frac{b_{r+1}}{2}} < \kappa$$

where κ is the constant introduced in (46).

Consider the set

$$\mathcal{M}(\mathbf{b}) \triangleq \mathcal{M} \cap \{(\mathbf{U}, \boldsymbol{\lambda}, \mathbf{z}) \mid |\lambda_i| \leq \epsilon^{\frac{b_i}{2}}\}.$$

The set $\mathcal{M}(\mathbf{b})$ is chosen such that any matrix $\mathbf{A} \in \mathcal{A}(\mathbf{a}, \mathbf{b})$ can be expressed as $\mathbf{A} = \mathbf{U}\boldsymbol{\Lambda}$ for some $(\mathbf{U}, \boldsymbol{\lambda}, \mathbf{z}) \in \mathcal{M}(\mathbf{b})$. Thus, the parametrization of $\mathcal{M}(\mathbf{b})$ will also provide a parametrization of $\mathcal{A}(\mathbf{a}, \mathbf{b})$.

Let $\{\psi^{(l)}\}_{l=1}^{L}$ be a set of parameterizations (around rank r points) such that

$$\mathcal{M}(\mathbf{b}) \subset \bigcup_{l=1}^{L} \mathcal{R}^{(l)} \tag{53}$$

where $\mathcal{R}^{(l)} \triangleq \psi^{(l)}(\mathcal{D})$. The assumption that $\epsilon^{\frac{b_{r+1}}{2}} \leq \kappa$ ensures that it is suffice to consider parameterizations around rank r points, $\bar{\boldsymbol{\omega}} \in \mathcal{N}$, in order to cover $\mathcal{M}(\mathbf{b})$. Note also that by the assumption in (46) the coordinate neighborhoods of $\psi^{(l)}$ are all equal to $\mathcal{D}$. Further, since $\mathcal{M}(\mathbf{b}) \subset \mathcal{N}$ is compact (and since $\mathcal{R}^{(l)}$ is open) it can be assumed that L is finite [31]. Define $\mathcal{D}^{(l)}(\mathbf{b})$ according to

$$\mathcal{D}^{(l)}(\mathbf{b}) \triangleq \psi^{-1}(\mathcal{M}(\mathbf{b}) \cap \mathcal{R}^{(l)})$$

and note that $\mathcal{D}^{(l)}(\mathbf{b}) \subset \mathcal{D}$. Finally, define

$$\mathcal{P}^{(l)}(\mathbf{b}) \triangleq \{\mathbf{A} \mid \exists \mathbf{z},\ (\mathbf{U}, \boldsymbol{\lambda}, \mathbf{z}) \in \mathcal{M}(\mathbf{b}) \cap \mathcal{R}^{(l)},\ \mathbf{A} = \mathbf{U}\boldsymbol{\Lambda}\}$$

where $\boldsymbol{\Lambda} \triangleq \mathrm{Diag}(\boldsymbol{\lambda})$ and note that

$$\mathcal{A}(\mathbf{a}, \mathbf{b}) \subset \bigcup_{l=1}^{L} \mathcal{P}^{(l)}(\mathbf{b}). \tag{54}$$

So far, the existence of a specific parametrization, given by $\mathcal{I}$, has been proven. However, not much has been said regarding the properties of this particular parametrization. Thus, to specify the benefits of the particular parametrization chosen, let in the parameter vector $\boldsymbol{\xi}$ the components obtained by selecting a subset of $(\mathbf{u}_1, \lambda_1, z_1, \ldots, \mathbf{u}_r, \lambda_r, z_r)$ be denoted by $\boldsymbol{\theta}_r \in \mathbb{R}^{q_r - p_r}$. Similarly, let the components obtained from $\mathbf{u}_k$, for $k = r+1, \ldots, m$ be denoted by $\boldsymbol{\theta}_k \in \mathbb{R}^{m-k}$. That is,

$$\boldsymbol{\xi} = (\boldsymbol{\theta}_r, \boldsymbol{\theta}_{r+1}, \lambda_{r+1}, z_{r+1}, \ldots, \boldsymbol{\theta}_m, \lambda_m, z_m).$$

Further, introduce $\hat{\boldsymbol{\xi}}$ and $\tilde{\boldsymbol{\xi}}$ and partition these analogously. Assume that $\boldsymbol{\xi}, \hat{\boldsymbol{\xi}} \in \mathcal{D}^{(l)}(\mathbf{b})$, let $(\mathbf{U}, \boldsymbol{\lambda}, \mathbf{z}) = \psi^{(l)}(\boldsymbol{\xi})$ and $(\hat{\mathbf{U}}, \hat{\boldsymbol{\lambda}}, \hat{\mathbf{z}}) = \psi^{(l)}(\hat{\boldsymbol{\xi}})$ and let $\mathbf{A} = \mathbf{U}\boldsymbol{\Lambda}$ and $\hat{\mathbf{A}} = \hat{\mathbf{U}}\hat{\boldsymbol{\Lambda}}$ where $\hat{\boldsymbol{\Lambda}} \triangleq \mathrm{Diag}(\hat{\boldsymbol{\lambda}})$. Further, let $\tilde{\mathbf{A}} = \hat{\mathbf{A}} - \mathbf{A}$, i.e. $\tilde{\mathbf{A}}$ is the perturbation in $\mathbf{A}$ resulting from a perturbation, $\tilde{\boldsymbol{\xi}} \triangleq \hat{\boldsymbol{\xi}} - \boldsymbol{\xi}$, of $\boldsymbol{\xi}$. The objective is now to show that if $\tilde{\boldsymbol{\xi}} \in \mathcal{C}$ where

$$\mathcal{C} \triangleq \{\tilde{\boldsymbol{\xi}} \mid \|\tilde{\boldsymbol{\theta}}_r\|_\infty \leq c\epsilon^{\frac{1}{2}},\ \|\tilde{\boldsymbol{\theta}}_k\|_\infty \leq c\epsilon^{\frac{1-b_k}{2}},\ |\tilde{\lambda}_k| \leq c\epsilon^{\frac{1}{2}},$$
$$|\tilde{z_k}| \leq c\epsilon^{\frac{1}{2}},\ k = r+1, \ldots, m\}$$

and c is some (yet to be defined) constant it will follow that

$$\|\hat{\mathbf{A}} - \mathbf{A}\| = \|\tilde{\mathbf{A}}\| \leq \epsilon^{\frac{1}{2}}. \tag{55}$$

In the above and in the following, $\hat{\lambda}_k$, $\tilde{\lambda}_k$, $\hat{z}_k$ and $\tilde{z}_k$ refer to the kth component of $\hat{\boldsymbol{\lambda}}$, $\tilde{\boldsymbol{\lambda}}$, $\hat{\mathbf{z}}$ and $\tilde{\mathbf{z}}$ respectively.

