The Diversity Order of the Semidefinite Relaxation Detector

Joakim Jaldén, *Member, IEEE*, and Björn Ottersten, *Fellow, IEEE*

***Abstract*—In this paper, 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 independent identically distributed (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 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 also possible in certain scenarios when using a noncombinatorial receiver structure.**

***Index Terms*—Detection, diversity, multiple-input–multiple-output (MIMO), semidefinite relaxation (SDR).**

I. Introduction

IN THIS PAPER, we consider the detection of binary symbols transmitted over an $n \times m$ multiple-input–multiple-output (MIMO) channel modeled according to

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

where $\boldsymbol{s} \in \mathcal{B}^m \triangleq \{\pm 1\}^m$, $\boldsymbol{H} \in \mathbb{R}^{n \times m}$, and $\boldsymbol{v}, \boldsymbol{y} \in \mathbb{R}^n$. In what follows, $\boldsymbol{y}$ is referred to as the vector of *received signals*, $\boldsymbol{H}$ as the *channel matrix*, $\boldsymbol{s}$ as the *transmitted message*, and $\boldsymbol{v}$ as the additive *noise* based on their physical interpretations in the digital communications context [1]. 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 $\boldsymbol{H}$ is known to the receiver.

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 code division multiple access (CDMA) [2] and communications over a multiple antenna channel [1]. However, while the detection problem is the same for many areas, the structure and assumptions regarding the channel matrix $\boldsymbol{H}$ will typically differ depending on the specific context. In the interest of simplicity, we will assume that the channel matrix may be modeled using independent identically distributed (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 [1]. The signal-to-noise ratio (SNR) of the channel is equal to ρ and the analysis will be focussed on the high SNR regime. The maximum-likelihood (ML) estimate of $\boldsymbol{s}$, $\hat{\boldsymbol{s}}_{\mathrm{ML}}$, is given by

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

where $\|\cdot\|$ denotes the Euclidean norm, i.e., the ML detector selects the message $\hat{\boldsymbol{s}}$, which minimizes the distance between the received signals and the hypothesized noise-free message $\boldsymbol{H}\hat{\boldsymbol{s}}$. An error is declared whenever $\hat{\boldsymbol{s}}_{\mathrm{ML}} \neq \boldsymbol{s}$ and it is well known that the ML detector is optimal in the sense that it minimizes the probability of error given that all transmitted messages are *a priori* equally likely. However, for a general channel matrix $\boldsymbol{H}$ and vector of received signals $\boldsymbol{y}$, 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 for the sphere decoding algorithm.

The prohibitive complexity of the ML detector motivates the study of suboptimal (but computationally advantageous) alternatives. Examples of such suboptimal alternatives are the zero forcing (ZF) and linear minimum mean square error (LMMSE) detectors [1] and their decision feedback counterparts and the lattice reduction-aided (LRA) detectors [5], [6]. Herein, we study the semidefinite relaxation (SDR) detector that obtains an estimate of $\boldsymbol{s}$ in polynomial time. The SDR detector was (in the communications literature) first proposed in [7]–[9] for CDMA multiuser detection but is straightforwardly applicable to the detection problem considered herein. The basis of the SDR detector is a convex relaxation technique where (2) is simplified by expanding the feasible set (relaxing some of the constraints). An estimate of $\boldsymbol{s}$ is then obtained by mapping the solution to the simplified optimization problem back into $\mathcal{B}^m$ by a suitable heuristic. Also, although generalizations of the SDR detector to higher order constellations have appeared in

Manuscript received June 13, 2006; revised June 21, 2007.

J. Jaldén was with the Signal Processing Lab, School of Electrical Engineering, Royal Institute of Technology (KTH), Stockholm SE-100 44, Sweden. He is now with the Institute of Communications and Radio-Frequency Engineering, Vienna University of Technology, Vienna A-1040, Austria (e-mail: joakim.jalden@nt.tuwien.at).

B. Ottersen is with the Signal Processing Lab, School of Electrical Engineering, Royal Institute of Technology (KTH), Stockholm SE-100 44, Sweden (e-mail: bjorn.ottersten@ee.kth.se).

Communicated by H. Boche, Associate Editor for Communications.

Digital Object Identifier 10.1109/TIT.2008.917634

0018-9448/$25.00 © 2008 IEEE

the literature [10]–[13], we will herein only consider the binary case.

In this paper, we focus on the error probability performance of the SDR detector in the high SNR regime and provide an analytical proof of that the detector achieves maximal diversity. This result is formally stated by Theorem 1 in Section II-B and represents a nontrivial extension of previously known performance guarantees available for the SDR detector; see, e.g., [8], [14], and [15]. It is also interesting to note that a similar result (regarding the maximal diversity) was recently provided for the LRA detector [16]. However, the design philosophies underlying the LRA and SDR detectors are fundamentally different. Whereas the LRA is combinatorial in nature the SDR detector is based on the minimization of a continuous function over a convex set.

After a review of the SDR receiver we introduce the main contribution of this work, namely, Theorem 1 in Section II. A short outline of the proof is given in Section III, while the full proof is saved for Sections IV and V. Following is Section VI, where we discuss possible generalizations of the result and provide numerical examples.

II. Semidefinite Relaxation

The use of SDR for bounding the optimal value of a combinatorial optimization problem was first considered in the late 1970s [17] (where it was used to bound the Shannon capacity of a graph). Theoretical work in the 1990s [18] along with the introduction of practical methods for solving semidefinite programs [19]–[21] made the SDR a viable method for finding approximate solutions to many combinatorial problems.

A. The SDR Detector

The (nonconvex) optimization problem given by

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

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

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

is equivalent to (2) in the sense that the solution to (2) is easily obtained from the solution to (3) and vice verse [7], [8], [22]. The optimal point of (2) is related to the optimal point of (3) through $\boldsymbol{x}$ as indicated by (4). Naturally, as (3) and (2) are equivalent they are also equally difficult to solve from a complexity theoretic point of view. In particular, it follows from [3] that (3) is also NP-hard in general.

The SDR detector is based on solving

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

in the place of (3). In (5), $\boldsymbol{X} \succeq \boldsymbol{0}$ indicates that $\boldsymbol{X}$ is symmetric and positive definite and since $\boldsymbol{X} = \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}$ implies $\boldsymbol{X} \succeq \boldsymbol{0}$ it follows that (5) represents a relaxation of (3). Because (5) is a *convex* problem it can be efficiently solved in polynomial time [20], [23]. In particular, there is an interior point algorithm which solves (5) to any fixed precision in $O(m^{3.5})$ time [24]; see, also, [7], where this algorithm is presented in the digital communications context. Additionally, there are algorithms and implementations specifically optimized for the data model considered in this work; see, e.g., [25]. When the optimal solution to (5) is rank one it is also an optimal solution to (3). The opposite is, however, not generally true and the solution to (5) can at most serve as a basis for obtaining an approximate solution to (2) or (3) [7], [8].

There are several suggestions for obtaining an estimate of $\boldsymbol{s}$ based on the solution of (5). Among the more powerful approaches are a randomization technique [8], [26] and an approximation based on the dominant eigenvector of the optimal $\boldsymbol{X}$ in (5)[7]. Numerical evidence suggests that the randomization technique results in lower error probability. We will, however, herein only consider the strategy of simply using the signs of the last column of $\boldsymbol{X}^{\star}$ where $\boldsymbol{X}^{\star}$ is the optimal point of (5). This approach was mentioned in [7], but discarded in favor of the (superior) eigenvector approach. However, as the sign-based approach already achieves the maximum diversity and is somewhat easier to analyze, we will only consider this method in detail. It should, however, be noted that our proof extends to the dominant eigenvector method in a fairly straightforward manner and to the randomization technique given that the simple estimate (obtained by considering the signs) is included in the list of candidate solutions.

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

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

where

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

is the sign function, i.e., $\hat{\boldsymbol{s}}_{\mathrm{SDR}}$ is given by the signs of the last column of $\boldsymbol{X}^{\star}$. Note also that because

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

the procedure is guaranteed to yield the ML solution whenever $\boldsymbol{X}^{\star}$ is rank one.

B. SDR Performance

The extraordinary performance of the SDR technique in many areas have been a motivating reason for its study and there are several results in the literature regarding the quality of the SDR approximation. These include the bound of [14], which is a generalization of a previous result for the *max cut* problem [26]. There are also results relating the SDR to other relaxations [27]. In the context of digital communications it has been shown that several low-complexity detectors may be viewed as further relaxations of the SDR detector [8]. Notably, these low-complexity detectors include both the ZF and LMMSE detectors and give strong support for the SDR approach although the results in [8] relate to the objective values of the relaxations rather than directly to the quality of the estimates. Further, a probabilistic bound on the difference in optimal objective value between (5) and (3) was given in [15] for the large system limit. Necessary and sufficient conditions for the existence of rank one solutions to (5) were given in [28], where it was also established that the detector is free of an error floor when $\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}$ is full rank. However, the result in [28] does not extend to a statement regarding diversity. Specifically, it is possible to show that an alternative SDR receiver which calls an error whenever (5) is not of rank one would not achieve the maximum diversity [29, Th. 7.3]. 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 $\boldsymbol{H} \in \mathbb{R}^{n\times m}$ in (1) consists of i.i.d. Gaussian entries of zero mean and fixed (nonzero) variance. Assume further that $n \geq m$. Then

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

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 [30] to cover the real-valued case. In the case of ZF and LMMSE, the diversity is $\frac{n-m+1}{2}$, which can be seen by following the argument of [1, Sec. 8.5.1] with a real-valued channel matrix.

Following [31], throughout this work, we will make use of the symbol $\doteq$ to denote *exponential equality*, defined according to

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

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 A. Using (7) generally allows for a more compact (and suggestive) notation and in this notation the statement of Theorem 1 becomes

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

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

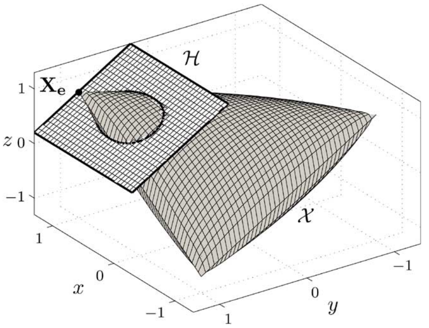

Fig. 1. Illustration of the feasible set $\mathcal{X}$ of the SDR detector in (5). The hyperplane $\mathcal{H}$ separates points in the feasible set that are close to and far from $\boldsymbol{X_e}$.

