On the Complexity of Sphere Decoding in Digital Communications

IEEE Transactions of Signal Processing

Joakim Jaldén, Björn Ottersten

April 2005

IR-S3-SB-0513

ROYAL INSTITUTE
OF TECHNOLOGY
Department of
Signals, Sensors & Systems
Signal Processing
S-100 44 STOCKHOLM

KUNGL TEKNISKA HÖGSKOLAN
Institutionen för
Signaler, Sensorer & System
Signalbehandling
100 44 STOCKHOLM

On the Complexity of Sphere Decoding in Digital Communications

Joakim Jaldén*, *Student member, IEEE,* Björn Ottersten, *Fellow, IEEE.*

Dept. of Signals, Sensors and Systems
Royal Institute of Technology
Osquldas väg 10
SE-100 44 Stockholm, Sweden

***Abstract*— Sphere decoding has been suggested by a number of authors as an efficient algorithm to solve various detection problems in digital communications. In some cases the algorithm is referred to as an algorithm of polynomial complexity without clearly specifying what assumptions are made about the problem structure. Another claim is that, although worst case complexity is exponential, the expected complexity of the algorithm is polynomial. Herein we study the expected complexity where the problem size is defined to be the number of symbols jointly detected and our main result is that the expected complexity is exponential for fixed SNR, contrary to previous claims. The sphere radius, a parameter of the algorithm, must be chosen to ensure a non-vanishing probability of solving the detection problem. This causes the exponential complexity since the squared radius must grow linearly with problem size. The rate of linear increase is however dependent on the noise variance and thus the rate of the exponential function is strongly dependent on the SNR. Therefore sphere decoding can be efficient for some SNR and problems of moderate size, even though the number of operations required by the algorithm strictly speaking always grows as an exponential function of the problem size.**

***Index Terms*— ML detection, sphere decoding, expected complexity, large deviation theory.**

I. Introduction

MAXIMUM likelihood (ML) detection of digital messages in general requires joint detection of an entire block of symbols [1]. For problems of certain structure efficient algorithms, such as the Viterbi algorithm, can successfully be applied. In general however, when no exploitable structure is at hand, the detection problem is very computationally intensive. Such hard instances of the detection problem arise in for example multiuser detection (MUD) problems in code division multiple access (CDMA) [2], [3] and linear dispersive space time block coding (LD-STBC) [4].

Consequently, there has recently been a growing interest in sphere decoding for ML detection in digital communications [5]–[7]. Sphere decoding, or the Fincke-Pohst algorithm [8], [9], offers large reductions in computational complexity for the class of computationally hard combinatorial problems that arise in the aforementioned (ML) detection problems. In [8] it is shown that the complexity of sphere decoding, under certain assumptions, is polynomial in the problem size meaning that there is a polynomial function of the problem size that bounds the number of operations required by the algorithm. The assumptions made in [8] were however made in another context and are not generally applicable to the ML detection problem encountered in digital communications [10].

The primary topic treated in this paper, as in [10], [11], is the expected number of operations required by the algorithm where the expected value is computed over the channel and noise realizations as well as the possible transmitted messages. The problem size will be defined as the number of symbols, m, that are jointly detected and the constellation size for each symbol, L, will be kept fixed. As shown in [10] the expected complexity, $C(m) = C(m, \rho)$, of the sphere decoder is dependent both on the size, m, and the SNR, ρ. It is also shown in [10] that, when the SNR is high, the expected number of operations required by the sphere decoder can be approximated by a polynomial function for small m. However, there does not exist, for any fixed ρ, a polynomial upper bound on $C(m)$ which holds for all m. Therefore, the algorithm is strictly speaking not of polynomial expected complexity under the usual definition [12], [13]. This result is proven herein by deriving exponential lower bounds on $C(m)$ for a large class of ML detection problems.

In [10] an exact expression for the expected number of operations required by the sphere decoder is obtained. However, this expression is hard to interpret and increasingly difficult to compute for larger problem sizes. Therefore it is not straightforward to see whether the expression tends to infinity as an exponential or a polynomial function. This paper differs from [10] since here, rather than studying an exact expression, the asymptotic behavior of $C(m)$ is considered. Specifically we show that the expected number of operations tends to infinity as $L^{\gamma m}$ where $\gamma \in (0, 1]$ is some small factor dependent on ρ. This can be compared to full search for which the number of operations tend to infinity as L^m. However, for large SNR the factor $\gamma \ll 1$ and therefore $L^{\gamma m}$ is close to 1 when m is small. This means that for large ρ and small m the complexity $C(m)$ is dominated by polynomial terms which is consistent with the results of [10].

A main contribution of this paper is to derive a way to compute γ for the specific case considered in [10]. Determining γ is valuable since it provides useful insight into which problem sizes, m, can be considered small. The problem structure considered herein is properly defined in Section II and sphere decoding algorithm is explained in Section III. After a brief

discussion about the definition of expected complexity in Section IV an expression, suitable for asymptotic analysis, is developed in Section V. This expression serves as the basis for a theorem in Section VI, which shows that $C(m)$ tends to infinity as an exponential function for a large class of detection problems. This class of problems includes the problem considered in [10] as a special case. In Section V, attention is restricted to the specific problem in [10] and a method for computing the asymptotic approximation of $C(m)$ is developed using the theory of large deviations. These results are illustrated by examples in Section VIII.

II. Problem Definition

In this section a generic model for the communication system is introduced. The generality of the model is chosen such that it includes the examples of CDMA [3] and LD-STBC [4] mentioned in the introduction. Furthermore it extends to multiple input multiple output (MIMO) systems, both in the flat fading and the frequency selective, block transmission, scenario [7]. Also, intersymbol interference (ISI) problems on a finite impulse response channel are included although in some of these cases sphere decoding may not be the best choice of algorithm. Note that all results, up to Section VII, hold in this general context.

Consider the maximum likelihood (ML) detection of a message $\bar{\mathbf{s}}$, drawn from an m-dimensional L-PAM constellation $\mathcal{D}_L^m$, which is sent across a linear channel and disturbed by additive noise. The received signal, $\mathbf{x} \in \mathbb{R}^m$, is given by

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

where $\mathbf{H} \in \mathbb{R}^{m \times m}$ is the channel matrix and $\mathbf{v} \in \mathbb{R}^m$ is the additive noise. The channel matrix, $\mathbf{H}$, is assumed randomly drawn from some distribution and known to the receiver. The noise, $\mathbf{v}$, is assumed to be white Gaussian noise, i.e. each component of $\mathbf{v}$ is assumed independently drawn from a normal, $\mathcal{N}(0, \sigma^2)$, distribution. The message $\bar{\mathbf{s}}$ belongs to the set $\mathcal{D}_L^m$ defined as

$$\mathcal{D}_L^m = \left\{ -\frac{L-1}{2}, -\frac{L-3}{2}, \ldots, \frac{L-3}{2}, \frac{L-1}{2} \right\}^m.$$

That is, each element of $\bar{\mathbf{s}}$ takes one of L different values, which are at integer spacing and centered around 0.