Let $\mathbf{u}_k$ and $\hat{\mathbf{u}}_k$ denote the kth columns of $\mathbf{U}$ and $\hat{\mathbf{U}}$. Let

$$(\tilde{\mathbf{U}}, \tilde{\boldsymbol{\lambda}}, \tilde{\mathbf{z}}) = (\hat{\mathbf{U}}, \hat{\boldsymbol{\lambda}}, \hat{\mathbf{z}}) - (\mathbf{U}, \boldsymbol{\lambda}, \mathbf{z})$$

and let $\tilde{\mathbf{u}}_k$ denote the kth column of $\tilde{\mathbf{U}}$. The first step is to prove that $\|\tilde{\mathbf{u}}_k\|_\infty \leq cK_k\epsilon^{\frac{1-b_k}{2}}$ for some constant K_k. Note that since $b_1 \leq \ldots \leq b_m$ it follows immediately from the Lipschitz continuity of ψ that $\|\tilde{\mathbf{u}}_m\| \leq cK_m\epsilon^{\frac{1-b_m}{2}}$ for some constant K_m. This is since $\epsilon^{\frac{1-b_k}{2}} \leq \epsilon^{\frac{1-b_m}{2}}$ for $k \leq m$ implies that $\|\tilde{\boldsymbol{\xi}}\|_\infty \leq c\epsilon^{\frac{1-b_m}{2}}$ and K_m could simply be selected as the Lipschitz constant (in ∞-norm) of ψ.

For $k < m$, let $\mathbf{U}_k \in \mathbb{R}^{m\times r}$ be the matrix consisting of the first k columns of $\mathbf{U}$, let $\boldsymbol{\lambda}_k \in \mathbb{R}^k$ the vector of the first k elements of $\boldsymbol{\lambda}$ and let $\mathbf{z}_k \in \mathbb{R}^k$ be the vector of the first k elements of $\mathbf{z}$. Assume that $\|\tilde{\mathbf{u}}_i\| \leq cK_i\epsilon^{\frac{1-b_i}{2}}$ for some $k < i \leq m$ and note that $(\mathbf{U}_k, \boldsymbol{\lambda}_k, \mathbf{z}_k)$ must satisfy (48) for

$$\boldsymbol{\gamma} = \sum_{i=k+1}^{m} \lambda_i^2 \mathrm{diag}(\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}) - \mathbf{u}_i z_i$$

and

$$\eta = \sum_{i=k+1}^{m} \lambda_i^2.$$

Note also that, by the structure of $\mathbf{P}$ in (51) it follows that

$$\begin{aligned}&(\mathbf{U}_k, \boldsymbol{\lambda}_k, \mathbf{z}_k) = \\ &\psi_k(\boldsymbol{\theta}_r, \boldsymbol{\theta}_{r+1}, \lambda_{r+1}, z_{r+1}, \ldots, \boldsymbol{\theta}_k, \lambda_k, z_k, \boldsymbol{\gamma}, \eta)\end{aligned} \tag{56}$$

where ψ_k is the function given by the implicit function theorem in (50). By expanding

$$\begin{aligned}\hat{\boldsymbol{\gamma}} &\triangleq \sum_{i=k+1}^{m} \hat{\lambda}_i^2 \mathrm{diag}(\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i^{\mathrm{T}}) - \hat{\mathbf{u}}_i\hat{z}_i \\ &= \sum_{i=k+1}^{m} (\lambda_i + \tilde{\lambda}_i)^2 \mathrm{diag}((\mathbf{u}_i + \tilde{\mathbf{u}}_i)(\mathbf{u}_i + \tilde{\mathbf{u}}_i)^{\mathrm{T}}) \\ &\quad - (\mathbf{u}_i + \tilde{\mathbf{u}}_i)(z_i + \tilde{z}_i)\end{aligned}$$

and

$$\hat{\eta} \triangleq \sum_{i=k+1}^{m} \lambda_i^2 = \sum_{i=k+1}^{m} (\lambda_i + \tilde{\lambda}_i)^2$$

it is straightforward to show that $\tilde{\boldsymbol{\gamma}} \triangleq \hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}$ and $\tilde{\eta} \triangleq \hat{\eta} - \eta$ satisfies

$$\|\tilde{\boldsymbol{\gamma}}\|_\infty \leq c\tilde{K}_k\epsilon^{\frac{1}{2}} \quad \text{and} \quad |\eta| \leq c\tilde{K}_k\epsilon^{\frac{1}{2}}$$

for some constant $\tilde{K}_k$. In essence, the potentially large perturbation (on the order or $\epsilon^{\frac{1-b_i}{2}}$) in $\boldsymbol{\theta}_i$ for i, $k < i \leq m$ is always multiplied by factors on the order of $\epsilon^{\frac{b_i}{2}}$ which results in a perturbation, $\tilde{\boldsymbol{\gamma}}$, on the order of $\epsilon^{\frac{1}{2}}$. Note also that it is implicitly assumed that ϵ is such that $c\tilde{K}_k\epsilon^{\frac{1}{2}} \leq \kappa$ or otherwise $(\boldsymbol{\omega}_r, \boldsymbol{\gamma}, \eta) \notin \mathcal{D}_r$. However, as ϵ can be assumed arbitrary small this is not a problem.