III. SDR Diversity Proof—Outline

Due to the symmetry of the problem (and the detector), it can be assumed without loss of generality that $\boldsymbol{s} = \boldsymbol{e}$ 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 $\boldsymbol{X} \in \mathcal{X}$ as in [32] or [7], i.e., according to

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

The feasible set $\mathcal{X}$ is illustrated in Fig. 1. The rank one matrix $\boldsymbol{X_e}$ that corresponds to the transmitted message $\boldsymbol{s} = \boldsymbol{e}$ 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) $\boldsymbol{X}^{\star}$ is close to $\boldsymbol{X_e}$, then the rounding procedure described in Section II will be able to recover the correct rank one matrix, namely, $\boldsymbol{X_e}$. It is only when the optimal point of (5) is far from $\boldsymbol{X_e}$ that an error can occur. In order to particularize the notion of "close to" in the proof of Theorem 1, we makes use of a hyperplane $\mathcal{H}$ as shown in Fig. 1 to single out the points in $\mathcal{X}$ that are "close to" $\boldsymbol{X_e}$. Specifically, we let $\mathcal{X}_{\boldsymbol{e}}$ be the points in $\mathcal{X}$ that are on the same side of $\mathcal{H}$ as $\boldsymbol{X_e}$ and choose $\mathcal{H}$ such that $\hat{\boldsymbol{s}}_{\mathrm{SDR}} = \boldsymbol{e}$ whenever $\boldsymbol{X}^{\star} \in \mathcal{X}_{\boldsymbol{e}}$. In the zero noise case, i.e., when $\boldsymbol{v} = \boldsymbol{0}$, $\boldsymbol{X_e}$ is always optimal in (5) with a criterion value equal to 0. It can also be shown that $\mathrm{Tr}(\boldsymbol{LX}) \geq \tau$ for $\boldsymbol{X} \notin \mathcal{X}_{\boldsymbol{e}}$ if

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

By allowing for $\boldsymbol{v} \neq \boldsymbol{0}$ while assuming that $\|\boldsymbol{v}\|$ is significantly smaller than τ, it will follow by continuity that $\mathrm{Tr}(\boldsymbol{LX_e})$ is still close to zero and that $\mathrm{Tr}(\boldsymbol{LX})$ is not significantly smaller than τ for any $\boldsymbol{X} \notin \mathcal{X}_{\boldsymbol{e}}$. This implies that there is a point in $\mathcal{X}_{\boldsymbol{e}}$ with a criterion value close to zero, while all points $\boldsymbol{X} \notin \mathcal{X}_{\boldsymbol{e}}$ have an objective value on the order of τ, and therefore, the optimum over $\mathcal{X}$ must belong to $\mathcal{X}_{\boldsymbol{e}}$ in which case $\hat{\boldsymbol{s}}_{\mathrm{SDR}} = \boldsymbol{e}$. In short, it is sufficient that τ is large in comparison with the noise in order for the SDR detector to make a correct decision. This argument is made rigorously in the proof of Lemma 1 in Section IV.

The overall proof of Theorem 1 is based on the heuristic argument described previously 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 $\boldsymbol{v}$ is atypically large. This statement is formalized by Lemma 2 in Section IV. The second part of the proof, contained in Section V, is concerned with bounding the probability that τ is atypically small. In order for τ to be small there must be at least one $\boldsymbol{X} \in \mathcal{X} \cap \mathcal{H}$ for which $\mathrm{Tr}(\boldsymbol{LX})$ is small and in essence the technique used to establish our bound 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 $\boldsymbol{X}$ for which $\mathrm{Tr}(\boldsymbol{LX})$ 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.

One does need to be careful, however, and not naively apply the union bound. This is because the probability that each ϵ-ball contains an $\boldsymbol{X}$ for which $\mathrm{Tr}(\boldsymbol{LX})$ is small depends on *where* in $\mathcal{X} \cap \mathcal{H}$ the ϵ-ball is located. Consequently, in order to obtain a sufficiently tight bound, $\mathcal{X} \cap \mathcal{H}$ must first be split into subsets with equiprobable coverings and the technically most challenging part of the proof relates to counting the number of ϵ-balls required to cover each such subset. The analysis of each particular ϵ-ball is provided by Lemma 3 and the counting argument is captured by Lemma 4 in Section V. The proof of Theorem 1, given at the end of Section V, then follows by combining Lemmas 3 and 4.

IV. SDR Diversity Proof—Part I

We begin by giving rigorous justification to the first part of the heuristic argument given in Section III and show that the noise $\boldsymbol{v}$ can effectively be removed from (or integrated out of) the analysis of the receiver diversity. Let the feasible set $\mathcal{X}$ of (5) be given by

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

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

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

where

$$\boldsymbol{M} \triangleq [\boldsymbol{I} \quad -\boldsymbol{e}] \in \mathbb{R}^{m\times m+1}. \tag{10}$$

It will be established later that $\mathcal{H}$ chosen this way is sufficient for drawing the conclusion that $\hat{\boldsymbol{s}}_{\mathrm{SDR}} = \boldsymbol{e}$ whenever $\boldsymbol{X}^{\star} \in \mathcal{X}_{\boldsymbol{e}}$. The optimal value of $\mathrm{Tr}(\boldsymbol{LX})$ over the intersection set $\mathcal{X} \cap \mathcal{H}$ is under the zero noise assumption given by

$$\tau \triangleq \min_{\boldsymbol{X}\in\mathcal{X}\cap\mathcal{H}} \mathrm{Tr}(\boldsymbol{L_0X}) \tag{11}$$

where

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

and $\boldsymbol{Q} \triangleq \boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}$. Note that $\boldsymbol{L_0}$ is equal to $\boldsymbol{L}$ in (4) when $\boldsymbol{v} = \boldsymbol{0}$ and $\boldsymbol{s} = \boldsymbol{e}$.

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. The statement is captured by Lemma 1 (note again that $\boldsymbol{s} = \boldsymbol{e}$ is assumed to be the transmitted message).

Lemma 1: Let τ be given by (11). Then

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

Proof: It follows by the linearity of the objective function that the optimal point of (5) must be on the boundary of $\mathcal{X}$ and rank deficient. Thus, consider an $\boldsymbol{X} \in \mathcal{X}$ for which $\boldsymbol{X} \not\succ \boldsymbol{0}$ ($\boldsymbol{X}$ is positive semidefinite but not positive definite) and partition $\boldsymbol{X}$ as

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{A}^{\mathrm{T}} \\ \boldsymbol{a}^{\mathrm{T}} \end{bmatrix} [\boldsymbol{A} \quad \boldsymbol{a}] = \begin{bmatrix} \boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} & \boldsymbol{A}^{\mathrm{T}}\boldsymbol{a} \\ \boldsymbol{a}^{\mathrm{T}}\boldsymbol{A} & \boldsymbol{a}^{\mathrm{T}}\boldsymbol{a} \end{bmatrix}$$

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

$$\boldsymbol{L} \triangleq \begin{bmatrix} \boldsymbol{H}^{\mathrm{T}}\boldsymbol{H} & -\boldsymbol{H}^{\mathrm{T}}\boldsymbol{y} \\ -\boldsymbol{y}^{\mathrm{T}}\boldsymbol{H} & \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} \end{bmatrix} = \begin{bmatrix} \boldsymbol{H}^{\mathrm{T}} \\ -\boldsymbol{y}^{\mathrm{T}} \end{bmatrix} [\boldsymbol{H} \quad -\boldsymbol{y}].$$

Thus

$$\begin{aligned} \mathrm{Tr}(\boldsymbol{LX}) &= \mathrm{Tr}\left(\begin{bmatrix} \boldsymbol{H}^{\mathrm{T}} \\ -\boldsymbol{y}^{\mathrm{T}} \end{bmatrix} [\boldsymbol{H} \quad -\boldsymbol{y}] \begin{bmatrix} \boldsymbol{A}^{\mathrm{T}} \\ \boldsymbol{a}^{\mathrm{T}} \end{bmatrix} [\boldsymbol{A} \quad \boldsymbol{a}]\right) \\ &= \mathrm{Tr}\left([\boldsymbol{H} \quad -\boldsymbol{y}] \begin{bmatrix} \boldsymbol{A}^{\mathrm{T}} \\ \boldsymbol{a}^{\mathrm{T}} \end{bmatrix} [\boldsymbol{A} \quad \boldsymbol{a}] \begin{bmatrix} \boldsymbol{H}^{\mathrm{T}} \\ -\boldsymbol{y}^{\mathrm{T}} \end{bmatrix}\right) \\ &= \mathrm{Tr}((\boldsymbol{HA}^{\mathrm{T}} - \boldsymbol{ya}^{\mathrm{T}})(\boldsymbol{HA}^{\mathrm{T}} - \boldsymbol{ya}^{\mathrm{T}})^{\mathrm{T}}) \\ &= \|\boldsymbol{HA}^{\mathrm{T}} - \boldsymbol{ya}^{\mathrm{T}}\|^2 \end{aligned}$$

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

$$\mathrm{Tr}(\boldsymbol{LX}) = \|\boldsymbol{H}(\boldsymbol{A}^{\mathrm{T}} - \boldsymbol{ea}^{\mathrm{T}}) - \boldsymbol{va}^{\mathrm{T}}\|^2.$$

Note that

$$\begin{aligned} \|\boldsymbol{H}(\boldsymbol{A} - \boldsymbol{ea}^{\mathrm{T}}) - \boldsymbol{va}^{\mathrm{T}}\| &\geq \|\boldsymbol{H}(\boldsymbol{A} - \boldsymbol{ea}^{\mathrm{T}})\| - \|\boldsymbol{va}^{\mathrm{T}}\| \\ &= \|\boldsymbol{H}(\boldsymbol{A} - \boldsymbol{ea}^{\mathrm{T}})\| - \|\boldsymbol{v}\| \end{aligned}$$

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

$$\|\boldsymbol{H}(\boldsymbol{A} - \boldsymbol{ea}^{\mathrm{T}})\| > 2\|\boldsymbol{v}\| \Leftrightarrow \|\boldsymbol{H}(\boldsymbol{A} - \boldsymbol{ea}^{\mathrm{T}})\|^2 > 4\|\boldsymbol{v}\|^2$$

it follows that

$$\mathrm{Tr}(\boldsymbol{LX}) > \|\boldsymbol{v}\|^2. \tag{12}$$

At the same time, for

$$\boldsymbol{X_e} \triangleq \begin{bmatrix} \boldsymbol{e} \\ 1 \end{bmatrix} [\boldsymbol{e}^{\mathrm{T}} \quad 1]$$

it follows that

$$\begin{aligned}\mathrm{Tr}(\boldsymbol{L}\boldsymbol{X}_{\boldsymbol{e}}) &= \mathrm{Tr}\left(\begin{bmatrix}\boldsymbol{H}^{\mathrm{T}}\\ -\boldsymbol{y}^{\mathrm{T}}\end{bmatrix}[\boldsymbol{H}\quad -\boldsymbol{y}]\begin{bmatrix}\boldsymbol{e}\\ 1\end{bmatrix}[\boldsymbol{e}^{\mathrm{T}}\quad 1]\right)\\ &= \mathrm{Tr}\left([\boldsymbol{H}\quad -\boldsymbol{y}]\begin{bmatrix}\boldsymbol{e}\\ 1\end{bmatrix}[\boldsymbol{e}^{\mathrm{T}}\quad 1]\begin{bmatrix}\boldsymbol{H}^{\mathrm{T}}\\ -\boldsymbol{y}^{\mathrm{T}}\end{bmatrix}\right)\\ &= \mathrm{Tr}((\boldsymbol{H}\boldsymbol{e}-\boldsymbol{y})(\boldsymbol{H}\boldsymbol{e}-\boldsymbol{y})^{\mathrm{T}})\\ &= \|\boldsymbol{H}\boldsymbol{e}-\boldsymbol{y}\|^2 = \|\boldsymbol{v}\|^2. \end{aligned} \tag{13}$$

Thus, by (12) and (13), it follows that

$$\|\boldsymbol{H}(\boldsymbol{A}-\boldsymbol{e}\boldsymbol{a}^{\mathrm{T}})\|^2 > 4\|\boldsymbol{v}\|^2 \Rightarrow \mathrm{Tr}(\boldsymbol{L}\boldsymbol{X}) > \mathrm{Tr}(\boldsymbol{L}\boldsymbol{X}_{\boldsymbol{e}}) \tag{14}$$

which implies that $\boldsymbol{X}$ cannot be optimal for (5) if

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

Now, note that

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

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

$$\begin{aligned}\|\boldsymbol{H}(\boldsymbol{A}-\boldsymbol{e}\boldsymbol{a}^{\mathrm{T}})\|^2 &= \mathrm{Tr}\left(\boldsymbol{H}\boldsymbol{M}\begin{bmatrix}\boldsymbol{A}^{\mathrm{T}}\\ \boldsymbol{a}^{\mathrm{T}}\end{bmatrix}[\boldsymbol{A}\quad \boldsymbol{a}]\boldsymbol{M}^{\mathrm{T}}\boldsymbol{H}^{\mathrm{T}}\right)\\ &= \mathrm{Tr}\left(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{M}\begin{bmatrix}\boldsymbol{A}^{\mathrm{T}}\\ \boldsymbol{a}^{\mathrm{T}}\end{bmatrix}[\boldsymbol{A}\quad \boldsymbol{a}]\boldsymbol{M}^{\mathrm{T}}\right)\\ &= \mathrm{Tr}(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{M}\boldsymbol{X}\boldsymbol{M}^{\mathrm{T}}) = \mathrm{Tr}(\boldsymbol{Q}\boldsymbol{M}\boldsymbol{X}\boldsymbol{M}^{\mathrm{T}})\end{aligned} \tag{15}$$

where $\boldsymbol{Q} \triangleq \boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}$. Let $\boldsymbol{X}^\star \in \mathcal{X}$ be the optimal point for (5) and note that

$$\mathrm{Tr}(\boldsymbol{Q}\boldsymbol{M}\boldsymbol{X}^\star\boldsymbol{M}^{\mathrm{T}}) \leq 4\|\boldsymbol{v}\|^2\,.$$

If this was not true, $\boldsymbol{X}^\star$ would not be optimal due to (14) and (15).