Under the above assumptions the ML estimate of $\bar{\mathbf{s}}$ is well known to be [1]

$$\hat{\mathbf{s}}_{\mathrm{ML}} = \underset{\mathbf{s} \in \mathcal{D}_L^m}{\operatorname{argmin}} \ \|\mathbf{x} - \mathbf{H}\mathbf{s}\|^2 \tag{2}$$

and for a general $\mathbf{H}$ this problem is known to be NP-hard [14].

In (1) the channel matrix, $\mathbf{H}$, is assumed to be a square and real valued matrix. This assumption simplifies the theoretical development of this paper. However, the results extend to the case of tall matrices, i.e. if $\mathbf{H}$ is an $\mathbb{R}^{n \times m}$ matrix for some $n \geq m$. The reason for this is that such a problem can be rewritten in the form of (1) by projecting the problem onto the column space of $\mathbf{H}$. The case of $m > n$, even though some of the results hold with slight modifications, will not be considered and from here on it shall be assumed that $m = n$.

The case of complex valued $\mathbf{H}$, $\bar{\mathbf{s}}$, and $\mathbf{v}$ can be rewritten in the above form under the additional assumption that each component of $\mathbf{v}$ is circularly symmetric complex Gaussian. More specifically, by expanding the problem dimensionality a complex valued problem can be written as

$$\begin{bmatrix} \Re(\mathbf{x}) \\ \Im(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} \Re(\mathbf{H}) & \Im(\mathbf{H}) \\ -\Im(\mathbf{H}) & \Re(\mathbf{H}) \end{bmatrix} \begin{bmatrix} \Re(\bar{\mathbf{s}}) \\ \Im(\bar{\mathbf{s}}) \end{bmatrix} + \begin{bmatrix} \Re(\mathbf{v}) \\ \Im(\mathbf{v}) \end{bmatrix} \tag{3}$$

where $\Re(\mathbf{x})$ and $\Im(\mathbf{x})$ denote the real and imaginary parts of $\mathbf{x}$ respectively. This means that QAM constellations are included in the framework of this paper.

Herein we define the SNR, ρ, at the receiver as

$$\rho = \frac{\mathrm{E}\left\{\|\mathbf{H}\bar{\mathbf{s}}\|^2\right\}}{\mathrm{E}\left\{\|\mathbf{v}\|^2\right\}} \tag{4}$$

where in the above, and in the following, all messages $\bar{\mathbf{s}}$ are assumed to be drawn with the same probability. That is, $\bar{\mathbf{s}}$ is uniformly distributed on $\mathcal{D}_L^m$. Also, an additional assumption will be made concerning the distribution of $\mathbf{H}$. To be specific, it will be assumed that there exist some constant c, independent of m, such that

$$\mathrm{E}\left\{\|\mathbf{h}_i\|^2\right\} \leq c^2 \quad \forall\ i \in [1, m] \tag{5}$$

where $\mathbf{h}_i$ is the ith column of $\mathbf{H}$. The interpretation of this assumption is that each symbol is transmitted with finite energy. While it is clear that this assumption is satisfied for most systems of practical interest it is important to explicitly state when the asymptotic properties of $C(m)$ are considered.

III. Sphere Decoding

This section is intended to give sufficient understanding of the sphere decoding algorithm for the reader to follow the complexity computations of the following sections. It will also introduce some important notations and concepts. The purpose of this section is not to explain the implementational aspects of sphere decoding, these are discussed in for instance [8] or [10], [11].

Sphere decoding solves (2) by searching only over those points that satisfy a constraint of the form

$$\|\mathbf{x} - \mathbf{H}\mathbf{s}\|^2 \leq r^2. \tag{6}$$

In other words, sphere decoding only considers points that lie inside a hypersphere of radius r. An efficient way to check this criterion, referred to as the Phost Strategy [8], is as follows.

Let $\mathbf{QR} = \mathbf{H}$ be the QR factorization of the channel matrix $\mathbf{H}$, i.e. $\mathbf{R}$ is an upper right triangular matrix and $\mathbf{Q}$ an orthogonal matrix. Due to the invariance of the ℓ_2 norm to orthogonal transforms the constraint of (6) can be rewritten as

$$\|\mathbf{R}\mathbf{s} - \mathbf{Q}^{\mathrm{T}}\mathbf{x}\|^2 \leq r^2.$$

Let $\mathbf{p} \in \mathbb{R}^m$ be given by

$$\mathbf{p} = \mathbf{p}(\mathbf{s}) = \mathbf{R}\mathbf{s} - \mathbf{Q}^{\mathrm{T}}\mathbf{x} \tag{7}$$

where emphasis has been placed on the fact that $\mathbf{p}$ is a function of $\mathbf{s}$. Equation (6) can now be written as

$$\sum_{i=1}^{m} p_i^2 \leq r^2 \tag{8}$$

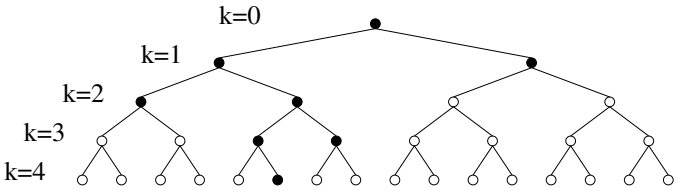

Fig. 1. Illustration of sphere decoding search tree for 2-PAM constellation with $m = 4$. The nodes visited by sphere decoding are shown in black.

where p_i is the ith entry of $\mathbf{p}$. Due to the upper triangular structure of $\mathbf{R}$ the ith entry of $\mathbf{p}$, p_i, is a function only of s_j, $j = i, \ldots, m$, where s_j is the jth entry of $\mathbf{s}$. Since p_i^2 is positive

$$\sum_{i=m-k+1}^{m} p_i^2 \leq r^2, \qquad k = 1, \ldots, m \tag{9}$$

follows from (8) and in particular for $k = 1$

$$p_m^2(s_m) \leq r^2. \tag{10}$$

Since $p_m^2(s_m)$ is a (positive semidefinite) quadratic function in s_m this gives upper and lower bounds for the values that can be assigned to s_m without violating (6). For each choice of s_m

$$p_{m-1}^2(s_{m-1}, s_m) \leq r^2 - p_m^2(s_m) \tag{11}$$

provides upper and lower bounds on s_{m-1} and in general

$$p_i^2(s_i, \ldots, s_m) \leq r^2 - \sum_{j=i+1}^{m} p_j^2(s_j, \ldots, s_m) \tag{12}$$

gives upper and lower bounds on s_i given $s_{i+1}, \ldots, s_m$.

If full search is illustrated as a search tree where each path trough the tree correspond to a possible message, $\mathbf{s}$, then sphere decoding can be viewed as a pruning algorithm on this tree where a subtree can be rejected at some depth k based on violation of the constraint given by (12). This is shown in Figure 1 for a 2-PAM constellation.

Sphere decoding traces down branches of the search tree until (12) is violated. At this point the algorithm backtracks and proceeds down a different branch. The efficiency of sphere decoding hinges on the ability to reject entire subtrees high up in search tree. The number of operations required by the sphere decoding algorithm is given by the number of nodes visited in the tree as well as the number of operations required per node. However, in this paper the complexity of the algorithm is defined as follows.