By the Lipschitz continuity of ψ_k in (50), it follows that

$$\|\tilde{\mathbf{u}}_k\|^2 \leq cK_k\epsilon^{\frac{1-b_k}{2}}$$

for some constant K_k since the argument in (56) is bounded by

$$\max(c\epsilon^{\frac{1-b_k}{2}}, c\tilde{K}_k\epsilon^{\frac{1}{2}}) \leq c\tilde{K}_k\epsilon^{\frac{1-b_k}{2}}.$$

By induction it follows that $\|\tilde{\mathbf{u}}_k\|^2 \leq cK_k\epsilon^{\frac{1-b_k}{2}}$ for $k = r+1,\ldots,m$ and $\|\tilde{\mathbf{u}}_k\| \leq cK_r\epsilon^{\frac{1}{2}}$ for $k = 1,\ldots,r$ where K_k, $k = r,\ldots,m$, are constants independent of ϵ and c. Now, by expanding

$$\begin{aligned}\hat{\mathbf{A}} =&\hat{\mathbf{U}}\hat{\mathbf{\Lambda}} = (\mathbf{U}+\tilde{\mathbf{U}})(\mathbf{\Lambda}+\tilde{\mathbf{\Lambda}})\\ =&\mathbf{U}\mathbf{\Lambda}+\mathbf{U}\tilde{\mathbf{\Lambda}}+\tilde{\mathbf{U}}\mathbf{\Lambda}+\tilde{\mathbf{U}}\tilde{\mathbf{\Lambda}}\end{aligned}$$

it follows that $\tilde{\mathbf{A}} \triangleq \hat{\mathbf{A}}-\mathbf{A}$ satisfies $\|\tilde{\mathbf{A}}\| \leq cK\epsilon^{\frac{1}{2}}$ for some constant, K. Finally, by selecting c according to $c = K^{-1}$ it follows that

$$\|\tilde{\mathbf{A}}\| = \|\hat{\mathbf{A}}-\mathbf{A}\| \leq \epsilon^{\frac{1}{2}}.$$

What has been shown so far is that a perturbation, $\tilde{\boldsymbol{\xi}}$, around a point, $\boldsymbol{\xi}$, in the parameter space $\mathcal{D}^{(l)}$ will, given that $\tilde{\boldsymbol{\xi}} \in \mathcal{C}$, result in a perturbation of $\mathbf{A}$, $\tilde{\mathbf{A}}$, which satisfies $\|\tilde{\mathbf{A}}\| \leq \epsilon^{\frac{1}{2}}$. This implies that given a set of $\boldsymbol{\xi} \in \mathcal{D}^{(l)}(\mathbf{b})$, $\{\boldsymbol{\xi}^{(l,i)}\}_{i=1}^{I}$, for which

$$\mathcal{D}^{(l)}(\mathbf{b}) \subset \bigcup_{i=1}^{I} \mathcal{C}(\boldsymbol{\xi}^{(l,i)})$$

where

$$\mathcal{C}(\boldsymbol{\xi}) \triangleq \mathcal{C}+\boldsymbol{\xi},$$

we will also have a covering of $\mathcal{P}^{(l)}(\mathbf{b})$ given by

$$\mathcal{P}^{(l)}(\mathbf{b}) \subset \bigcup_{i=1}^{I} \mathcal{A}_\epsilon(\mathbf{A}^{(l,i)}) \tag{57}$$

where $\mathbf{A}^{(l,i)} = \mathbf{U}^{(l,i)}\mathbf{\Lambda}^{(l,i)}$,

$$(\mathbf{U}^{(l,i)},\boldsymbol{\lambda}^{(l,i)},\mathbf{z}^{(l,i)}) \triangleq \psi^{(l)}(\boldsymbol{\xi}^{(l,i)}),$$

$\mathbf{\Lambda}^{(l,i)} \triangleq \mathrm{Diag}(\boldsymbol{\lambda}^{(l,i)})$ and where $\mathcal{A}_\epsilon(\mathbf{A})$ is defined in (27). However, as $\mathcal{C}(\boldsymbol{\xi})$ is simply a (rectangular) box centered at $\boldsymbol{\xi}$ and since

$$\begin{aligned}\mathcal{D}^{(l)}(\mathbf{b}) \subset \{\boldsymbol{\xi} \mid &\|\boldsymbol{\theta}_r\|_\infty \leq 2,\ \|\boldsymbol{\theta}_k\|_\infty \leq 1,\ |\lambda_k| \leq \epsilon^{\frac{b_k}{2}},\\ &|z_k| \leq 2\epsilon^{\frac{b_k}{2}},\ k = r+1,\ldots,m\}\end{aligned} \tag{58}$$

it follows that $\{\boldsymbol{\xi}^{(l,i)}\}_{i=1}^{I}$ could be chosen such that

$$I \dot{\leq} \epsilon^{-\mu}$$

where

$$\mu = \frac{(q_r-p_r)}{2} + \sum_{k=r+1}^{m} \frac{(m-k)(1-b_k)^+}{2} + \frac{2(1-b_k)^+}{2}.$$

This follows from the general statement that in order to cover a large M-dimensional box with side lengths ϵ^{β_i}, $i = 1, \ldots, M$, with small boxes of side length ϵ^{α_i}, $i = 1, \ldots, M$, one needs (in the $\doteq$ sense)

$$\prod_{i=1}^{M} \epsilon^{-(\alpha_i - \beta_i)^+} = \epsilon^{-\sum_{i=1}^{M} (\alpha_i - \beta_i)^+}$$

small boxes in total. Note also that if $\alpha_i < \beta_i$ the "small" boxes are actually wider than the large box in the ith dimension which is the reason for the $(\alpha_i - \beta_i)^+$ expression as opposed to $(\alpha_i - \beta_i)$.