The assumption of the lemma, i.e.,

$$\tau > 4\|\boldsymbol{v}\|^2$$

implies that $\mathrm{Tr}(\boldsymbol{Q}\boldsymbol{M}\boldsymbol{X}^\star\boldsymbol{M}^{\mathrm{T}}) > 4\|\boldsymbol{v}\|^2$ for any $\boldsymbol{X} \in \mathcal{X} \cap \mathcal{H}$. The same conclusion could be drawn for $\boldsymbol{X} \in \mathcal{X}$ for which $\mathrm{Tr}(\boldsymbol{M}\boldsymbol{X}\boldsymbol{M}^{\mathrm{T}}) \geq 1$. This follows due to the linearity of the cost function and since $\mathrm{Tr}(\boldsymbol{M}\boldsymbol{X}_{\boldsymbol{e}}\boldsymbol{M}^{\mathrm{T}}) = 0$. That is, if there were $\boldsymbol{X} \in \mathcal{X}$ for which $\mathrm{Tr}(\boldsymbol{M}\boldsymbol{X}\boldsymbol{M}^{\mathrm{T}}) \geq 1$ and $\mathrm{Tr}(\boldsymbol{Q}\boldsymbol{M}\boldsymbol{X}\boldsymbol{M}^{\mathrm{T}}) \leq 4\|\boldsymbol{v}\|^2$, then $\boldsymbol{X}_\gamma \triangleq \gamma\boldsymbol{X}+(1-\gamma)\boldsymbol{X}_{\boldsymbol{e}} \in \mathcal{X}\cap\mathcal{H}$ for some $\gamma \in (0,1]$ and $\mathrm{Tr}(\boldsymbol{Q}\boldsymbol{M}\boldsymbol{X}_\gamma\boldsymbol{M}^{\mathrm{T}}) \leq 4\|\boldsymbol{v}\|^2$ contrary to the assumption. In short

$$\tau \leq 4\|\boldsymbol{v}\|^2 \Rightarrow \mathrm{Tr}(\boldsymbol{M}\boldsymbol{X}^\star\boldsymbol{M}^{\mathrm{T}}) < 1.$$

Now, partition $\boldsymbol{X}^\star$ as

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

and note that

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

As

$$\mathrm{Tr}(\boldsymbol{M}\boldsymbol{X}^\star\boldsymbol{M}^{\mathrm{T}}) = 2m - 2\boldsymbol{e}^{\mathrm{T}}\boldsymbol{b}$$

since $\mathrm{diag}(\boldsymbol{B}) = \boldsymbol{e}$ due to $\mathrm{diag}(\boldsymbol{X}^\star) = \boldsymbol{e}$ it follows by $\mathrm{Tr}(\boldsymbol{M}\boldsymbol{X}^\star\boldsymbol{M}^{\mathrm{T}}) < 1$ that

$$\boldsymbol{e}^{\mathrm{T}}\boldsymbol{b} \geq m - \frac{1}{2}$$

which implies that all elements of $\boldsymbol{b}$ are in the range of $(\frac{1}{2}, 1]$. Thus, the rounding procedure given in (6) will round the last column of $\boldsymbol{X}^\star$ to $\boldsymbol{e}$ and it follows that $\hat{\boldsymbol{s}}_{\mathrm{SDR}} = \boldsymbol{e}$. □

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 $\boldsymbol{v}$ 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

$$\mathrm{P}\left(\tau \leq \rho^{-1}\right) \,\dot{\leq}\, \rho^{-d} \Rightarrow \mathrm{P}\left(\hat{\boldsymbol{s}}_{\mathrm{SDR}} \neq \boldsymbol{e}\right) \,\dot{\leq}\, \rho^{-d}. \tag{16}$$

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

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

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

$$\begin{aligned} p_e &\leq \mathrm{P}\left(\tau \leq 4\rho^{-1}\|\boldsymbol{w}\|^2\right)\\ &= \int_0^\infty \mathrm{P}\left(\tau \leq 4\rho^{-1}\|\boldsymbol{w}\|^2 \mid \|\boldsymbol{w}\|^2 = \gamma\right) f_{\|\boldsymbol{w}\|^2}(\gamma)d\gamma\\ &= \int_0^\infty \mathrm{P}\left(\tau \leq 4\rho^{-1}\gamma\right) f_{\|\boldsymbol{w}\|^2}(\gamma)d\gamma\\ &\leq c4^{d-\delta}\rho^{-d+\delta}\int_0^\infty \gamma^{d-\delta} f_{\|\boldsymbol{w}\|^2}(\gamma)d\gamma\\ &= c4^{d-\delta}\rho^{-d+\delta}\mathrm{E}\{\|\boldsymbol{w}\|^{2(d-\delta)}\} = c'\rho^{-d+\delta}\end{aligned}$$

for some c' independent of ρ. Note that $c' < \infty$ follows because $\|\boldsymbol{w}\|$ 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. SDR Diversity Proof—Part II

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

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

in the high SNR limit. To this end, it is useful to again consider the definition of τ (11). By noting that

$$\mathrm{Tr}(\boldsymbol{L_0 X}) = \mathrm{Tr}(\boldsymbol{M}^{\mathrm{T}}\boldsymbol{QMX}) = \mathrm{Tr}(\boldsymbol{QMXM}^{\mathrm{T}})$$

if follows that τ is given by

$$\tau = \inf_{\boldsymbol{Y}\in\mathcal{Y}} \mathrm{Tr}(\boldsymbol{QY}) \tag{17}$$

where

$$\mathcal{Y} \triangleq \boldsymbol{M}(\mathcal{X}\cap\mathcal{H})\boldsymbol{M}^{\mathrm{T}}. \tag{18}$$

We can thus equivalently view (17) as our definition of τ. The main reason for doing so is simply that the objective function in (17) has a somewhat simpler form than the one in (11).

Now, note that in order for $\tau \leq \rho^{-1}$, there must be at least one $\boldsymbol{Y}\in\mathcal{Y}$ for which $\mathrm{Tr}(\boldsymbol{QY}) \leq \rho^{-1}$. However, the probability that $\mathrm{Tr}(\boldsymbol{QY}) \leq \rho^{-1}$ for some particular $\boldsymbol{Y}\in\mathcal{Y}$ will generally depend on the specific $\boldsymbol{Y}$ considered (as briefly mentioned in Section III). In order to deal with this, we will first partition $\mathcal{Y}$ into a finite number of subsets $\{\mathcal{Y}_i\}$

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

such that $\mathrm{P}\left(\mathrm{Tr}(\boldsymbol{QY}) \leq \rho^{-1}\right)$ is more or less constant for all $\boldsymbol{Y}$ within one such subset. Then, the probability that $\tau \leq \rho^{-1}$ will be bounded by applying the union bound according to

$$\mathrm{P}\left(\tau \leq \rho^{-1}\right) \leq \sum_i \mathrm{P}\left(\tau_i \leq \rho^{-1}\right) \tag{19}$$

where

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

and where by property (36b) in Appendix A it is known that the sum in (19) will be given (or completely dominated), in the exponential equality sense, 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 [31]. However, in [31], typical events were identified by classifying particularly bad channels $\boldsymbol{H}$, while here, we will use the concept to identify particularly troublesome subsets of $\mathcal{Y}$. In essence, we will partition $\mathcal{Y}$ based on the eigenvalues of $\boldsymbol{Y}\in\mathcal{Y}$ (or how close to singular $\boldsymbol{Y}$ is). The subset which dominates (19) will be found by optimizing over the possible eigenvalue combinations. 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 $\boldsymbol{Y}\in\mathcal{Y}$, i.e., the set given by

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

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

$$\mathrm{Tr}(\boldsymbol{QY}) = \sigma\boldsymbol{u}^{\mathrm{T}}\boldsymbol{Qu}. \tag{20}$$

As $\sigma = 1$ due to the constraint $\mathrm{Tr}(\boldsymbol{Y}) = 1$ it follows by (36d) in Appendix A that

$$\mathrm{P}\left(\mathrm{Tr}(\boldsymbol{QY}) \leq \rho^{-1}\right) = \mathrm{P}\left(\|\boldsymbol{Hu}\|^2 \leq \rho^{-1}\right) \doteq \rho^{-\frac{n}{2}}$$

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

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

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

$$\mathrm{P}\left(\tau_{\mathrm{R1}} \leq \rho^{-1}\right) \doteq \rho^{-\frac{n}{2}}$$

where

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

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) $\hat{\boldsymbol{s}} \in \mathcal{B}^m\backslash\boldsymbol{e}$ that are searched over by the ML detector. This is also the reason why

$$\mathrm{P}\left(\tau_{\mathrm{R1}} \leq \rho^{-1}\right) \doteq \mathrm{P}\left(\hat{\boldsymbol{s}}_{\mathrm{ML}} = \boldsymbol{e}\right).$$

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

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

for some constant $c > 0$, and let

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

As the criterion function $\mathrm{Tr}(\boldsymbol{QY})$ may be bounded according to

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

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

$$\mathrm{P}\left(\tau_{\mathrm{FR}} \leq \rho^{-1}\right) \overset{.}{\leq} \rho^{-\frac{mn}{2}}$$

by applying property (36d) in Appendix A. This result can also be strengthened to show that

$$\mathrm{P}\left(\tau_{\mathrm{FR}} \leq \rho^{-1}\right) \doteq \rho^{-\frac{mn}{2}}.$$

3) Discussion: The implication of the result in Sections V-A1 and V-A2 is that the event that $\tau \leq \rho^{-1}$ 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 $\rho^{-\frac{mn}{2}}$ while the later is only $\rho^{-\frac{n}{2}}$ and $\rho^{-\frac{mn}{2}} \ll \rho^{-\frac{n}{2}}$ when ρ is large (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 remains to be answered however is if there is some other set of $\boldsymbol{Y}$, somewhere between the full rank and rank one matrices, which can cause $\tau \leq \rho^{-1}$ to occur with a probability substantially larger than $\rho^{-\frac{n}{2}}$. The answer to this question is somewhat surprisingly *no* provided that $n \geq m$ (but *maybe* in some $n < m$ cases). In fact, most of the remaining part of this 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

$$\mathcal{Y}(\boldsymbol{a},\boldsymbol{b}) \triangleq \mathcal{Y} \cap \{\boldsymbol{Y} \mid \rho^{-a_k} \leq \sigma_k(\boldsymbol{Y}) \leq \rho^{-b_k}\} \tag{21}$$

where $\boldsymbol{a} = (a_1, \dots, a_m)$, $\boldsymbol{b} = (b_1, \dots, b_m)$, and $\sigma_k(\boldsymbol{Y})$ denotes the kth eigenvalue of $\boldsymbol{Y}$. For notational convenience, in (21), we will also interpret ρ^{-a_k} as 0 for $a_k = \infty$ in order to allow one or more eigenvalues to be identically equal to zero. We can assume without loss of generality that the eigenvalues are ordered and that $0 \leq a_1 \leq \cdots \leq a_m$, $0 = b_1 \leq \cdots \leq b_m$, and $b_k \leq a_k$ for $k = 1, \dots, m$. Note that the assumption that $b_1 = 0$ can be made because (21) would be empty otherwise, due to the $\mathrm{Tr}(\boldsymbol{Y}) = 1$ constraint of $\mathcal{Y}$ in (18). Further, we define the random variable

$$\tau(\boldsymbol{a},\boldsymbol{b}) \triangleq \inf_{\boldsymbol{Y} \in \mathcal{Y}(\boldsymbol{a},\boldsymbol{b})} \mathrm{Tr}(\boldsymbol{QY}). \tag{22}$$