Definition 1: The complexity of sphere decoding is the expected number of nodes visited in the search tree where the expected value is computed over the channel $\mathbf{H}$, noise, $\mathbf{v}$, and transmitted message, $\bar{\mathbf{s}}$.

It should be clear that the radius, r, of the sphere must be chosen such that at least one of the possible paths through the tree reaches the bottom or no points $\mathbf{s}$ will satisfy (6). In a communications application it is often enough to require that this happens with high probability.

In [10] a good way to choose r is given as follows. Note that for the message sent, $\bar{\mathbf{s}}$,

$$\|\mathbf{x} - \mathbf{H}\bar{\mathbf{s}}\|^2 = \|\mathbf{v}\|^2 \tag{13}$$

is a χ_m^2 distributed variable. The radius could thus be chosen such that

$$\Pr\left\{\|\mathbf{v}\|^2 \leq r^2\right\} = 1 - \varepsilon \tag{14}$$

for some $\varepsilon \ll 1$ related to the target error probability. For this reason the radius, r, will in this paper be assumed to satisfy

$$r^2 \geq \mathrm{E}\left\{\|\mathbf{v}\|^2\right\} = m\sigma^2. \tag{15}$$

If this is not fulfilled, the probability of finding $\bar{\mathbf{s}}$ inside the sphere would tend to zero as m grew large. Note that one of the assumptions in [8] which leads to a polynomial complexity result is that r remains fixed and independent of m. From (15) it can be seen that such an assumption is not realistic in the digital communications scenario since it would render the algorithm useless as m grows. The constraint of (15) is however unfortunately what causes the exponential complexity as will be shown in Section VI.

Finally, it is debatable whether the chosen complexity measure, i.e. the expected complexity, is the most judicious choice. However, since there exist choices of $\mathbf{H}$ for which the number of nodes visited range from m to L^m the best or worst case complexity does not provide good insight into the behavior of the algorithm. For this reason the expected number of nodes is used in the definition above. Additional justification of this complexity measure is also given by remark 4 at the end of Section VII.

IV. Complexity

Before entering into a discussion about the complexity of the sphere decoder it is useful to provide a formal definition of the concept of polynomial expected complexity. It is also important to remember that the expected complexity of the sphere decoder is dependent on the probability distribution of the channel matrix, $\mathbf{H}$, and whenever the expected complexity is discussed it is implicitly assumed that there is some underlying distribution of $\mathbf{H}$.

A function $f(m)$ is said to be $O(g(m))$ if there exist some finite c and m' such that [13]

$$f(m) \leq cg(m) \qquad \text{for all} \qquad m \geq m'. \tag{16}$$

Assume that some sequence of probability distributions of the channel matrixes $\mathbf{H} \in \mathbb{R}^{m \times m}$ for $m = 1, 2, \ldots$ is given and let $C = C(m)$ be the expected, or average, number of nodes visited by the sphere decoder. Then the sphere decoder is said to be of polynomial expected complexity [12], [13] if $C(m)$ is $O(m^k)$ for some fixed k. Note that this can not be true if there exist an exponential lower bound on $C(m)$ for all m and that the existence of such a lower bound is shown herein.

Some care should be taken when considering $C(m)$ for finite m and comparing this to asymptotic statements such as polynomial expected complexity. This is illustrated by the fact that the function $C(m)$ may be well approximated by a polynomial function for all m of practical interest even when the algorithm is, strictly speaking, not of polynomial expected complexity. Consider for example the function

$$(m^2 + m + 1)2^{\gamma m}$$

for some small $\gamma > 0$. This function is well approximated by a polynomial function for moderate m but is not $O(m^k)$ for any k. That in fact the expected number of operations required by the sphere decoder, for large ρ and moderate m, can be approximated by a polynomial function is what is shown in [10], [11] and experimentally verified in [15]. There is however no finite SNR, ρ, for which the exponential term vanishes completely. The point is that there is always some m for which the sphere decoder is more computationally intensive then a, possibly suboptimal, polynomial time algorithm but that this m might be large.

Finally, $C(m)$ is the expected number of nodes visited by the sphere decoder and is referred to as the expected complexity of the algorithm. However, recall that all statements about polynomial or exponential complexity always refer to the asymptotic nature of $C(m)$.

V. Computing the Expected Complexity

To compute the expected complexity of the algorithm it is convenient to first view the number of nodes, N, visited by the algorithm as a function of $\mathbf{H}$, $\mathbf{v}$ and $\bar{\mathbf{s}}$ and thereafter compute the expected complexity, C, as

$$C = \mathrm{E}\left\{N\right\}. \tag{17}$$

In this section it is shown that the number of nodes, N, itself can be written as the expected value of a function of $\mathbf{s}$, where $\mathbf{s}$ is a randomly selected path through the search tree. By doing so an expression for C which now includes an expected value computed over $\mathbf{H}$, $\mathbf{v}$, $\bar{\mathbf{s}}$ and $\mathbf{s}$ can be obtained. To prove this is largely an exercise in reordering summations but the benefit will in the following sections be made clear by the use of standard tools from probability theory to obtain useful results about the complexity.

To simplify the notation the following notation will be used. Let $\mathbf{x}$ be a vector in $\mathbb{R}^m$. Then $\mathbf{x}_i^j$ will denote the vector in $\mathbb{R}^{j-i+1}$ of the components x_i through x_j. i.e.

$$\mathbf{x}_i^j = \begin{bmatrix} x_i & \cdots & x_j \end{bmatrix}^{\mathrm{T}}.$$

Furthermore, no difference in notation will be made between stochastic variables and their realization. Thus, e.g.

$$\mathrm{E}_{\mathbf{s}}\left\{g(\mathbf{s})\right\} = \sum_{\mathbf{s}} f(\mathbf{s}) g(\mathbf{s})$$

were $f(\mathbf{s})$ is the probability mass function of $\mathbf{s}$ when in fact the $\mathbf{s}$ on the left hand side of the equation is a stochastic variable and the right hand side is the sum over all realizations of $\mathbf{s}$. The stochastic variable over which the expected value is computed will be indicated as above when necessary. Furthermore, the notation

$$\sum_{\mathbf{s}_i^j} \tag{18}$$

will be used to denote the sum over all possible combinations of symbols i through j of the vector $\mathbf{s}$, i.e. a sum over L^{j-i+1} components.

As stated in Section III the complexity of the algorithm is given by the expected number of nodes in the search tree. It turns out that it is convenient to define the depth in the search tree at which a particular path is cut off.

Definition 2: Given $\mathbf{H}$, $\mathbf{v}$, $\bar{\mathbf{s}}$ and some path through the tree, $\mathbf{s}$, the *search depth*, d, is defined as

$$d = \sup\{k \in \mathbb{Z} \mid k \in [0, m] \ , \ \|\mathbf{p}_{m-k+1}^m\|^2 \leq r^2\} \tag{19}$$

where $\mathbf{p}_{m+1}^m = 0$ by definition and $\mathbf{p}$ is given by (7).