By noting that

$$q_r - p_r = (m+2)r - m - \frac{r(r+1)}{2} - 1 = \sum_{k=2}^{r} m - k + 2$$

and using the assumption that $b_k = 0$ for $k = 1, \ldots, r$ it follows that μ can be written as

$$\mu = \sum_{k=2}^{m} \frac{(m-k+2)(1-b_k)^+}{2}.$$

Thus, it has so far been shown that it is possible to cover $\mathcal{P}^{(l)}$ by $I \dot{\leq} \epsilon^{-\mu}$ sets $\mathcal{A}_\epsilon(\mathbf{A}_i)$. By (54) and since L was finite this result extends to the covering of $\mathcal{A}(\mathbf{a}, \mathbf{b})$. That is, it has been shown that there exists a covering, $\mathbf{A}$, which satisfies

$$\mathcal{A}(\mathbf{a}, \mathbf{b}) \subset \bigcup_{\mathbf{A}_i \in \mathbf{A}} \mathcal{A}_\epsilon(\mathbf{A}_i)$$

and

$$|\mathbf{A}| \dot{\leq} \epsilon^{-\mu}$$

as was asserted by Lemma 4. ■

REFERENCES

[1] S. Verdú, *Multiuser Detection*. Cambridge University Press, 1998.

[2] D. Tse and P. Wiswanath, *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.

[3] S. Verdú, "Computational complexity of multiuser detection," *Algorithmica*, vol. 4, pp. 303–312, 1989.

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

[5] P. Tan and L. Rasmussen, "The application of semidefinite programming for detection in CDMA," *IEEE Journal on Selected Areas in Communications*, vol. 19, no. 8, pp. 1442–1449, Aug. 2001.

[6] W.-K. Ma, T. N. Davidson, K. Wong, Z.-Q. Luo, and P.-C. Ching, "Quasi-maximum-likelihood multiuser detection using semi-definite relaxation with application to synchronous CDMA," *IEEE Transactions on Signal Processing*, vol. 50, no. 4, pp. 912–922, Apr. 2002.

[7] M. Abdi, H. E. Nahas, A. Jard, and E. Moulines, "Semidefinite positive relaxation of the maximum-likelihood criterion applied to multiuser detection in a CDMA context," *IEEE Signal Processing Letters*, vol. 9, no. 6, pp. 165–167, June 2002.

[8] Y. E. Nesterov, "Quality of semidefinite relaxation for nonconvex quadratic optimization," CORE, Universite Catholique de Louvain, Belgium, Tech. Rep., 1997.

[9] M. Kisialiou and Z.-Q. Lou, "Perfromance analysis of quasi-maximum-liklihood detector based on semi-definite programming," in *Proc. IEEE ICASSP'05*, 2005.

[10] H. Yao and G. W. Wornell, "Lattice-reduction-aided detectors for MIMO communication systems," in *Proc. of IEEE GLOBECOM 2002*, Nov. 2002.

[11] C. Windpassinger and R. F. H. Fisher, "Low-complexity near-maximum-likelihood detection and precoding for MIMO systems using lattice reduction," in *Proc. of IEEE ITW 2003*, Apr. 2003.

[12] M. Taherzadeh, A. Mobasher, and A. K. Khandani, "LLL reduction achieves the receive diversity in MIMO decoding," *Submitted to IEEE Transactions on Information Theory*, 2006.

[13] L. Lovász, "On the Shannon capacity of a graph," *IEEE Transactions on Information Theory*, vol. 25, no. 1, pp. 1–7, Jan. 1979.

[14] L. Lovász and A. Schrijver, "Cones of matrices and set-functions an 0-1 optimization," *SIAM Journal on Optimization*, vol. 1, no. 2, pp. 166–190, May 1991.

[15] F. Jarre, "An interior-point method for minimizing the maximum eigenvalue of a linear combination of matrices," *SIAM Journal on Control and Optimization*, vol. 31, no. 5, pp. 1360–1377, Sept. 1993.

[16] Y. Nesterov and A. Nemirovski, *Interior Point polynomial algorithms in convex programming*. SIAM, 1994.

[17] L. Vandenberghe and S. Boyd, "A primal-dual potential reduction method for problems involving matrix inequalities," *Mathematical Programming*, vol. 69, no. 1, p. 205, 1995.

[18] M. X. Goemans and D. P. Williamson, "Improved approximation algorithms for maximum cut and satisfiability problem using semi-definite programming," *Journal of the ACM*, vol. 42, no. 6, pp. 1115–1145, 1995.

[19] H. Wolkowicz, R. Saigal, and L. Vandenberghe, Eds., *Handbook of Semidefinite Programming*. Kluwer Academic Publishers, 2000.

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

[21] C. Helmberg, F. Rendl, R. Vanderbei, and H. Wolkowicz, "An interior-point method for semidefinite programming," *SIAM Journal on Optimization*, vol. 6, pp. 342–361, 1996.

[22] J. Jaldén, C. Martin, and B. Ottersten, "Semidefinite programming for detection in linear systems – optimality conditions and space-time decoding," in *Proc. IEEE ICASSP'03*, Apr. 2003.

[23] W.-K. Ma, P.-C. Ching, and Z. Ding, "Semidefinite relaxation based multiuser detection for M-ary PSK multiuser systems," *IEEE Transactions on Signal Processing*, vol. 52, no. 10, pp. 2862–2872, Oct. 2004.

[24] A. Wiesel, Y. C. Eldar, and S. Shamai, "Semidefinite relaxation for detection of 16-QAM signaling in MIMO channels," *IEEE Signal Processing Letters*, vol. 12, no. 9, pp. 653–656, Sept. 2005.

[25] S. Poljak, F. Rendl, and H. Wolkowicz, "A recipe for semidefinite relaxation for (0,1)-quadratic programming," *Journal of Global Optimization*, vol. 7, no. 1, pp. 51–73, July 1995.

[26] R. V. Nee, A. V. Zelst, and G. Awater, "Maximum likelihood decoding in a space division multiplexing system," in *IEEE VTC00*, Tokyo, Japan, May 2000.

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

[28] C. Helmberg, *Semidefinite Programming for Combinatorial Optimization*. Berlin, Germany: Konrad-Zuse-Zentrum, 2000.

[29] G. Pataki, "On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eignevalues," *Mathematics of Operations Research*, vol. 23, pp. 339–358, 1998.

[30] J. W. Milnor, *Topology from the differentiable viewpoint*. Princeton University Press, 1965.