In what follows, a bound on the probability of $\tau(\boldsymbol{a},\boldsymbol{b}) \leq \rho^{-1}$ is obtained by first partitioning $\mathcal{Y}(\boldsymbol{a},\boldsymbol{b})$ into even smaller sets (essentially ϵ-balls) and then using the union bound to bound $\mathrm{P}\left(\tau(\boldsymbol{a},\boldsymbol{b}) \leq \rho^{-1}\right)$. It will be more convenient to work with a square root factorization of $\boldsymbol{Y} \in \mathcal{Y}$ instead of with $\boldsymbol{Y}$ directly. Thus, we define a function

$$\varphi : \mathbb{S}_{+}^{m} \mapsto \mathbb{R}^{m \times m} \tag{23}$$

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

$$\mathcal{A}(\boldsymbol{a},\boldsymbol{b}) \triangleq \varphi(\mathcal{Y}(\boldsymbol{a},\boldsymbol{b})) \tag{24}$$

i.e., $\mathcal{A}(\boldsymbol{a},\boldsymbol{b})$ is the set of square root factors which can be obtained from $\boldsymbol{Y} \in \mathcal{Y}(\boldsymbol{a},\boldsymbol{b})$. Note that $\mathrm{Tr}(\boldsymbol{QY}) = \|\boldsymbol{HA}\|^2$, because $\boldsymbol{Q} = \boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}$ and $\boldsymbol{A} = \varphi(\boldsymbol{Y})$. The random variable $\tau(\boldsymbol{a},\boldsymbol{b})$, defined in (22), can thus be equivalently defined by

$$\tau(\boldsymbol{a},\boldsymbol{b}) = \inf_{\boldsymbol{A} \in \mathcal{A}(\boldsymbol{a},\boldsymbol{b})} \|\boldsymbol{HA}\|^2. \tag{25}$$

We are now ready to provide the first lemma regarding the probability that $\|\boldsymbol{H}\tilde{\boldsymbol{A}}\|^2 \leq \rho^{-1}$ for any $\tilde{\boldsymbol{A}}$ in a spherical neighborhood of some given center point $\boldsymbol{A} \in \mathcal{A}(\boldsymbol{a},\boldsymbol{b})$.

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

$$\mathcal{A}_{\rho}(\boldsymbol{A}) \triangleq \{\tilde{\boldsymbol{A}} \mid \|\tilde{\boldsymbol{A}} - \boldsymbol{A}\| \leq \rho^{-\frac{1}{2}}\}. \tag{26}$$

Further, let

$$\tau(\boldsymbol{A}) \triangleq \inf_{\tilde{\boldsymbol{A}} \in \mathcal{A}_{\rho}(\boldsymbol{A})} \|\boldsymbol{H}\tilde{\boldsymbol{A}}\|^2. \tag{27}$$

Then

$$\mathrm{P}\left(\tau(\boldsymbol{A}) \leq \rho^{-1}\right) \stackrel{.}{\leq} \rho^{-\nu} \quad \text{where} \quad \nu \triangleq \sum_{k=1}^{m} \frac{n(1-a_k)^{+}}{2}.$$

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

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

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

$$\|\boldsymbol{H}\| \leq \rho^{\delta} \tag{28}$$

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

$$\|\boldsymbol{h}_k\| \geq 2\rho^{-\frac{1-a_k}{2}+\delta}. \tag{29}$$

First, we will show that this event implies that $\tau(\boldsymbol{A}) > \rho^{-1}$ and next that the event fails to occur with a probability which is not larger (in the $\stackrel{.}{\leq}$ sense) than $\rho^{-\nu+nm\delta}$. Hence

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

Note first that (29) implies

$$\left\|\boldsymbol{h}_k \sigma_k^{\frac{1}{2}}\right\| \geq 2\rho^{-\frac{1}{2}+\delta}$$

for at least one k because $\sigma_k \geq \rho^{-a_k}$. Note also that this implies

$$\|\boldsymbol{HA}\| = \|\boldsymbol{H}\boldsymbol{\Sigma}^{\frac{1}{2}}\| \geq 2\rho^{-\frac{1}{2}+\delta}.$$

Now, consider $\|\boldsymbol{H}\tilde{\boldsymbol{A}}\|$ for any $\tilde{\boldsymbol{A}}$ satisfying $\|\tilde{\boldsymbol{A}} - \boldsymbol{A}\| \leq \rho^{-\frac{1}{2}}$. Under the additional assumption of (28), it follows that

$$\begin{aligned}
\|\boldsymbol{H}\tilde{\boldsymbol{A}}\| &= \|\boldsymbol{HA} - \boldsymbol{H}(\boldsymbol{A} - \tilde{\boldsymbol{A}})\| \\
&\geq \|\boldsymbol{HA}\| - \|\boldsymbol{H}(\boldsymbol{A} - \tilde{\boldsymbol{A}})\| \\
&\geq 2\rho^{-\frac{1}{2}+\delta} - \rho^{-\frac{1}{2}+\delta} \\
&= \rho^{-\frac{1}{2}+\delta} > \rho^{-\frac{1}{2}}
\end{aligned}$$

where the last inequality holds whenever $\rho \geq 1$. Note also that $\|\boldsymbol{H}\tilde{\boldsymbol{A}}\| > \rho^{-\frac{1}{2}}$ implies $\|\boldsymbol{H}\tilde{\boldsymbol{A}}\|^2 > \rho^{-1}$. Therefore, (28) and (29) implies that $\tau(\boldsymbol{A}) > \rho^{-1}$.

Now, consider the probability that (29) fails to hold, e.g., that

$$\|\boldsymbol{h}_k\| < 2\rho^{-\frac{1-a_k}{2}+\delta}$$

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

$$\mathrm{P}\Big(\|\boldsymbol{h}_k\| < 2\rho^{-\frac{1-a_k}{2}+\delta}\ \forall\, k\Big) = \prod_{k=1}^{m} \mathrm{P}\left(\|\boldsymbol{h}_k\| < 2\rho^{-\frac{1-a_i}{2}+\delta}\right) \dot{\leq} \prod_{k=1}^{m} \rho^{-\frac{n(1-a_k-2\delta)^+}{2}} \leq \rho^{-\nu+nm\delta}$$

where we have used

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

according to (36d) in Appendix A. The probability that (28) fails to hold can be upper bounded as

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

according to (36e) in Appendix A. Therefore, by applying the union bound

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

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

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

which concludes the proof. □

The next lemma provides a bound on the number of such $\rho^{-\frac{1}{2}}$-balls [defined as in (26)], which are required to completely cover the set $\mathcal{A}(\boldsymbol{a}, \boldsymbol{b})$. Lemma 4 is the technically most difficult result of this work and we discuss this lemma in the following, but save the stringent proof for Appendix B.

Lemma 4: Let $\mathcal{A}(\boldsymbol{a}, \boldsymbol{b})$ and $\mathcal{A}_\rho(\boldsymbol{A})$ be defined as in (24) and (26), respectively. Then, there is a collection of points $\mathfrak{A} = \{\boldsymbol{A}_i\}$, for which

$$\mathcal{A}(\boldsymbol{a}, \boldsymbol{b}) \subset \bigcup_{\boldsymbol{A}_i \in \mathfrak{A}} \mathcal{A}_\rho(\boldsymbol{A}_i)$$

and

$$|\mathfrak{A}| \ \dot{\leq}\ \rho^{\mu}$$

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

$$\mu \triangleq \sum_{k=2}^{m} \frac{(m-k+2)(1-b_k)^+}{2}. \tag{30}$$

Proof: Given in Appendix B. □

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 $\boldsymbol{A} \in \mathcal{A}$, i.e.,

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

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

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

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 [33] and not specific to $\mathcal{A}_{\mathrm{R}r}$). The volume V covered by one d_r-dimensional $\rho^{-\frac{1}{2}}$-ball is on the order of

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

and, therefore, one needs on the order of

$$N \doteq \frac{1}{V} \doteq \rho^{\frac{d_r}{2}} \tag{31}$$

such $\rho^{-\frac{1}{2}}$-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 $\mathcal{A}_{\mathrm{R}r}$.

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

$$|\mathfrak{A}_r| \ \dot{\leq}\ \rho^{\mu_r}$$

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 $\boldsymbol{A}$ are bounded by powers of ρ^{-1} yields the result stated in Lemma 4. Note also that this is similar to the discussion following Theorem 4 in [31].

Now, Lemmas 3 and 4 can be combined in order to bound the probability that $\mathcal{A}(\boldsymbol{a}, \boldsymbol{b})$ contains an $\boldsymbol{A}$ for which $\|\boldsymbol{H}\boldsymbol{A}\|^2 \leq \rho^{-1}$. Then, by optimizing over $\boldsymbol{a}$ and $\boldsymbol{b}$, one can find the set of the form of $\mathcal{A}(\boldsymbol{a}, \boldsymbol{b})$ most likely to contain such an $\boldsymbol{A}$. 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

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

where

$$\zeta \triangleq \inf_{1 \geq c_2 \geq \cdots \geq c_m \geq 0} \frac{n}{2} + \sum_{k=2}^{m} \frac{(n-m+k-2)c_k}{2}. \tag{32}$$

Proof: Consider picking some $\boldsymbol{b} = (b_1, \ldots, b_m)$ for which $b_1 = 0$ and $b_1 \leq b_2 \leq \cdots \leq b_m \leq 1$ and choose a $\delta > 0$. Let $\boldsymbol{a} = (a_1, \ldots, a_m)$ 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 $\tau(\boldsymbol{a},\boldsymbol{b}) \le \rho^{-1}$ where $\tau(\boldsymbol{a},\boldsymbol{b})$ is defined in (22) can be bounded, using the union bound according as

$$\mathrm{P}\left(\tau(\boldsymbol{a},\boldsymbol{b}) \le \rho^{-1}\right) \le \sum_{\boldsymbol{A}_i \in \mathfrak{A}} \mathrm{P}\left(\tau(\boldsymbol{A}_i) \le \rho^{-1}\right)$$

where $\mathfrak{A}$ is chosen according to Lemma 4 and where $\tau(\boldsymbol{A}_i)$ is given by (27). Each term in the sum is upper bounded by

$$\mathrm{P}\left(\tau(\boldsymbol{A}_i) \le \rho^{-1}\right) \;\dot{\le}\; \rho^{-\nu}$$

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

$$|\mathfrak{A}| \;\dot{\le}\; \rho^{\mu}$$

where μ is given by (30). Thus, the probability that $\tau(\boldsymbol{a},\boldsymbol{b}) \le \rho^{-1}$ is bounded as

$$\mathrm{P}\left(\tau(\boldsymbol{a},\boldsymbol{b}) \le \rho^{-1}\right) \;\dot{\le}\; \rho^{-(\nu-\mu)}$$

where

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

and where the property

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

(for a_k chosen as previously) was used to establish the first inequality. The second inequality follows by the definition of ζ in (32) along with $b_k \ge 0$.