The search depth, d, is an integer valued function of $\mathbf{H}$, $\mathbf{v}$, $\bar{\mathbf{s}}$ and $\mathbf{s}$ with values in the range 0 to m. This will sometimes be emphasized by writing out the arguments, i.e. $d = d(\mathbf{s})$ if the dependence on the search path is important for the argument. It should also be noted that d is dependent on the search radius.

Lemma 1: For fixed $\mathbf{H}$, $\mathbf{v}$ and $\bar{\mathbf{s}}$, the number of nodes visited, N, in the search tree is given by

$$N = \frac{\mathrm{E}_{\mathbf{s}}\left\{L^{d(\mathbf{s})+1}\right\} - 1}{L - 1} \tag{20}$$

where $d = d(\mathbf{s})$ is defined as above and $\mathbf{s}$ is a random variable uniformly distributed on $\mathcal{D}_L^m$.

Proof: Let $P_k(\mathbf{s}_{m-k+1}^m)$ be an indicator function that equals 1 if $\|\mathbf{p}(\mathbf{s})_{m-k+1}^m\|^2 \leq r^2$ and 0 otherwise. Note that P_k is a function of $s_k, \ldots, s_m$ only. Also, for purely notational purposes, let

$$\sum_{\mathbf{s}_{m+1}^m} x = \sum_{\mathbf{s}_1^0} x = x.$$

Then, by summing over all possible nodes and using the indicator function, $P_k(\mathbf{s}_{m-k+1}^m)$, the number of nodes is given by

$$\begin{aligned}
N &\overset{\text{a}}{=} \sum_{k=0}^{m} \left[\sum_{\mathbf{s}_{m-k+1}^m} P_k(\mathbf{s}_{m-k+1}^m) \right] \\
&\overset{\text{b}}{=} \sum_{k=0}^{m} \left[\sum_{\mathbf{s}_1^{m-k}} L^{-(m-k)} \sum_{\mathbf{s}_{m-k+1}^m} P_k(\mathbf{s}_{m-k+1}^m) \right] \\
&= L^{-m} \sum_{k=0}^{m} L^k \left[\sum_{\mathbf{s}} P_k(\mathbf{s}_{m-k+1}^m) \right] \\
&= \sum_{\mathbf{s}} L^{-m} \sum_{k=0}^{m} L^k P_k(\mathbf{s}_{m-k+1}^m) \\
&\overset{\text{c}}{=} \sum_{\mathbf{s}} L^{-m} \sum_{k=0}^{d(\mathbf{s})} L^k \\
&\overset{\text{d}}{=} \sum_{\mathbf{s}} L^{-m} \ (L^{d(\mathbf{s})+1} - 1)/(L-1).
\end{aligned} \tag{21}$$

In the above, (a) is the summation of all nodes visited at depth k in the search tree for $k = 0, \ldots, m$, (b) follows since

$$\sum_{\mathbf{s}_1^{m-k}} L^{-(m-k)} = 1 \tag{22}$$

and (c) follows from

$$P_k(\mathbf{s}_{m-k+1}^m) = 1 \Rightarrow P_l(\mathbf{s}_{m-l+1}^m) = 1 \text{ if } k \geq l, \tag{23}$$

and

$$P_k(\mathbf{s}_{m-k+1}^m) = 0 \Rightarrow P_l(\mathbf{s}_{m-l+1}^m) = 0 \text{ if } k \leq l. \tag{24}$$

The last equality, (d), is the expression for the sum of a geometric series and the last line equals the definition of the expected value in the lemma. This concludes the proof. ■

Theorem 1: The expected complexity, C, of sphere decoding, where the expected value is computed over $\mathbf{H}$, $\bar{\mathbf{s}}$ and $\mathbf{v}$, is

$$C = \frac{\mathrm{E}\left\{L^{d+1}\right\} - 1}{L - 1} \tag{25}$$

where d is defined as above and $\mathbf{s}$ is uniformly distributed over $\mathcal{D}_L^m$.

Proof: The complexity, C, is by definition the expected number of nodes in the search tree. That is

$$C = \mathrm{E}_{\mathbf{H},\bar{\mathbf{s}},\mathbf{v}}\left\{N\right\} \tag{26}$$

and from Lemma 1

$$C = \mathrm{E}_{\mathbf{H},\bar{\mathbf{s}},\mathbf{v}}\left\{\frac{\mathrm{E}_{\mathbf{s}}\left\{L^{d(\mathbf{s})+1}\right\} - 1}{L - 1}\right\} = \frac{\mathrm{E}\left\{L^{d+1}\right\} - 1}{L - 1} \tag{27}$$

which concludes the proof. ■

Remark 1: The complexity is given by the statistics of the search depth, d, along random paths. This implies, by the above and Jensen's inequality, that whenever the expected search depth, $\mathrm{E}\{d\}$, grows linearly with the problem size the algorithm will be of exponential complexity.

VI. Lower Bound On The Expected Complexity

As commented at the end of the previous section, the expected complexity grows exponentially in m, if it can be shown that $\mathrm{E}\{d\}$ grows linearly with m. This is done in this section under an additional assumption that each symbol is transmitted with finite energy. As pointed out earlier these assumptions are valid for most communications problems of practical interest. In particular, they hold for the systems considered in [6], [7], [10], [11].

Theorem 2: Assume that the noise $\mathbf{v}$, channel $\mathbf{H}$, and sent symbol $\bar{\mathbf{s}}$ are independently drawn, that all symbols $\bar{\mathbf{s}}$ are equally likely and that there exists some c such that the random channel matrix $\mathbf{H} \in \mathbb{R}^{m\times m}$ satisfies

$$\mathrm{E}\left\{\|\mathbf{h}_i\|^2\right\} \leq c^2 \quad \forall\ i \in [1, m] \tag{28}$$

where $\mathbf{h}_i$ is the ith column vector of $\mathbf{H}$. The expected complexity of the sphere decoding algorithm is then bounded below by

$$C(m) \geq \frac{L^{\eta m} - 1}{L - 1} \quad , \quad \eta = \frac{1}{2}\left(\frac{c^2(L^2-1)}{6\sigma^2} + 1\right)^{-1}. \tag{29}$$

Hence, under these assumptions, the complexity grows exponentially in the problem size.