[31] W. Rudin, *Principles of Mathematical Analysis*, 3rd ed. McGraw-Hill International Editions, 1996.

[32] W. M. Boothby, *An Introduction to Differential Manifolds and Riemannian Geometry*, 2nd ed. New York: Academic Press, 1986.

[33] R. G. Bartle, *The Elements of Real Analysis*. John Wiley & Sons, Inc., 1964.

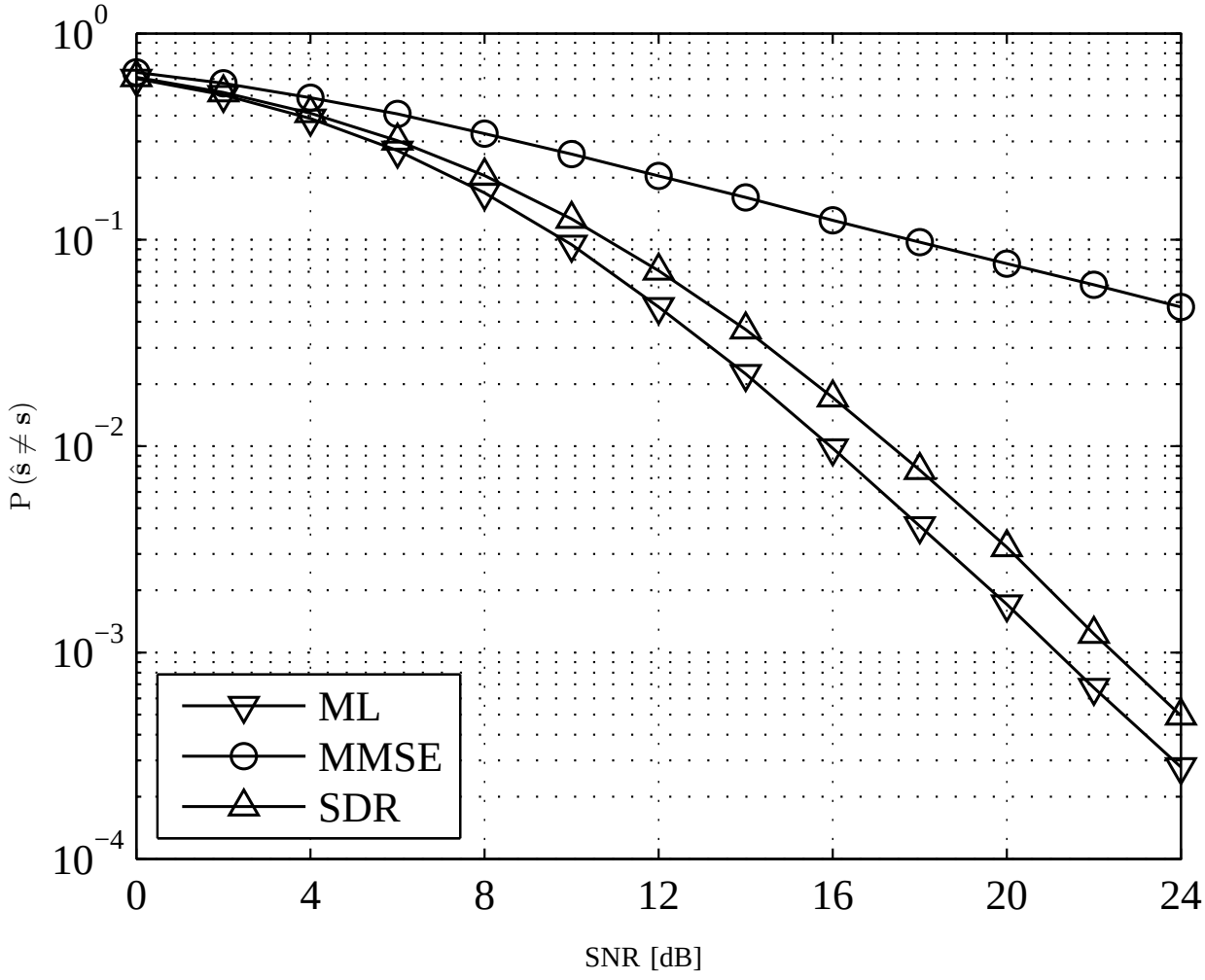

Fig. 1. The probability of error when $\mathbf{H} \in \mathbb{R}^{n \times m}$ has i.i.d. real valued Gaussian entries, and where $m = n = 4$.