Now, let

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

where φ is given by (23). Note that we can pick a finite set of $\boldsymbol{b} \in [0,1]^m$, $\mathfrak{B} = \{\boldsymbol{b}_i\}$, such that

$$\mathcal{A} \subset \bigcup_{\boldsymbol{b} \in \mathfrak{B}} \mathcal{A}(\boldsymbol{a},\boldsymbol{b}) \tag{33}$$

where $\boldsymbol{a} = \boldsymbol{a}(\boldsymbol{b})$ according to the aforementioned. This follows, because by specifying $\boldsymbol{b} = (b_1,\ldots,b_m)$, we include the matrices $\boldsymbol{Y} \in \mathcal{Y}$ for which the kth eigenvalue satisfies $\rho^{-b_k-\delta} \le \sigma_k \le \rho^{-b_k}$ if $b_k < 1$ and $\sigma_k \le \rho_{-1}$ 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 $\mathrm{Tr}(\boldsymbol{Y}) = 1$, which implies that $\sigma_1 \in [\rho^{-\delta}, 1]$ for sufficiently large ρ given that $\delta > 0$, which is why $b_1 = 0$ can be assumed without loss of generality.

Using the union bound, it follows that

$$\begin{aligned}\mathrm{P}\left(\tau \le \rho^{-1}\right) &\le \sum_{\boldsymbol{b} \in \mathfrak{B}} \mathrm{P}\left(\tau(\boldsymbol{a},\boldsymbol{b}) \le \rho^{-1}\right) \\ &\dot{\le} \rho^{-\zeta + \frac{mn\delta}{2}}\end{aligned}$$

because each term in the sum satisfies

$$\mathrm{P}\left(\tau(\boldsymbol{a},\boldsymbol{b}) \le \rho^{-1}\right) \;\dot{\le}\; \rho^{-\zeta + \frac{mn\delta}{2}}$$

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

$$\mathrm{P}\left(\tau(\boldsymbol{a},\boldsymbol{b}) \le \rho^{-1}\right) \;\dot{\le}\; \rho^{-\zeta}$$

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 (32) and apply Lemma 2. We give the proof as follows.

Proof (of Theorem 1): For the case where $n \ge m$, all terms in the sum appearing in (32) are nonnegative. Thus, the minimum in (32) is achieved for $c_2 = \cdots = c_m = 0$ and it follows that

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

This, combined with Lemma 2, proves that

$$\mathrm{P}\left(\hat{\boldsymbol{s}}_{\mathrm{SDR}} \ne \boldsymbol{e}\right) \;\dot{\le}\; \rho^{-\frac{n}{2}}.$$

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

$$\mathrm{P}\left(\hat{\boldsymbol{s}}_{\mathrm{SDR}} \ne \boldsymbol{e}\right) \ge \mathrm{P}\left(\hat{\boldsymbol{s}}_{\mathrm{ML}} \ne \boldsymbol{e}\right) \doteq \rho^{-\frac{n}{2}}$$

because the ML detector achieves the minimum probability of error. Therefore, it follows that

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

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

VI. Examples and Extensions

We conclude by providing numerical examples illustrating the results obtained and discuss possible extensions and future work.

A. Numerical Example

The overall performance of the SDR detector is illustrated in Fig. 2 for the case when $n = m = 4$. For all the examples in this section, the variances of the elements in $\boldsymbol{H}$ are chosen to be n^{-1}, yielding unit energy symbols at the receiver. The performance of the ML detector, the LMMSE detector, and a version of the SDR detector with randomized rounding (denoted SDRR) are also included for comparison. In SDRR, the final estimate of $\boldsymbol{s}$ is obtained by, in addition to the estimate already obtained, adding $2m$ random candidates generated according to the procedure outlined in [8] and choosing the one with the smallest ML metric as the final estimate. The probability that (5) does not have a rank one solution is indicated by the dashed line.

As predicted by Theorem 1, it can be seen that the SDR detector achieves the same diversity order as the ML detector, a

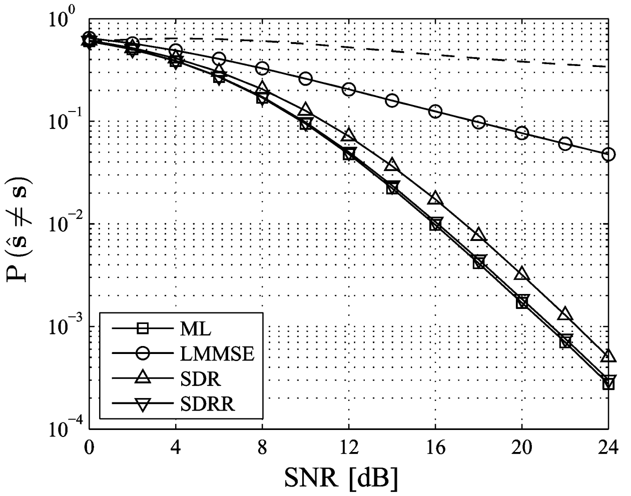

Fig. 2. Probability of error when $\boldsymbol{H} \in \mathbb{R}^{n \times m}$ has i.i.d. real-valued Gaussian entries, and where $m = n = 4$. The dashed line correspond to $\mathrm{P}\left(\mathrm{Rank}(\boldsymbol{X}^{\star}) \neq 1\right)$, where $\boldsymbol{X}^{\star}$ is the optimizer of (5).

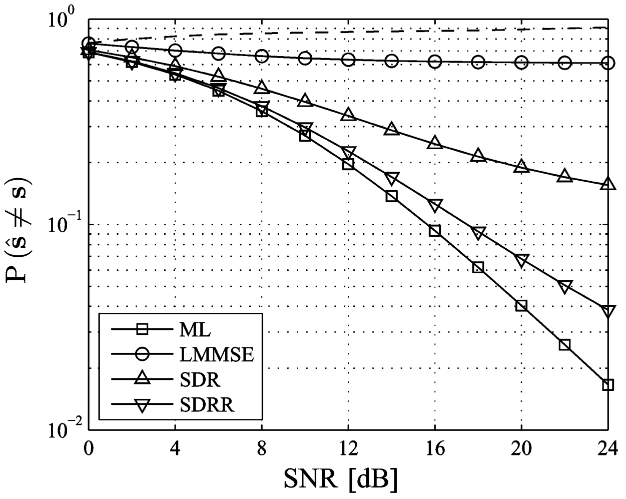

Fig. 3. Probability of error when $\boldsymbol{H} \in \mathbb{R}^{n \times m}$ has i.i.d. real-valued Gaussian entries, and where $m = 4$ and $n = 2$. The dashed line correspond to $\mathrm{P}\left(\mathrm{Rank}(\boldsymbol{X}^{\star}) \neq 1\right)$, where $\boldsymbol{X}^{\star}$ is the optimizer of (5).

property not shared by the simpler LMMSE detector. In addition, it can be seen that the SDRR detector provides a significant improvement over the somewhat simpler SDR detector considered herein. Also, as mentioned in Section II-B, we see that the rank one solutions alone are not sufficient for explaining the SDR performance.

B. The $n < m$ Case

By Theorem 1, full diversity has been established so far under the condition that $n \geq m$. However, a careful inspection of the proofs shows that the only part which explicitly relies on this assumption is when it is argued that $c_2 = \cdots = c_m = 0$ is an optimal point for (32) in the $n \geq m$ case. However, nontrivial bounds on the diversity will follow whenever ζ in (32) is strictly positive. To exemplify this, 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

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

where

$$d = \frac{1}{2}\left(m - \frac{r(r+3)}{2}\right). \tag{34}$$

Proof: The result is established by finding the optimum in (32) and applying Lemma 2. To this end, note that the optimum of (32) 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 (32) 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. □

It should be noted, however, that this result is only nontrivial if

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

In addition, 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 nontrivial. At the same time, we do not expect the bound to be very loose in the sense that the SDR detector would maintain ML diversity in the general $n < m$ case. The latter belief is supported by Fig. 3, where the error probability of the SDR is significantly larger than that of the ML detector. Also, in this case, the situation is improved by the SDRR implementation although there is still a significant gap to ML.

C. Complex Channel Matrices

Throughout this work, we have also assumed that the channel matrix is real valued. It is well known, however, that the SDR receiver is also applicable to the case where 4-quadratic-amplitude modulation (4-QAM) symbols are transmitted over a complex-valued MIMO channel; see, e.g., [7]. The most direct strategy is to rewrite the problem in an equivalent real-valued form according to

$$\begin{bmatrix} \Re(\boldsymbol{y}_c) \\ \Im(\boldsymbol{y}_c) \end{bmatrix} = \begin{bmatrix} \Re(\boldsymbol{H}_c) & -\Im(\boldsymbol{H}_c) \\ \Im(\boldsymbol{H}_c) & \Re(\boldsymbol{H}_c) \end{bmatrix} \begin{bmatrix} \Re(\boldsymbol{s}_c) \\ \Im(\boldsymbol{s}_c) \end{bmatrix} + \begin{bmatrix} \Re(\boldsymbol{v}_c) \\ \Im(\boldsymbol{v}_c) \end{bmatrix} \tag{35}$$

where $\boldsymbol{y}_c \in \mathbb{C}^N$, $\boldsymbol{H}_c \in \mathbb{C}^{N \times M}$, $\boldsymbol{s}_c \in \mathbb{C}^M$, and $\boldsymbol{v}_c \in \mathbb{C}^N$ are the corresponding complex-valued quantities and where $\Re(\cdot)$ and

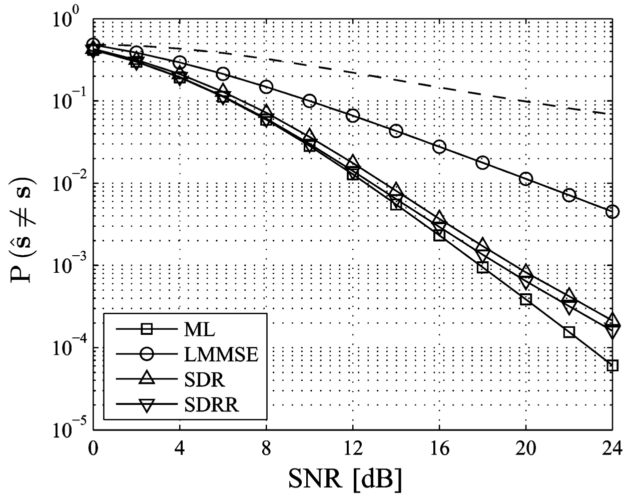

Fig. 4. Probability of error when $\boldsymbol{H}_c \in \mathbb{C}^{N\times M}$ has i.i.d. complex-valued Gaussian entries, and where $N = M = 2$. The dashed line correspond to $\mathrm{P}\,(\mathrm{Rank}(\boldsymbol{X}^\star) \neq 1)$, where $\boldsymbol{X}^\star$ is the optimizer of (5).

$\Im(\cdot)$ denote the real and imaginary parts. However, the proof of Theorem 1 does not extend to cover this case. The specific reason is found in Lemma 3, where the rotational symmetry of $\boldsymbol{H}$ is explicitly used. This symmetry is lost in the formulation given in (35), even in the case where $\boldsymbol{H}_c$ is i.i.d. complex, 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. 4, where the basic SDR detector considered herein as well as the SDRR detector appears to experience a small loss in diversity.

Assuming that there indeed is a loss in diversity in the complex case, an interesting topic for future work would be to investigate whether strengthening the relaxation as suggested in [7] or [13] could increase the diversity order. Numerical results in [13] suggest that this may be possible and that (35) may not be the optimal way of dealing with the complex case. It may also be that the suggested loss in diversity is in fact due to the simple rounding procedure used to obtain estimates of $\boldsymbol{s}$ from the solution of (5). However, at this stage,it is not clear if these questions could be fully answered using the analytic tools developed herein.

VII. Conclusion

In this paper, we have shown that when applied to a fading channel, modeled by a real-valued matrix with i.i.d. Gaussian entries of zero mean and finite variance, the SDR detector achieves the maximum possible diversity. This provides a strong performance guarantee for the SDR approach, when applied in the communications context.

Appendix A
Exponential Equality

For the readers' convenience, we list the most important properties associated with the definition of *exponential equality* in (7) (for this work). These properties are easily derived from the definition in (7) and can also be found (often implicitly) in many texts; see, e.g., [1] and [31]. Thus, we state the properties without proof.