Proof: Imposing the system model (1) on (7) yields

$$\mathbf{p} = \mathbf{R}(\mathbf{s} - \bar{\mathbf{s}}) - \mathbf{Q}^{\mathrm{T}}\mathbf{v} = \mathbf{R}(\mathbf{s} - \bar{\mathbf{s}}) + \bar{\mathbf{v}} \tag{30}$$

where the components of $\bar{\mathbf{v}}$ are i.i.d. Gaussian with variance σ^2. Note that

$$\Pr\left\{d < k\right\} = \Pr\left\{\|\mathbf{p}_{m-k+1}^m\|^2 > r^2\right\}. \tag{31}$$

By Markov's inequality [16]

$$\Pr\left\{\|\mathbf{p}_{m-k+1}^m\|^2 > r^2\right\} \leq \frac{\mathrm{E}\left\{\|\mathbf{p}_{m-k+1}^m\|^2\right\}}{r^2} \tag{32}$$

which implies

$$\begin{aligned}\Pr\left\{d \geq k\right\} &= 1 - \Pr\left\{d < k\right\} \\ &\geq 1 - \frac{\mathrm{E}\left\{\|\mathbf{p}_{m-k+1}^m\|^2\right\}}{r^2} \\ &\geq 1 - \frac{\mathrm{E}\left\{\|\mathbf{p}_{m-k+1}^m\|^2\right\}}{m\sigma^2}\end{aligned} \tag{33}$$

where the last inequality comes from $r^2 \geq m\sigma^2$ as given by (15). Due to the assumption of independent symbols

$$\begin{aligned}&\mathrm{E}\left\{\|\mathbf{p}_{m-k+1}^m\|^2\right\} = \mathrm{E}\left\{\|(\mathbf{R}(\mathbf{s} - \bar{\mathbf{s}}) + \bar{\mathbf{v}})_{m-k+1}^m\|^2\right\} \\ &= \sum_{i=m-k+1}^{m} \mathrm{E}\left\{(s_i - \bar{s}_i)^2\right\}\mathrm{E}\left\{\|(\mathbf{r}_i)_{m-k+1}^m\|^2\right\} + \mathrm{E}\left\{\bar{v}_i^2\right\}\end{aligned} \tag{34}$$

were $\mathbf{r}_i$ is the ith column of $\mathbf{R}$. Since $\mathbf{r}_i = \mathbf{Q}^{\mathrm{T}}\mathbf{h}_i$ and $\mathbf{Q}$ is an orthogonal matrix it follows from the assumptions that

$$\mathrm{E}\left\{\|(\mathbf{r}_i)_{m-k+1}^m\|^2\right\} \leq c^2. \tag{35}$$

Also

$$\mathrm{E}\left\{(s_i - \bar{s}_i)^2\right\} = \frac{L^2 - 1}{6} \tag{36}$$

and $\mathrm{E}\left\{\bar{v}_i^2\right\} = \sigma^2$ which yields

$$\mathrm{E}\left\{\|\mathbf{p}_{m-k+1}^m\|^2\right\} \leq k\left(\frac{L^2-1}{6}c^2 + \sigma^2\right) = \beta k \tag{37}$$

for

$$\beta = \left(\frac{L^2-1}{6}c^2 + \sigma^2\right) \tag{38}$$

which together with (33) yields

$$\Pr\left\{d \geq k\right\} \geq 1 - \frac{\beta k}{m\sigma^2}. \tag{39}$$

Introducing $n = \lfloor m\sigma^2/\beta \rfloor$ and the stochastic variable ν with a probability distribution

$$\Pr\left\{\nu = k\right\} = \frac{1}{n} \quad \text{for} \quad k = 0, \ldots, n-1 \tag{40}$$

yields

$$\begin{aligned}\Pr\left\{d \geq k\right\} &\geq 1 - \frac{k}{m\sigma^2/\beta} \geq 1 - \frac{k}{\lfloor m\sigma^2/\beta \rfloor} \\ &= 1 - \frac{k}{n} = \Pr\left\{\nu \geq k\right\}\end{aligned} \tag{41}$$

for $k = 0, \ldots, n-1$. The result trivially holds for $k > n-1$. From this it follows that

$$\begin{aligned}\mathrm{E}\left\{d\right\} &\geq \mathrm{E}\left\{\nu\right\} = \sum_{k=0}^{n-1}\frac{k}{n} = \frac{1}{2}(n-1) \\ &\geq \frac{1}{2}(\frac{m\sigma^2}{\beta} - 2) = \frac{m}{2\beta/\sigma^2} - 1 = \eta m - 1.\end{aligned} \tag{42}$$

By Jensen's inequality [16] and since L^x is a convex function

$$\mathrm{E}\left\{L^{d+1}\right\} \geq L^{\mathrm{E}\{d+1\}} \geq L^{\eta m} \tag{43}$$

which concludes the proof. ■

Remark 2: The constraint on $\mathbf{H}$ in Theorem 2 implies that the SNR is bounded. The contrary is not necessarily true, i.e. there may be choices of $\mathbf{H}$ where the SNR, according to the definition herein, remains bounded but the expected norm of some columns tend to infinity as m grows large. The constraint is chosen such that these, degenerate cases are excluded. Also note that it is the expected value of the norm, not the norm itself, which is bounded in the theorem.

VII. The Complexity Exponent

Theorem 2 states that the complexity is exponential in m by giving an exponential lower bound. At the same time a trivial upper bound on the complexity is given by L^m. It is therefore reasonable to assume that the complexity of sphere decoding lies somewhere between these bounds and that there exist some $\gamma \in (0, 1]$ such that the complexity is given by

$$C(m) \asymp L^{\gamma m} \tag{44}$$

where the exact meaning of $\asymp$ is that for any $\varepsilon > 0$ there is an M such that

$$L^{(\gamma+\varepsilon)m} > C(m) > L^{(\gamma-\varepsilon)m} \quad \forall\ m \geq M.$$

Note that by this notation only the linear term in the exponent is considered and that $C(m)$ can be multiplied by any polynomial function without changing the asymptotic expression. The same expression will thus hold if the number of numerical operations required by the algorithm is considered as a measure of the complexity. This is a direct consequence of the fact that the number of operations per node is bounded by a polynomial function [8].

The existence of γ however depends on the particular problem, that is the statistics of $\mathbf{H}$, $\mathbf{v}$ and $\bar{\mathbf{s}}$ and how these vary with problem size. To be more precise, the limit

$$\gamma = \lim_{m\to\infty} \frac{1}{m} \log_L C(m) \tag{45}$$

must exist for the above to be applicable. From Theorem 2 it is known that for any system that satisfies the assumptions made in the theorem, the limit, γ, is strictly positive if it does exist. To compute γ is not a trivial task and it seems unlikely that there will exist closed form expressions for γ for any but trivial systems.

However, the virtue of obtaining an exact value for γ lies in that it indicates for which problem sizes, m, sphere decoding is applicable. An interesting interpretation of γ is as a reduction of effective problem size. That is, γm may be considered the effective problem size when the sphere decoder is compared to full search, i.e. if $\gamma = 1/2$ the problem sizes which are considered feasible may be doubled if the sphere decoder is applied instead of full search.

For some choices of $\mathbf{H}$ the theory of large deviations [17] may be effectively applied to prove the existence of and numerically compute γ. This will be done in this section under the assumptions that the elements of $\mathbf{H}$ are independently drawn from a normal distribution. This is the same assumption as is made in [10] and [11]. For the remainder of this section it shall be assumed that $\mathbf{H}$ is an $m \times m$ matrix of independent normally distributed elements with zero mean and variance m^{-1}. The reason for this particular choice of variance is to make the SNR, ρ, independent of problem size. That is

$$\rho = \frac{\mathrm{E}\left\{\|\mathbf{H}\bar{\mathbf{s}}\|^2\right\}}{\mathrm{E}\left\{\|\mathbf{v}\|^2\right\}} = \frac{L^2 - 1}{12\sigma^2}. \tag{46}$$

Note that the choice of scaling $\bar{\mathbf{s}}$, $\mathbf{H}$ or $\mathbf{v}$ to keep the SNR fixed is completely arbitrary and does not affect the results. Also, throughout this section it shall be assumed that $r^2 = m\sigma^2$, see Remark 3. The main result of the section is given by Theorem 3.