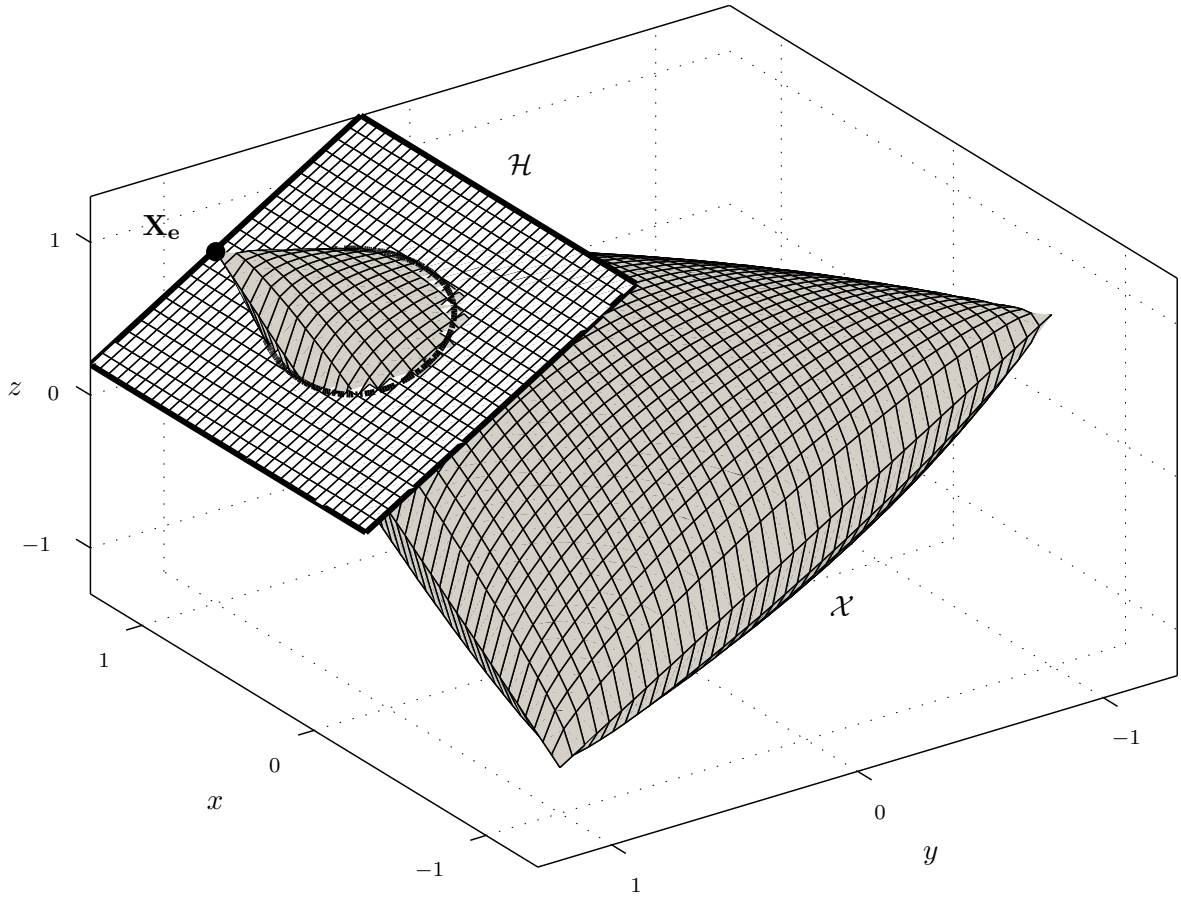

Fig. 2. Illustration of the feasible set, $\mathcal{X}$, of the SDR detector in (5). The hyperplane $\mathcal{H}$ separates points in the feasible set that are close to and far from $\mathbf{X_e}$.

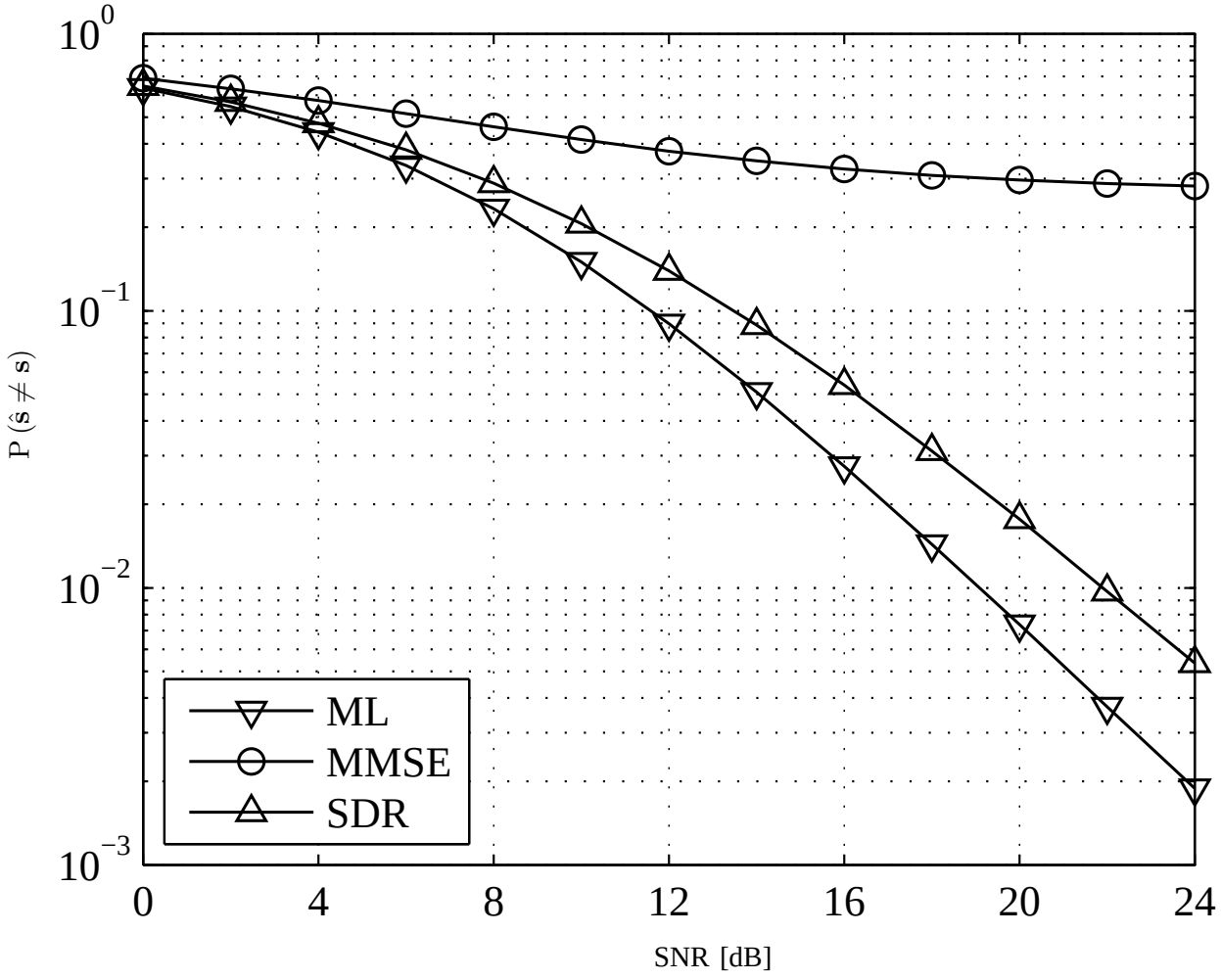

Fig. 3. The probability of error when $\mathbf{H} \in \mathbb{R}^{n \times m}$ has i.i.d. real valued Gaussian entries, and where $m = 4$ and $n = 3$.