1) *Scaling property:* For any $a \in [-\infty,\infty]$ and $c \in (-\infty,\infty)$, it holds that
$$f(\rho) \doteq \rho^{-a} \Rightarrow cf(\rho) \doteq \rho^{-a}. \tag{36a}$$
2) *Summation property:* For any $a, b \in [-\infty,\infty]$, it holds that
$$f(\rho) \doteq \rho^{-a},\ g(\rho) \doteq \rho^{-b} \Rightarrow f(\rho) + g(\rho) \doteq \rho^{-\min(a,b)}. \tag{36b}$$
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
$$f(\rho) \doteq \rho^{-a},\ g(\rho) \doteq \rho^{-b} \Rightarrow f(\rho)g(\rho) \doteq \rho^{-(a+b)} \tag{36c}$$
if the cases where $a + b$ is not well defined are excluded.
4) *Extremal realizations of Gaussian vectors:* Let $\boldsymbol{h} \in \mathbb{R}^d$ be a vector of i.i.d. Gaussian elements of finite nonzero variance. Then
$$\mathrm{P}\left(\|\boldsymbol{h}\|^2 \leq \rho^{-c}\right) \doteq \rho^{-\frac{dc^+}{2}} \tag{36d}$$
for $c \in (-\infty,\infty)$, where $c^+ \triangleq \max(c,0)$ and
$$\mathrm{P}\left(\|\boldsymbol{h}\|^2 \geq \rho^{c}\right) \doteq \rho^{-\infty} \tag{36e}$$
for $c > 0$. These properties follow by noting that $\|\boldsymbol{h}\|^2$ is χ^2 distributed with d degrees of freedom; see, e.g., [1, Sec. 5.4.2].

It should also be noted that the properties given in (36a)–(36c) also hold with $\dot{\leq}$ or $\dot{\geq}$ in place of $\doteq$.

Appendix B
Proof of Lemma 4

Before proving Lemma 4 we establish the following technical result regarding the feasible set of (17).

Lemma 6: The set $\mathcal{Y}$ defined in (18) satisfies
$$\mathcal{Y} = \{\boldsymbol{Y} \in \mathbb{S}^m \mid \mathrm{Tr}(\boldsymbol{Y}) = 1, \boldsymbol{Y} \succeq \frac{1}{4}\boldsymbol{d}\boldsymbol{d}^{\mathrm{T}}, \boldsymbol{d} = \mathrm{diag}(\boldsymbol{Y})\}. \tag{37}$$

Proof: Consider the transformation given by
$$\underbrace{\begin{bmatrix} \boldsymbol{Y} & \boldsymbol{a} \\ \boldsymbol{a}^{\mathrm{T}} & c \end{bmatrix}}_{\boldsymbol{P}} = \underbrace{\begin{bmatrix} \boldsymbol{I} & -\boldsymbol{e} \\ \boldsymbol{0}^{\mathrm{T}} & 1 \end{bmatrix}}_{\boldsymbol{T}} \boldsymbol{X} \begin{bmatrix} \boldsymbol{I} & 0 \\ -\boldsymbol{e}^{\mathrm{T}} & 1 \end{bmatrix} \tag{38}$$
or inversely
$$\boldsymbol{X} = \underbrace{\begin{bmatrix} \boldsymbol{I} & \boldsymbol{e} \\ \boldsymbol{0}^{\mathrm{T}} & 1 \end{bmatrix}}_{\boldsymbol{R}} \begin{bmatrix} \boldsymbol{Y} & \boldsymbol{a} \\ \boldsymbol{a}^{\mathrm{T}} & c \end{bmatrix} \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{e}^{\mathrm{T}} & 1 \end{bmatrix} \tag{39}$$
since $\boldsymbol{T}^{-1} = \boldsymbol{R}$. Note also that $\boldsymbol{Y}$ is given by $\boldsymbol{Y} = \boldsymbol{M}\boldsymbol{X}\boldsymbol{M}^{\mathrm{T}}$ as $\boldsymbol{M} = [\boldsymbol{I} \quad -\boldsymbol{e}]$ by (10). Because $\mathrm{Tr}(\boldsymbol{M}\boldsymbol{X}\boldsymbol{M}^{\mathrm{T}}) = 1$ for $\boldsymbol{X} \in \mathcal{H}$, it follows that $\mathrm{Tr}(\boldsymbol{Y}) = 1$ for $\boldsymbol{Y} \in \mathcal{Y}$.

Completing the matrix multiplication in (39) yields

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

Thus, the constraint $\mathrm{diag}(\boldsymbol{X}) = \boldsymbol{e}$ for $\boldsymbol{X} \in \mathcal{X}$ implies that $c = 1$ for $\boldsymbol{Y} \in \mathcal{Y}$. Further, using $c = 1$

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

which implies that

$$\boldsymbol{a} = -\frac{1}{2}\mathrm{diag}(\boldsymbol{Y}). \tag{40}$$

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

Because $\boldsymbol{T}$ (and $\boldsymbol{R}$) is invertible, the constraint $\boldsymbol{X} \succeq \boldsymbol{0}$ is equivalent to $\boldsymbol{P} \succeq \boldsymbol{0}$. However, $\boldsymbol{P} \succeq \boldsymbol{0}$ if and only if its Schur complement [23] is positive semidefinite, i.e., if

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

Thus, by combining (40) with $c = 1$ and identifying $\boldsymbol{d} = -2\boldsymbol{a}$, it is established that the set $\mathcal{Y} = \boldsymbol{M}(\mathcal{X} \cap \mathcal{H})\boldsymbol{M}^{\mathrm{T}}$, originally defined in (18), is equivalently given by (37). □

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

Lemma 4: Let $\mathcal{A}(\boldsymbol{a}, \boldsymbol{b})$ and $\mathcal{A}_\rho(\boldsymbol{A})$ be defined as in (24) and (26), respectively. Then, there is a collection of points $\mathfrak{A} = \{\boldsymbol{A}_i\}$, for which

$$\mathcal{A}(\boldsymbol{a}, \boldsymbol{b}) \subset \bigcup_{\boldsymbol{A}_i \in \mathfrak{A}} \mathcal{A}_\rho(\boldsymbol{A}_i)$$

and

$$|\mathfrak{A}| \ \dot{\leq} \ \rho^{\mu}$$

where

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

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

$$\mathrm{Tr}(\boldsymbol{\Lambda}^2) = 1 \tag{41a}$$

$$\mathrm{diag}(\boldsymbol{U}\boldsymbol{\Lambda}^2\boldsymbol{U}^{\mathrm{T}}) = \boldsymbol{U}\boldsymbol{z} \tag{41b}$$

$$\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \boldsymbol{I} \tag{41c}$$

$$\boldsymbol{\Lambda}^2 - \frac{1}{4}\boldsymbol{z}\boldsymbol{z}^{\mathrm{T}} \succeq \boldsymbol{0} \tag{41d}$$

where $\boldsymbol{\Lambda} \triangleq \mathrm{diag}(\boldsymbol{\lambda})$. In what follows, the set of solutions to (41) will be denoted by $\mathcal{M}$. The set of solutions to (41a)–(41c) but not necessarily (41d) is denoted by $\mathcal{N}$ and it follows that $\mathcal{M} \subset \mathcal{N}$. From (41a) and (41c) it follows that $\boldsymbol{\lambda}$ and $\boldsymbol{U}$ in the solution set are bounded. However, as $\boldsymbol{U}$ is full rank due to (41c) it follows through (41b) that $\boldsymbol{z}$ is also bounded. Therefore, both $\mathcal{N}$ and $\mathcal{M}$ are compact (closed and bounded) sets.

The constraints of (41) are such that any solution $(\boldsymbol{U}, \boldsymbol{\lambda}, \boldsymbol{z})$ of (41) satisfies $\boldsymbol{U}\boldsymbol{\Lambda}^2\boldsymbol{U}^{\mathrm{T}} \in \mathcal{Y}$ and any eigenvalue decomposition $\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^{\mathrm{T}}$ of $\boldsymbol{Y} \in \mathcal{Y}$ solves (41) for $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{\frac{1}{2}}$ and some (unique) $\boldsymbol{z}$. To see this, consider the eigenvalue decomposition $\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^{\mathrm{T}}$ of some $\boldsymbol{Y} \in \mathcal{Y}$, where $\mathcal{Y}$ is given by (18). Note also that $\boldsymbol{Y}$ belongs to $\mathcal{Y}$ if and only if it satisfies the constraints of (37) as proven in Lemma 6. The orthogonality of $\boldsymbol{U} \in \mathbb{R}^{m \times m}$ is a property of the eigenvalue decomposition, and therefore, (41c) is satisfied. For $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{\frac{1}{2}}$ and $\boldsymbol{z} = \boldsymbol{U}^{\mathrm{T}}\mathrm{diag}(\boldsymbol{U}\boldsymbol{\Lambda}^2\boldsymbol{U}^{\mathrm{T}})$, the constraint of (41b) is satisfied. As $\boldsymbol{Y} \in \mathcal{Y}$, it follows that $\boldsymbol{Y} - \frac{1}{4}\boldsymbol{d}\boldsymbol{d}^{\mathrm{T}} \succeq \boldsymbol{0}$, where $\boldsymbol{d} = \mathrm{diag}(\boldsymbol{Y})$. Therefore, $\mathrm{diag}(\boldsymbol{Y}) = \mathrm{diag}(\boldsymbol{U}\boldsymbol{\Lambda}^2\boldsymbol{U}^{\mathrm{T}}) = \boldsymbol{U}\boldsymbol{z}$ implies

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

which means that (41d) is satisfied. Finally, the constraint $\mathrm{Tr}(\boldsymbol{Y}) = 1$ in (37) implies that $\mathrm{Tr}(\boldsymbol{\Lambda}^2) = 1$ and (41a) is satisfied. Reversing the reasoning and applying Lemma 6 show that any solution to (41) must also have the property that $\boldsymbol{U}\boldsymbol{\Lambda}^2\boldsymbol{U}^{\mathrm{T}} \in \mathcal{Y}$.

The value of introducing (41) is that it will provide, through the implicit function theorem [34], a means of parameterizing the eigenvalues and vectors of $\boldsymbol{Y} \in \mathcal{Y}$. 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

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

Define

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

according to

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

and note that $H(\boldsymbol{\omega}) = \boldsymbol{0}$ corresponds to (41a)–(41c). In the above, $\mathrm{svec}(\cdot)$ refers to the vector obtained by stacking the upper triangular part of a symmetric matrix into a vector. Let

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

be a solution of (41) and $\mathcal{I}$ be an index set satisfying

$$\mathcal{I} \subset \{1, \ldots, q\} \tag{42}$$

and

$$|\mathcal{I}| = p. \tag{43}$$

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 [34] states that if

$$\left| \frac{\partial H(\boldsymbol{\omega})}{\partial \boldsymbol{\omega}_{\mathcal{I}}} \right|_{\boldsymbol{\omega}=\bar{\boldsymbol{\omega}}} \neq 0 \tag{44}$$