The model for $\mathbf{H}$ used throughout this section is reasonable for some communication problems, especially in wireless communications under Rayleigh fading assumptions. Furthermore the model makes the mathematics tractable which allows further insight into the complexity of sphere decoding.

It will now be shown how to compute the limit of (45) from the statistics of the search depth, d. It is convenient to first normalize d by the problem size. Therefore let the normalized search depth, z_m, be

$$z_m = \frac{d_m}{m} \tag{47}$$

where the notation $d = d_m$ is used to emphasize the dependence on m. Under the above assumptions on $\mathbf{H}$ it can be shown that

$$z_m \xrightarrow{\mathrm{P}} \mu \tag{48}$$

for some μ as $m \to \infty$, i.e. z_m converges in probability to μ. By Jensen's inequality [16] this provides a lower bound on the expected complexity since

$$\mathrm{E}\left\{L^{d_m}\right\} = \mathrm{E}\left\{L^{mz_m}\right\} \geq L^{m\mathrm{E}\{z_m\}} \geq L^{(\mu-\varepsilon)m} \tag{49}$$

for large m and small $\varepsilon > 0$. Unfortunately this lower bound is not tight, not even asymptotically. Tighter bounds may be obtained by considering that for any a a lower bound on $\mathrm{E}\left\{L^{mz_m}\right\}$ is given by

$$\mathrm{E}\left\{L^{mz_m}\right\} \geq \Pr\left\{z_m \geq a\right\} L^{am}. \tag{50}$$

Since z_m converges in probability to μ it is known that

$$\lim_{m\to\infty} \Pr\left\{z_m \geq a\right\} = 0 \qquad \text{for} \qquad a > \mu. \tag{51}$$

It is possible to show that the convergence to this limit is exponential. That is for $a \geq \mu$,

$$\lim_{m\to\infty} \frac{1}{m} \ln \Pr\left\{z_m \geq a\right\} = -I_z(a) \tag{52}$$

for some function $I_z(a)$ which is called the rate function for z_m where some properties of the rate function are $I_z(a) \geq 0$ and $I_z(\mu) = 0$. Introducing $I_z(a)$ into (50) yields

$$\mathrm{E}\left\{L^{mz_m}\right\} \geq e^{-(I_z(a)+\varepsilon)m} L^{am} = L^{(a-(I_z(a)+\varepsilon)/\ln L)m} \tag{53}$$

for large m and arbitrary small $\varepsilon > 0$. By choosing $a = \mu$ the bound of (49) is obtained. Better bounds may be obtained by optimizing (53) over a. Using large deviation theory it can be shown that the best bound obtained by this optimization is asymptotically tight. To be precise, by Varadhan's integral Lemma [17, 2.12],

$$\mathrm{E}\left\{e^{mg(z_m)}\right\} \asymp e^{\xi m} \tag{54}$$

for any bounded continuous function $g(a)$ where

$$\xi = \sup_a (g(a) - I_z(a)). \tag{55}$$

Let $g(a) = a \ln L$ for $a \in [0,1]$, then

$$\mathrm{E}\left\{L^{mz_m}\right\} = \mathrm{E}\left\{e^{mz_m \ln L}\right\} = \mathrm{E}\left\{e^{mg(z_m)}\right\}. \tag{56}$$

The complexity of sphere decoding is thus asymptotically given by

$$C(m) \asymp L^{\gamma m} \tag{57}$$

where

$$\begin{aligned}\gamma &= \sup_a (g(a) - I_z(a))/\ln L \\ &= \sup_{a\in[\mu,1]} \{a - I_z(a)/\ln L\}.\end{aligned} \tag{58}$$

The last equality comes from the fact that $I_z(\mu) = 0$ and $I_z(a) \geq 0$. The rest of this section will be devoted to the, numerical, computation of $I_z(a)$ for $a \in [\mu, 1]$. This derivation will heavily rely on results from large deviation theory. The most useful concepts of this theory, as well as an introduction to the subject, can be found in [17]. Finally, before presenting the proofs, two useful definitions will be given.

Definition 3: Let u_k be a stochastic variable defined by

$$u_k = \frac{1}{k}\sum_{i=1}^{k} \frac{6(s_i - \bar{s}_i)^2}{L^2 - 1} \tag{59}$$

for $k \geq 1$ where s_i and $\bar{s}_i$ are independent and uniformly distributed on $\mathcal{D}_L$.

The normalization of u_k is chosen such that $\mathrm{E}\{u_k\} = 1$ independent of k and L. Also by the law of large numbers

$$u_k \xrightarrow{\text{a.s.}} 1 \qquad \text{as} \qquad k \to \infty, \tag{60}$$

where $\xrightarrow{\text{a.s.}}$ denotes almost sure convergence.

Definition 4: Let w_k be a normalized χ^2_k distributed stochastic variable, that is

$$w_k = \frac{1}{k}\sum_{i=1}^{k} y_i^2 \tag{61}$$

where y_i are i.i.d. $\mathcal{N}(0,1)$.

By the law of large numbers

$$w_k \xrightarrow{\text{a.s.}} 1 \qquad \text{as} \qquad k \to \infty. \tag{62}$$

Lemma 2: For $r^2 = m\sigma^2$, the probability that the normalized search depth, z_m, is larger than some $a \geq 0$ is given by

$$\Pr\{z_m \geq a\} = \Pr\left\{(2\rho a^2 u_k + a)w_k \leq 1 + \mathcal{O}(k^{-1})\right\} \tag{63}$$

where

$$k = \lceil am \rceil = am + \tau \tag{64}$$

for some roundoff error $\tau \in [0,1)$ and where $\mathcal{O}(k^{-1})$ tends to zero as k grows large.

Proof: First note that

$$\Pr\{z_m \geq a\} = \Pr\{d_m \geq am\} = \Pr\{d_m \geq \lceil am \rceil\} \tag{65}$$

where the last equality comes from the fact that d_m is integer valued. With k given as above

$$P = \Pr\{d_m \geq k\} = \Pr\left\{\|\mathbf{p}^m_{m-k+1}\|^2 \leq m\sigma^2\right\}. \tag{66}$$

The first step is to express the distribution of $\|\mathbf{p}^m_{m-k+1}\|^2$ in terms of u_k and w_k. To accomplish this, let $\mathbf{Q}_1 \in \mathbb{R}^{m\times m-k}$ be the first $m-k$ columns of $\mathbf{Q}$, and $\mathbf{Q}_2 \in \mathbb{R}^{m\times k}$ be the k last columns of $\mathbf{Q}$. Let equivalent definitions apply to $\mathbf{H}$. Using $\mathbf{Q}_2^{\mathrm{T}}\mathbf{H}_1 = \mathbf{0}$ which is a consequence of the QR-factorization, $\mathbf{p}^m_{m-k+1}$ can be written as

$$\mathbf{p}^m_{m-k+1} = \mathbf{Q}_2^{\mathrm{T}}(\mathbf{H}_2(\mathbf{s}^m_{m-k+1} - \bar{\mathbf{s}}^m_{m-k+1}) + \mathbf{v}^m_{m-k+1}). \tag{67}$$