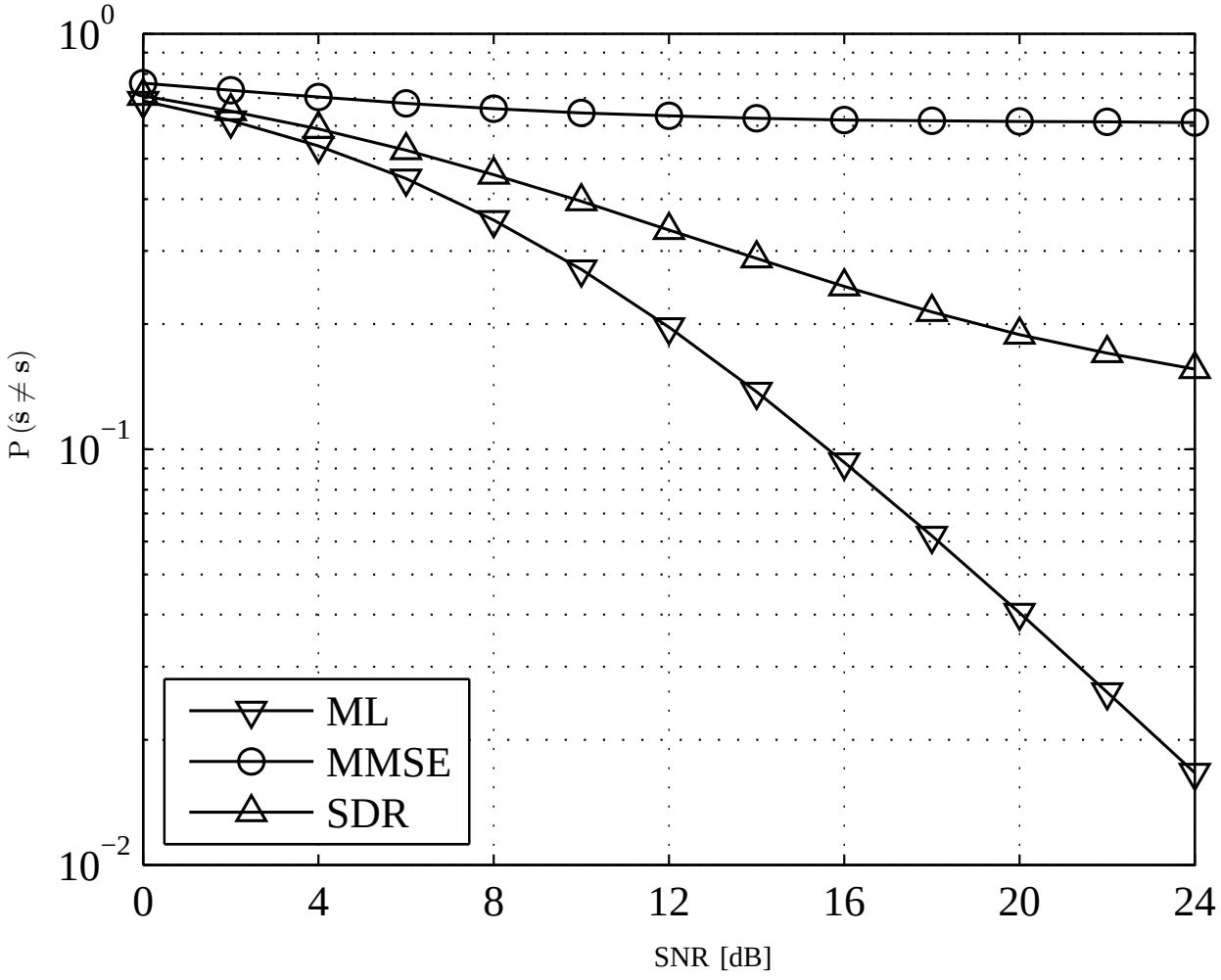

Fig. 4. The probability of error when $\mathbf{H} \in \mathbb{R}^{n \times m}$ has i.i.d. real valued Gaussian entries, and where $m = 4$ and $n = 2$.