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, (44) 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 (44) is satisfied, the solution set of (41) is locally parameterized by $q-p$ scalar parameters. In fact, it will be shown later that given *any* solution $\bar{\boldsymbol{\omega}}$ to (41) there will be some index set $\mathcal{I}$ satisfying (42) and (43) for which (44) is satisfied. This implies that $\mathcal{N}$ is a $q-p$-dimensional (smooth) manifold embedded in $\mathbb{R}^q$ [35]. Note, however, that the specific index set $\mathcal{I}$ required to satisfy (44) 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 be assumed without loss of generality that the domain of ψ is given by

$$\mathcal{D} = (-\kappa, \kappa)^{q-p} \tag{45}$$

i.e., that $\mathcal{D}$ is an open hypercube for some $\kappa > 0$ [35]. Further, because $\mathcal{N}$ is compact, it can be assumed that κ is independent of $\bar{\boldsymbol{\omega}}$. It can also be assumed, without loss of generality, that ψ is Lipschitz continuous [36] 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 (44) is satisfied, it is sufficient to prove that the Jacobian matrix $\boldsymbol{D}$

$$\boldsymbol{D} \triangleq \left. \frac{\partial H(\boldsymbol{\omega})}{\partial \boldsymbol{\omega}} \right|_{\boldsymbol{\omega}=\bar{\boldsymbol{\omega}}} \in \mathbb{R}^{p \times q} \tag{46}$$

is full rank. In this event, the index set $\mathcal{I}$ can be taken as the indexes of any p linearly independent columns of $\boldsymbol{D}$. For our purposes, however, we will 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 $\boldsymbol{Y} \in \mathcal{Y}$ (see the discussion in Section V-A3). To this end, consider some $\bar{\boldsymbol{\omega}} \in \mathcal{M}$, for which $\lambda_{r+1} = \cdots = \lambda_m = 0$, i.e., $\bar{\boldsymbol{\omega}}$ corresponds to a rank r matrix $\bar{\boldsymbol{Y}} \in \mathcal{Y}$. Here, and in what follows, λ_k and z_k refer to the kth component of $\boldsymbol{\lambda}$ and $\boldsymbol{z}$, respectively. For any $\bar{\boldsymbol{\omega}} \in \mathcal{M}$, it follows by (41d) that $|z_k| \leq 2|\lambda_k|$ for $k=1,\ldots,m$, and in particular, it follows that $z_k = 0$ whenever $\lambda_k = 0$. In what follows, we will refer to any $\bar{\boldsymbol{\omega}} \in \mathcal{N}$, which satisfies both $\lambda_{r+1} = \cdots = \lambda_m = 0$ and $z_{r+1} = \cdots = z_m = 0$ 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 (41d) 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 $\boldsymbol{u}_k$ denote the kth column of $\boldsymbol{U}$. In what follows, it will 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 $\boldsymbol{u}_k$ for $k = r+1, \ldots, m$, and a subset of $q_r - p_r$ parameters from

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

It is straightforward to verify that this amounts to a total of $q-p$ parameters. The specific parameters chosen from $\boldsymbol{u}_k$ 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

$$\mathrm{Tr}(\boldsymbol{\Lambda}_r) + \eta = 1 \tag{47a}$$

$$\mathrm{diag}(\boldsymbol{U}_r \boldsymbol{\Lambda}_r \boldsymbol{U}_r) + \boldsymbol{\gamma} = \boldsymbol{U}_r \boldsymbol{z}_r \tag{47b}$$

$$\boldsymbol{U}_r^{\mathrm{T}} \boldsymbol{U}_r = \boldsymbol{I} \tag{47c}$$

where $(\boldsymbol{U}_r, \boldsymbol{\lambda}_r, \boldsymbol{z}_r, \boldsymbol{\gamma}, \eta) \in \mathbb{R}^{m \times r} \times \mathbb{R}^r \times \mathbb{R}^r \times \mathbb{R}^m \times \mathbb{R}^1$ 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 (47). These will be used later to develop a perturbation analysis of the solutions to (41) around the rank r points.

Let

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

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

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

according to

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

and note that $H_r(\boldsymbol{\omega}_r, \boldsymbol{\gamma}, \eta) = \mathbf{0}$ is equivalent to (47). In order to establish that the solution set of (47) 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

$$\boldsymbol{D}_r = \left. \frac{\partial H_r(\bar{\boldsymbol{\omega}}_r)}{\partial \bar{\boldsymbol{\omega}}_r} \right|_{\boldsymbol{\omega} = \bar{\boldsymbol{\omega}}} \in \mathbb{R}^{p_r \times q_r} \tag{48}$$

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 as before, if $\boldsymbol{D}_r$ in (48) is full rank then this implies the existence of a Lipschitz continuous function

$$\psi_r : \mathcal{D}_r \mapsto \mathcal{R}_r \tag{49}$$

where $(\boldsymbol{U}_r, \boldsymbol{\lambda}_r, \boldsymbol{z}_r) = \psi_r(\boldsymbol{\xi}_r, \boldsymbol{\gamma}, \eta)$ 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 $\boldsymbol{D}_r$ consider the matrix

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

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

$$\boldsymbol{U}_r = [\, \boldsymbol{u}_1 \quad \cdots \quad \boldsymbol{u}_r \,] = [\, \boldsymbol{g}_1 \quad \cdots \quad \boldsymbol{g}_m \,]^{\mathrm{T}}.$$

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

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

where

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

and where $\bar{\boldsymbol{g}}_i^2$ denotes elementwise squaring of $\bar{\boldsymbol{g}}_i$. Assume first that $2\bar{\boldsymbol{g}}_i^{\mathrm{T}} \bar{\boldsymbol{\Lambda}}_r^2 - \bar{\boldsymbol{z}}_r^{\mathrm{T}} = \mathbf{0}$ for some i, $1 \leq i \leq m$. This implies through (47b) (and $\boldsymbol{\gamma} = \mathbf{0}$) that

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

and in turn $\bar{\boldsymbol{\Lambda}}_r^2 \bar{\boldsymbol{g}}_i = \mathbf{0}$ as $\bar{\boldsymbol{\Lambda}}_r^2 \succeq \mathbf{0}$. Further, it follows that $\bar{\boldsymbol{z}}_r = \mathbf{0}$ and that $\bar{\boldsymbol{\Lambda}}_r = \mathbf{0}$ by inserting $\bar{\boldsymbol{z}}_r = \mathbf{0}$ into (47b). However, this violates (47a) and contradicts that $\bar{\boldsymbol{\omega}}_r$ is a solution to (47). Thus, it can be assumed that $2\bar{\boldsymbol{g}}_i^{\mathrm{T}} \bar{\boldsymbol{\Lambda}}_r^2 - \bar{\boldsymbol{z}}_r^{\mathrm{T}} \neq \mathbf{0}$ for all $i = 1, \ldots, m$ which implies that the first $m + 1$ rows of $\tilde{\boldsymbol{D}}_r$ are linearly independent.

Establishing that the last $r(r + 1)/2$ rows of $\tilde{\boldsymbol{D}}_r$ 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 [35]. For this reason, we will not provide an explicit proof of this. In fact, the last $r(r+1)/2$ rows of $\tilde{\boldsymbol{D}}_r$ are not only linearly independent but also orthogonal. What now remains to be done, in order to show that $\tilde{\boldsymbol{D}}_r$ 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 $\tilde{\boldsymbol{D}}_r$ together with $\bar{\boldsymbol{\lambda}}_r \neq \mathbf{0}$. For the next m rows, the only potential problem would be if $\boldsymbol{g}_i = \mathbf{0}$ for some i. However, as

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

it follows that $\bar{\boldsymbol{G}}_i$ is linear in $\bar{\boldsymbol{g}}_i$ and equal to zero whenever $\bar{\boldsymbol{g}}_i = \mathbf{0}$. Together with the property that $2\bar{\boldsymbol{g}}_i^{\mathrm{T}} \bar{\boldsymbol{\Lambda}}_r^2 - \bar{\boldsymbol{z}}_r^{\mathrm{T}} \neq \mathbf{0}$, 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 $\tilde{\boldsymbol{D}}_r$ and $\boldsymbol{D}_r$ are full rank. Note that as

$$\boldsymbol{D} = \boldsymbol{D}_m$$

it also follows that the assertion of (44) has been proven.

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

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

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

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{D}_r & \mathbf{0} & \cdots & \mathbf{0} \\ \times & \bar{\boldsymbol{F}}_{r+1}^{\mathrm{T}} & \cdots & \mathbf{0} \\ \times & \times & \ddots & \vdots \\ \times & \times & \times & \bar{\boldsymbol{F}}_m^{\mathrm{T}} \end{bmatrix} \tag{50}$$

where

$$\bar{\boldsymbol{F}}_k = [\, \bar{\boldsymbol{u}}_1 \quad \cdots \quad \bar{\boldsymbol{u}}_{k-1} \quad 2\bar{\boldsymbol{u}}_k \,] \tag{51}$$

and where $\bar{\boldsymbol{u}}_i$ is the ith column of $\bar{\boldsymbol{U}}$ in $(\bar{\boldsymbol{U}}, \bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{z}}) = \bar{\boldsymbol{\omega}}$. The structure of (51) follows by differentiating $\operatorname{svec}\left(\boldsymbol{U}_r^{\mathrm{T}} \boldsymbol{U}_r - \boldsymbol{I} \right)$ with respect to the kth column of $\boldsymbol{U}_r$ (remember that svec forms a vector of the *upper* triangular part of its matrix argument). Note that $\boldsymbol{F}_k^{\mathrm{T}} \in \mathbb{R}^{k \times m}$ is full rank for any k, $1 \leq k \leq m$ (as the rows are orthogonal), and that $\boldsymbol{D}_r \in \mathbb{R}^{p_r \times q_r}$ is full rank as proven earlier. By considering the structure of $\boldsymbol{P}$, it follows that a linearly independent set of columns can be selected by choosing p_r columns form the set of columns containing $\boldsymbol{D}_r$ and k columns from each set containing $\boldsymbol{F}_k$ for $k = r+1, \ldots, m$. As elaborated on earlier, this is however equivalent to the statement that the set of solutions to (41) can be parameterized locally around $\bar{\boldsymbol{\omega}}$ by specifying $q_r - p_r$ parameters from $\boldsymbol{\omega}_r$, $m - k$ parameters from $\boldsymbol{u}_k$ along with λ_k and z_k for $k = r+1, \ldots, m$.

Now, we turn attention to the original problem posed by Lemma 4, that is, the problem of obtaining a covering of $\mathcal{A}(\boldsymbol{a},\boldsymbol{b})$ defined in (24) and where $\boldsymbol{a}=(a_1,\ldots,a_m)$, $\boldsymbol{b}=(b_1,\ldots,b_m)$, and $0\le b_1\le\cdots\le b_m$. Let r be the maximum integer for which

$$0=b_1=\cdots=b_r<b_{r+1}\le\cdots\le b_m.$$

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

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

where κ is the constant introduced in (45).

Consider the set

$$\mathcal{M}(\boldsymbol{b})\triangleq\mathcal{M}\cap\left\{(\boldsymbol{U},\boldsymbol{\lambda},\boldsymbol{z})\mid|\lambda_i|\le\rho^{-\frac{b_i}{2}}\right\}.$$

The set $\mathcal{M}(\boldsymbol{b})$ is chosen such that any matrix $\boldsymbol{A}\in\mathcal{A}(\boldsymbol{a},\boldsymbol{b})$ can be expressed as $\boldsymbol{A}=\boldsymbol{U\Lambda}$ for some $(\boldsymbol{U},\boldsymbol{\lambda},\boldsymbol{z})\in\mathcal{M}(\boldsymbol{b})$. Thus, the parametrization of $\mathcal{M}(\boldsymbol{b})$ will also provide a parametrization of $\mathcal{A}(\boldsymbol{a},\boldsymbol{b})$. Let $\{\psi^{(l)}\}_{l=1}^{L}$ be a set of parameterizations (around rank r points) such that

$$\mathcal{M}(\boldsymbol{b})\subset\bigcup_{l=1}^{L}\mathcal{R}^{(l)}\tag{52}$$

where $\mathcal{R}^{(l)}\triangleq\psi^{(l)}(\mathcal{D})$. The assumption that $\rho^{-\frac{b_{r+1}}{2}}\le\kappa$ ensures that it is suffice to consider parameterizations around rank r points $\bar{\boldsymbol{\omega}}\in\mathcal{N}$, in order to cover $\mathcal{M}(\boldsymbol{b})$. Note also that by the assumption in (45) the coordinate neighborhoods of $\psi^{(l)}$ are all equal to $\mathcal{D}$. Further, because $\mathcal{M}(\boldsymbol{b})\subset\mathcal{N}$ is compact (and because $\mathcal{R}^{(l)}$ is open) it can be assumed that L is finite [34]. Define $\mathcal{D}^{(l)}(\boldsymbol{b})$ according to