The vector $\mathbf{q} \in \mathbb{R}^m$ given by

$$\mathbf{q} = \mathbf{H}_2(\mathbf{s}^m_{m-k+1} - \bar{\mathbf{s}}^m_{m-k+1}) + \mathbf{v}^m_{m-k+1} \tag{68}$$

is a rotationally invariant, normally distributed, vector with a variance of

$$\sum_{i=m-k+1}^{m} \frac{(s_i - \bar{s}_i)^2}{m} + \sigma^2 \tag{69}$$

per dimension. Even though $\mathbf{q}$ is not statistically independent of $\mathbf{Q}_2$ it is statistically independent of the space spanned by the columns of $\mathbf{Q}_2$. This can be seen by considering the fact that the columns of $\mathbf{Q}_2$ span the orthogonal complement of the space spanned by the columns of $\mathbf{Q}_1$, that $\mathbf{Q}_1$ is uniquely given by $\mathbf{H}_1$ and that $\mathbf{H}_1$ is independent of $\mathbf{H}_2$. A rigorous proof of this is given in [11].

Multiplication by $\mathbf{Q}_2^{\mathrm{T}}$ is equivalent to projection of $\mathbf{q}$ onto a linear subspace of dimension k and therefore, by introducing $\mathbf{y} \in \mathbb{R}^k$, the statistics of $\|\mathbf{p}^m_{m-k+1}\|^2$ are the same as the statistics of

$$\left(\sum_{i=1}^{k} \frac{s_i - \bar{s}_i)^2}{m} + \sigma^2\right)\|\mathbf{y}\|^2 \tag{70}$$

if $\mathbf{y}$ is a vector of i.i.d. $\mathcal{N}(0,1)$ distributed entries. Inserting the above, together with the definitions of u_k and w_k into (66) yields

$$P = \Pr\left\{\left(\frac{k(L^2-1)}{6m}u_k + \sigma^2\right)kw_k \leq m\sigma^2\right\}. \tag{71}$$

Using

$$m = \frac{k - \tau}{a}, \tag{72}$$

(46), and dividing both sides by $k\sigma^2/a$ yields, after some algebra,

$$P = \Pr\left\{(2\rho a^2 u_k + a)w_k \leq 1 + \mathcal{O}(k^{-1})\right\}. \tag{73}$$

The term $\mathcal{O}(k^{-1})$ is due to the roundoff error τ. However, this effect tends to zero as m, and k, grows large. Ignoring the error term, (73) is obtained by substituting $m = k/a$ into (71) and normalizing by $k\sigma^2/a$. This concludes the proof. ■

From Lemma 2 and the almost sure convergence of u_k and w_k to 1 it follows that, assuming $a > 0$ and noting that $k \to \infty$ as $m \to \infty$,

$$\lim_{m\to\infty} \Pr\{z_m \geq a\} = 1 \tag{74}$$

if

$$(2\rho a^2 + a) < 1 \Leftrightarrow a < \frac{\sqrt{8\rho + 1} - 1}{4\rho} \tag{75}$$

and that

$$\lim_{m\to\infty} \Pr\{z_m \geq a\} = 0 \tag{76}$$

if

$$(2\rho a^2 + a) > 1 \Leftrightarrow a > \frac{\sqrt{8\rho+1}-1}{4\rho}. \tag{77}$$

In other words μ is given by

$$\mu = \frac{\sqrt{8\rho+1}-1}{4\rho}. \tag{78}$$

Now, assuming that $a > \mu$, let

$$\mathcal{S}_a = \left\{(u,w) \mid (2\rho a^2 u + a)w \leq 1\right\}. \tag{79}$$

The set $\mathcal{S}_a$ is the set of all combinations of (u_k,w_k) such that the criterion of (73), ignoring $\mathcal{O}(k^{-1})$, is satisfied. It can be shown, by applying the Gärtner-Ellis Theorem [17, D.11] to the vector valued stochastic variable (u_k, w_k), that

$$\begin{aligned}&\lim_{k\to\infty} \frac{1}{k} \ln \Pr\{(u_k,w_k) \in \mathcal{S}_a\} \\ &= -\inf_{(u,w)\in\mathcal{S}_a} \{I_u(u) + I_w(w)\}\end{aligned} \tag{80}$$

where, for $u \leq 1$,

$$I_u(u) = -\lim_{k\to\infty} \frac{1}{k} \ln \Pr\{u_k \leq u\} \tag{81}$$

and, for $w \leq 1$,

$$I_w(w) = -\lim_{k\to\infty} \frac{1}{k} \ln \Pr\{w_k \leq w\}. \tag{82}$$

$I_u(u)$ and $I_w(w)$ are the rate functions corresponding to u_k and w_k respectively. A technical note, to aid rigorous verification of the above claim, is that since u_k and w_k are independent, the rate function of (u_k, w_k) is the sum of $I_u(u)$ and $I_w(w)$ and that equality holds in (80) since $I_u(u)$ and $I_w(w)$ are continuous on the boundary of $\mathcal{S}_a$. The intuition behind (80) is that among all the combinations of (u_k, w_k) that satisfy $(u_k, w_k) \in \mathcal{S}_a$ the probability of the event is dominated by most probable combination.

Since u_k and w_k are normalized sums of i.i.d. random variables their respective rate functions can be obtained by Chernoff's Theorem [17] as

$$I_u(u) = \sup_\theta \left[u\theta - \ln \mathrm{E}\left\{e^{\theta u_1}\right\}\right] \tag{83}$$

and

$$I_w(w) = \sup_\theta \left[w\theta - \ln \mathrm{E}\left\{e^{\theta w_1}\right\}\right] \tag{84}$$

respectively. Unfortunately there is no closed form expression for $I_u(u)$ for $L > 2$. It is however possible to numerically compute it by writing the expected value as a sum over all possible values of u_1 and numerically maximizing the obtained expression. For the special case where $L = 2$, u_1 is simply a Bernoulli random variable which takes on the values 0 and 2 with equal probability. The rate function for this case is given by [17]

$$I_u(u) = \frac{u}{2}\ln(\frac{u}{2}) + (1-\frac{u}{2})\ln(1-\frac{u}{2}) + \ln 2. \tag{85}$$

For $I_w(w)$ the supremum can be computed by first noting

$$\mathrm{E}\left\{e^{\theta w_1}\right\} = \mathrm{E}\left\{e^{\theta y_1^2}\right\} = \frac{1}{\sqrt{1-2\theta}} \tag{86}$$

where y_1 is $\mathcal{N}(0,1)$. Maximizing

$$w\theta + \ln\sqrt{1-2\theta} \tag{87}$$

over θ yields

$$I_w(w) = \frac{w-1}{2} - \frac{1}{2}\ln w \tag{88}$$

for $w > 0$.