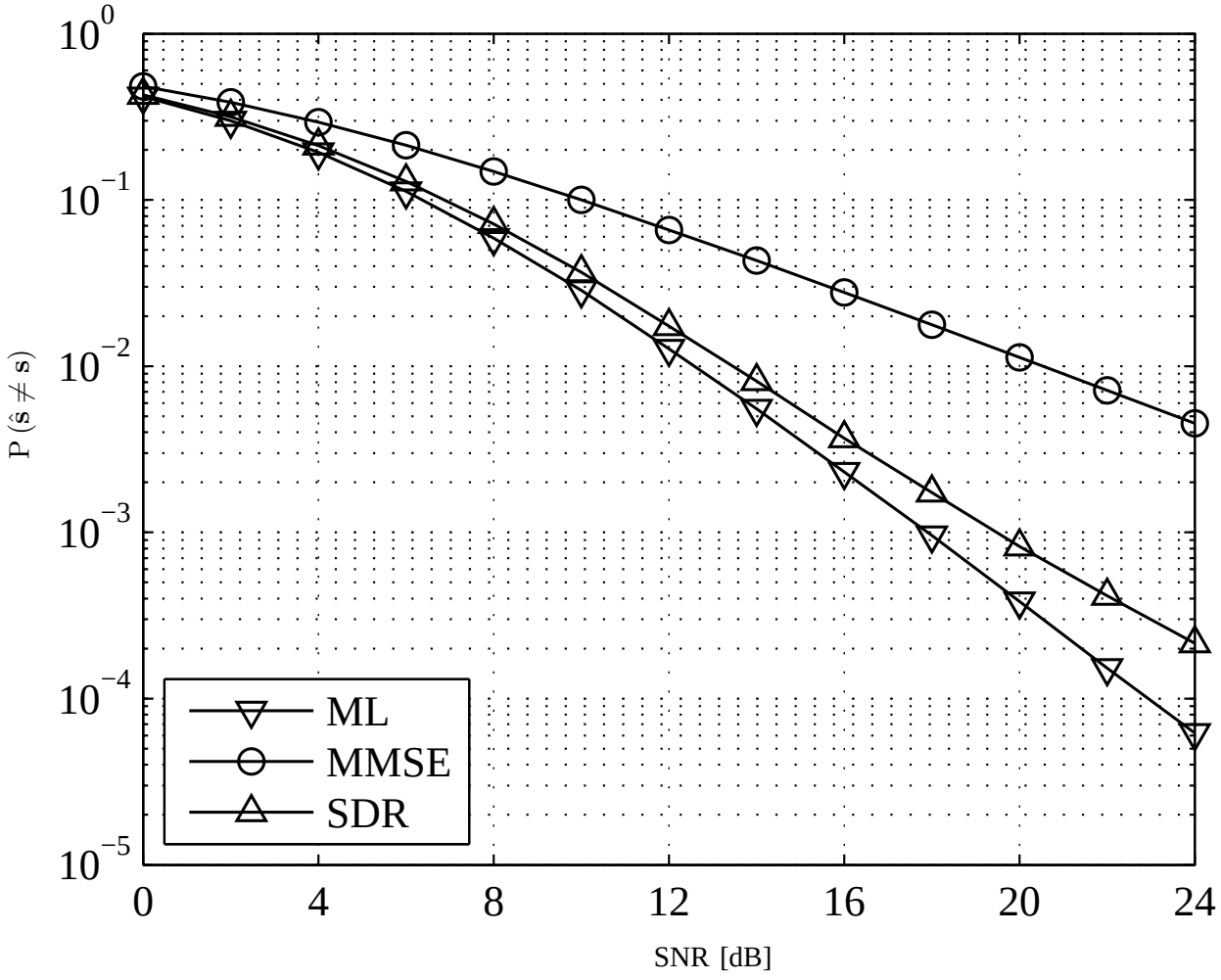

Fig. 5. The probability of error when $\mathbf{H} \in \mathbb{C}^{N \times M}$ has i.i.d. complex valued Gaussian entries, and where $N = M = 2$.

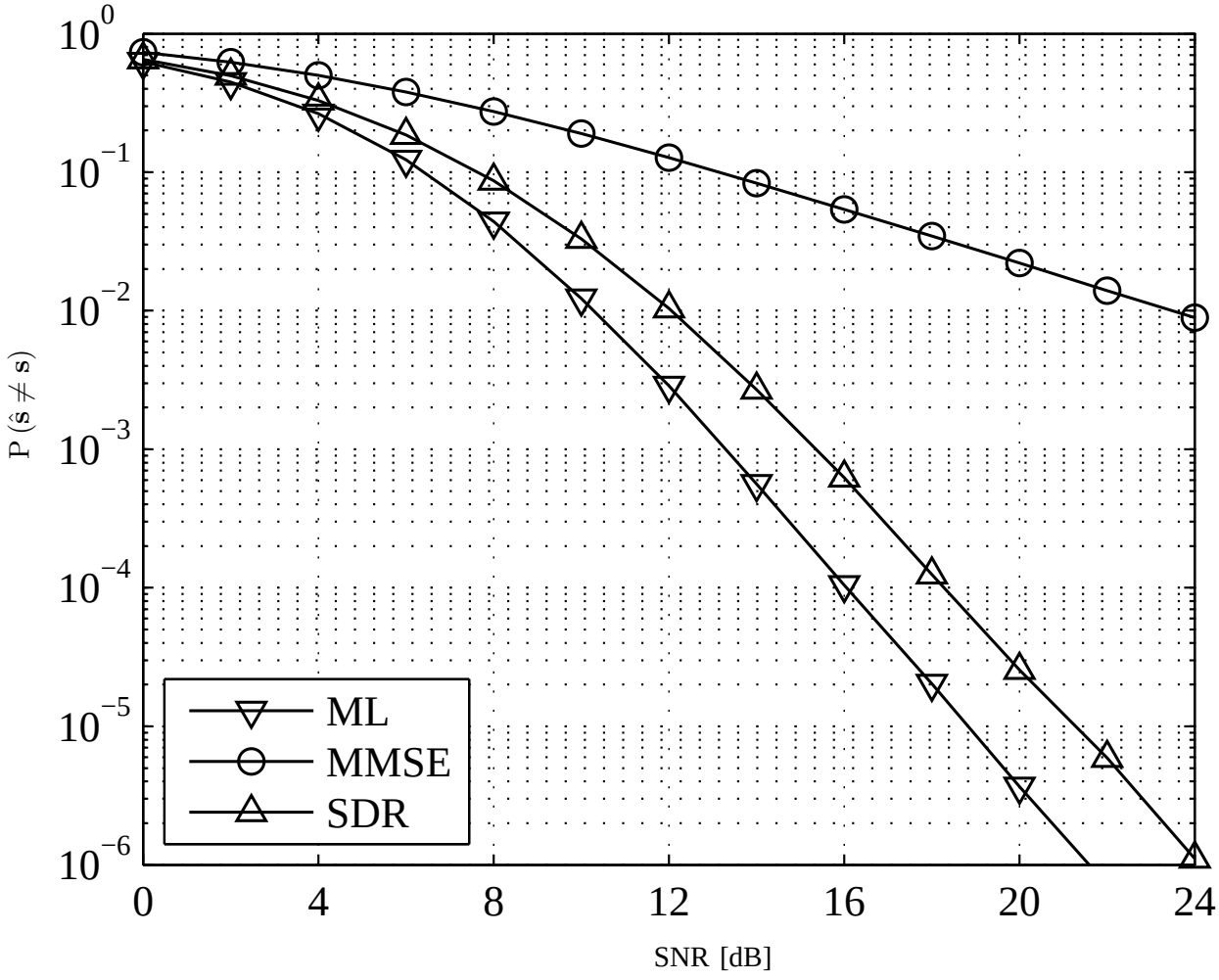

Fig. 6. The probability of error when $\mathbf{H} \in \mathbb{C}^{N \times M}$ has i.i.d. complex valued Gaussian entries, and where $N = M = 4$.