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

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

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

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

$$\mathcal{A}(\boldsymbol{a},\boldsymbol{b})\subset\bigcup_{l=1}^{L}\mathcal{P}^{(l)}(\boldsymbol{b}).\tag{53}$$

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 the components obtained by selecting a subset of $(\boldsymbol{u}_1,\lambda_1,z_1,\ldots,\boldsymbol{u}_r,\lambda_r,z_r)$, in the parameter vector $\boldsymbol{\xi}$, be denoted by $\boldsymbol{\theta}_r\in\mathbb{R}^{q_r-p_r}$. Similarly, let the components obtained from $\boldsymbol{u}_k$, 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. Assuming that $\boldsymbol{\xi},\hat{\boldsymbol{\xi}}\in\mathcal{D}^{(l)}(\boldsymbol{b})$, let $(\boldsymbol{U},\boldsymbol{\lambda},\boldsymbol{z})=\psi^{(l)}(\boldsymbol{\xi})$ and $(\hat{\boldsymbol{U}},\hat{\boldsymbol{\lambda}},\hat{\boldsymbol{z}})=\psi^{(l)}(\hat{\boldsymbol{\xi}})$ and let $\boldsymbol{A}=\boldsymbol{U\Lambda}$ and $\hat{\boldsymbol{A}}=\hat{\boldsymbol{U}}\hat{\boldsymbol{\Lambda}}$, where $\hat{\boldsymbol{\Lambda}}\triangleq\mathrm{Diag}(\hat{\boldsymbol{\lambda}})$. Further, let $\tilde{\boldsymbol{A}}=\hat{\boldsymbol{A}}-\boldsymbol{A}$, i.e., $\tilde{\boldsymbol{A}}$ be the perturbation in $\boldsymbol{A}$ 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

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

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

$$\|\hat{\boldsymbol{A}}-\boldsymbol{A}\|=\|\tilde{\boldsymbol{A}}\|\le\rho^{-\frac{1}{2}}.\tag{54}$$

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 $\hat{\boldsymbol{\lambda}}$, $\tilde{\boldsymbol{\lambda}}$, $\hat{\boldsymbol{z}}$, and $\tilde{\boldsymbol{z}}$, respectively.

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

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

and let $\tilde{\boldsymbol{u}}_k$ denote the kth column of $\tilde{\boldsymbol{U}}$. The first step is to prove that $\|\tilde{\boldsymbol{u}}_k\|_\infty\le cK_k\rho^{-\frac{1-b_k}{2}}$ for some constant K_k. Note that since $b_1\le\cdots\le b_m$ it follows immediately from the Lipschitz continuity of ψ that $\|\tilde{\boldsymbol{u}}_m\|\le cK_m\rho^{-\frac{1-b_m}{2}}$ for some constant K_m. This is because $\rho^{-\frac{1-b_k}{2}}\le\rho^{-\frac{1-b_m}{2}}$ for $k\le m$ implies that $\|\tilde{\boldsymbol{\xi}}\|_\infty\le c\rho^{-\frac{1-b_m}{2}}$ and K_m could simply be selected as the Lipschitz constant (in ∞-norm) of ψ. For $k<m$, let $\boldsymbol{U}_k\in\mathbb{R}^{m\times r}$ be the matrix consisting of the first k columns of $\boldsymbol{U}$, let $\boldsymbol{\lambda}_k\in\mathbb{R}^k$ the vector of the first k elements of $\boldsymbol{\lambda}$, and let $\boldsymbol{z}_k\in\mathbb{R}^k$ be the vector of the first k elements of $\boldsymbol{z}$. Assume that $\|\tilde{\boldsymbol{u}}_i\|\le cK_i\rho^{-\frac{1-b_i}{2}}$ for some $k<i\le m$ and note that $(\boldsymbol{U}_k,\boldsymbol{\lambda}_k,\boldsymbol{z}_k)$ must satisfy (47) for

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

and

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

Note also that, by the structure of $\boldsymbol{P}$ in (50), it follows that

$$(\boldsymbol{U}_k,\boldsymbol{\lambda}_k,\boldsymbol{z}_k)=\psi_k(\boldsymbol{\theta}_r,\boldsymbol{\theta}_{r+1},\lambda_{r+1},z_{r+1},\ldots,\boldsymbol{\theta}_k,\lambda_k,z_k,\boldsymbol{\gamma},\eta)\tag{55}$$

where ψ_k is the function given by the implicit function theorem in (49). By expanding

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

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 $\tilde{\eta}\triangleq\hat{\eta}-\eta$ satisfy

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

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

By the Lipschitz continuity of ψ_k in (49), it follows that

$$\|\tilde{\boldsymbol{u}}_k\|^2 \le cK_k\rho^{-\frac{1-b_k}{2}}$$

for some constant K_k because the argument in (55) is bounded by

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

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

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

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

$$\|\tilde{\boldsymbol{A}}\| = \|\hat{\boldsymbol{A}} - \boldsymbol{A}\| \le \rho^{-\frac{1}{2}}.$$

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

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

where

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

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

$$\mathcal{P}^{(l)}(\boldsymbol{b}) \subset \bigcup_{i=1}^{I} \mathcal{A}_\rho(\boldsymbol{A}^{(l,i)}) \tag{56}$$

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

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

$\boldsymbol{\Lambda}^{(l,i)} \triangleq \mathrm{Diag}(\boldsymbol{\lambda}^{(l,i)})$, and where $\mathcal{A}_\rho(\boldsymbol{A})$ is defined in (26). However, as $\mathcal{C}(\boldsymbol{\xi})$ is simply a (rectangular) box centered at $\boldsymbol{\xi}$ and because

$$\begin{aligned}\mathcal{D}^{(l)}(\boldsymbol{b}) \subset \Big\{\boldsymbol{\xi} \mid \|\boldsymbol{\theta}_r\|_\infty \le 2,\ \|\boldsymbol{\theta}_k\|_\infty \le 1,\ |\lambda_k| \le \rho^{-\frac{b_k}{2}}, \\ |z_k| \le 2\rho^{-\frac{b_k}{2}},\ k = r+1, \ldots, m\Big\}\end{aligned} \tag{57}$$

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

$$I \mathrel{\dot{\le}} \rho^{\mu}$$

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 $\rho^{-\beta_i}, i = 1, \ldots, M$, with small boxes of side length $\rho^{-\alpha_i}$, $i = 1, \ldots, M$, one needs (in the $\doteq$ sense)

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

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 been shown so far that it is possible to cover $\mathcal{P}^{(l)}$ by $I \mathrel{\dot{\le}} \rho^{\mu}$ sets $\mathcal{A}_\rho(\boldsymbol{A}_i)$. By (53) and since L was finite this result extends to the covering of $\mathcal{A}(\boldsymbol{a}, \boldsymbol{b})$. That is, it has been shown that there exists a covering $\mathfrak{A}$, which satisfies

$$\mathcal{A}(\boldsymbol{a}, \boldsymbol{b}) \subset \bigcup_{\boldsymbol{A}_i \in \mathfrak{A}} \mathcal{A}_\rho(\boldsymbol{A}_i)$$

and

$$|\mathfrak{A}| \mathrel{\dot{\le}} \rho^{\mu}$$

as was asserted by Lemma 4. □

References

[1] D. Tse and P. Wiswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[2] S. Verdú, *Multiuser Detection*. Cambridge, U.K.: Cambridge University Press, 1998.

[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 Trans. Signal Process.*, vol. 53, no. 4, pp. 1474–1484, Apr. 2005.

[5] H. Yao and G. W. Wornell, "Lattice-reduction-aided detectors for MIMO communication systems," in *Proc. IEEE GLOBECOM*, Nov. 2002, vol. 1, pp. 424–428.

[6] C. Windpassinger and R. F. H. Fisher, "Low-complexity near-maximum-likelihood detection and precoding for MIMO systems using lattice reduction," in *Proc. IEEE Inf. Theory Workshop*, Apr. 2003, pp. 345–348.

[7] P. Tan and L. Rasmussen, "The application of semidefinite programming for detection in CDMA," *IEEE J. Sel. Areas Commun.*, vol. 19, no. 8, pp. 1442–1449, Aug. 2001.

[8] 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 Trans. Signal Process.*, vol. 50, no. 4, pp. 912–922, Apr. 2002.

[9] 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 Process. Lett.*, vol. 9, no. 6, pp. 165–167, Jun. 2002.

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

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

[12] N. D. Sidiropoulos and Z.-Q. Luo, "A semidefinite relaxation approach to MIMO detection for high-order QAM constellations," *IEEE Signal Process. Lett.*, vol. 13, no. 9, pp. 525–528, Sep. 2006.

[13] A. Mobasher, M. Taherzadeh, R. Sotirov, and A. K. Khandani, "A near maximum likelihood decoding algorithm for MIMO systems based on semi-definite programming," *IEEE Trans. Inf. Theory*, vol. 53, no. 11, pp. 3869–3886, Nov. 2007.

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

[15] M. Kisialiou and Z.-Q. Luo, "Performance analysis of quasi-maximum-likelihood detector based on semi-definite programming," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Mar. 2005, vol. 3, pp. iii/433–iii/436.

[16] M. Taherzadeh, A. Mobasher, and A. K. Khandani, "LLL reduction achieves the receive diversity in MIMO decoding," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4801–4805, Dec. 2006.

[17] L. Lovász, "On the Shannon capacity of a graph," *IEEE Trans. Inf. Theory*, vol. IT-25, no. 1, pp. 1–7, Jan. 1979.

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

[19] F. Jarre, "An interior-point method for minimizing the maximum eigenvalue of a linear combination of matrices," *SIAM J. Control Optim.*, vol. 31, no. 5, pp. 1360–1377, Sep. 1993.

[20] Y. Nesterov and A. Nemirovski, *Interior Point Polynomial Algorithms in Convex Programming*. Philadelphia, PA: SIAM, 1994.

[21] L. Vandenberghe and S. Boyd, "A primal-dual potential reduction method for problems involving matrix inequalities," *Math. Programm.*, vol. 69, no. 1, pp. 205–236, Jul. 1995.

[22] H. Wolkowicz, R. Saigal, and L. Venberghe, Eds., *Handbook of Semidefinite Programming*. Norwell, MA: Kluwer, 2000.

[23] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

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

[25] M. Kisialiou and Z.-Q. Luo, "Efficient implementation of a quasi-maximum-likelihood detector based on semi-definite relaxation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Apr. 2007, vol. 4, pp. IV-1329–IV-1332.

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

[27] S. Poljak, F. Rendl, and H. Wolkowicz, "A recipe for semidefinite relaxation for (0,1)-quadratic programming," *J. Global Optim.*, vol. 7, no. 1, pp. 51–73, Jul. 1995.

[28] J. Jaldén, C. Martin, and B. Ottersten, "Semidefinite programming for detection in linear systems—Optimality conditions and space-time decoding," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Apr. 2003, vol. 4, pp. IV-9–IV-12.

[29] J. Jaldén, "Detection for multiple input multiple output channels: Analysis of sphere decoding and semidefinite relaxation," Ph.D. dissertation, Schl. Electr. Eng., Royal Inst. Technol. (KTH), Stockholm, Sweden, 2006.

[30] R. V. Nee, A. V. Zelst, and G. Awater, "Maximum likelihood decoding in a space division multiplexing system," in *Proc. IEEE Veh. Technol. Conf.*, Tokyo, Japan, May 2000, vol. 1, pp. 6–10.

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

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

[33] J. W. Milnor, *Topology From the Differentiable Viewpoint*. Princeton, NJ: Princeton Univ. Press, 1965.

[34] W. Rudin, *Principles of Mathematical Analysis*, 3rd ed. New York: McGraw-Hill, 1996.

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

[36] R. G. Bartle, *The Elements of Real Analysis*. New York: Wiley, 1964.