An important note to make is that $I_u(u)$ and $I_w(w)$ are continuous and nonicreasing in the range $(0, 1]$. This follows since they are rate functions for sums of i.i.d. random variables [17]. For this reason the infimum of (80) will be taken for (u, w) satisfying

$$(2\rho a^2 u + a)w = 1, \quad u \leq 1, \quad w \leq 1. \tag{89}$$

In other words, by parametrization this reduces to an optimization problem over one variable. However, the above does not guarantee that the criterion function is unimodal. In fact, there are choices of a, L and ρ for which it is not. Extensive testing however suggest that this does not pose a serious problem, at least not for the examples given in Section VIII. In other words, the rather simple optimization routines used to produce the numerical examples in the paper seem to find the global optimum without any difficulty. Also, for most parameter values considered here, the function is unimodal.

Theorem 3: Let the channel matrix, $\mathbf{H} \in \mathbb{R}^{m\times m}$ consist of i.i.d. normally distributed entries, $\mathbf{v}$ be white Gaussian and the sent message $\bar{\mathbf{s}}$ be uniformly distributed on $\mathcal{D}_L^m$. Also let the sphere radius, r, satisfy $r^2 = m\sigma^2$. Then the complexity of sphere decoding, for fixed SNR, is

$$C(m) \asymp L^{\gamma m} \tag{90}$$

where

$$\gamma = \sup_{a\in[\mu,1]} \{a - I_z(a)/\ln L\} \tag{91}$$

and

$$I_z(a) = a \inf_{(u,w)\in\mathcal{S}_a} \{I_u(u) + I_w(w)\} \tag{92}$$

for $\mathcal{S}_a$, $I_u(u)$ and $I_w(w)$ given by (79), (83), and (88) respectively and where μ is given by (78).

Proof: Considering the previous discussion the only thing left to prove is the expression for $I_z(a)$. By Lemma 2 and the previous results

$$\begin{aligned}-I_z(a) &= \lim_{m\to\infty} \frac{1}{m} \ln \Pr\{z_m \geq a\} \\ &= \lim_{k\to\infty} \frac{a}{k-\tau} \ln \Pr\left\{(2\rho a^2 u_k + a)w_k \leq 1 + \mathcal{O}(k^{-1})\right\} \\ &= \lim_{k\to\infty} \frac{a}{k} \ln \Pr\left\{(2\rho a^2 u_k + a)w_k \leq 1\right\} \\ &= \lim_{k\to\infty} \frac{a}{k} \ln \Pr\{(u_k,w_k) \in \mathcal{S}_a\} \\ &= -a \inf_{(u,w)\in\mathcal{S}_a} \{I_u(u) + I_w(w)\}.\end{aligned} \tag{93}$$

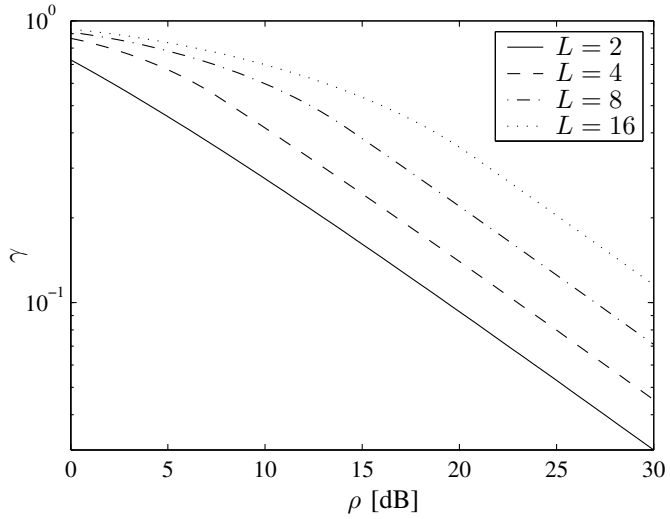

Fig. 2. The rate, γ, as a function of the SNR, ρ, for different values of constellation size, L.

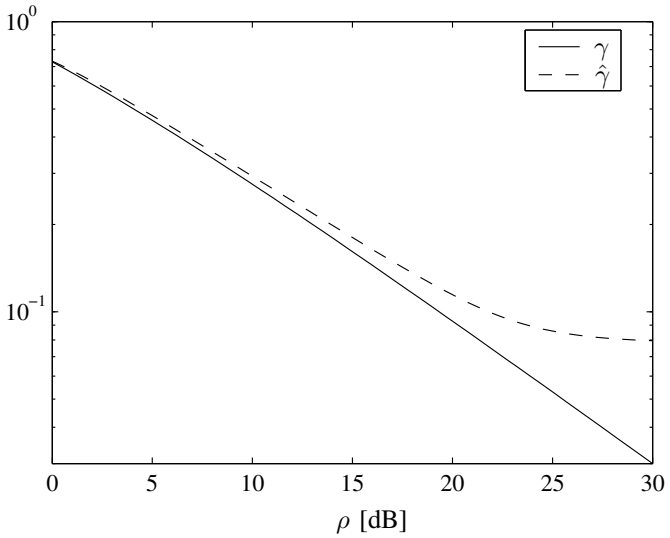

Fig. 3. The rate, γ, and $\hat{\gamma} = m^{-1} \log_L C(m)$ for $m = 80$ and $L = 2$ as a function of SNR, ρ.

The $\mathcal{O}(k^{-1})$ can be omitted at the second step since $I_u(u)$ and $I_w(w)$ are continuous functions on the boundary of $\mathcal{S}_a$. ■

Remark 3: In a real application the radius of the sphere would be chosen as

$$r^2 = \alpha m \sigma^2 \tag{94}$$

for some $\alpha > 1$ such that the probability of finding $\bar{\mathbf{s}}$ within the sphere is high. However, since

$$\Pr\left\{\|\mathbf{v}\|^2 \leq \alpha m \sigma^2\right\} \to 1 \qquad \text{as} \qquad m \to \infty \tag{95}$$

for any $\alpha > 1$ Theorem 3 remains true if $\alpha = \alpha(m)$ is chosen such that

$$\Pr\left\{\|\mathbf{v}\|^2 \leq \alpha m \sigma^2\right\} = 1 - \varepsilon \tag{96}$$

for some small ε, i.e. $\alpha \to 1$ as $m \to \infty$.

Remark 4: The expected complexity as a complexity measure can be further motivated by the results of this section. Assume that there is a $\gamma > 0$ such that $C(m)$ satisfies (44). Then for an arbitrarily small $\delta > 0$, by Markov's inequality [16],

$$\Pr\left\{N \geq L^{(\gamma+\delta)m}\right\} \leq \frac{\mathrm{E}\{N\}}{L^{(\gamma+\delta)m}} \tag{97}$$

where N is the number of nodes visited by the algorithm. Since $C = \mathrm{E}\{N\}$, there is an $\varepsilon < \delta$ and an M such that

$$\frac{\mathrm{E}\{N\}}{L^{(\gamma+\delta)m}} \leq \frac{L^{(\gamma+\varepsilon)m}}{L^{(\gamma+\delta)m}} = L^{(\varepsilon-\delta)m} \tag{98}$$

for all $m \geq M$. Thus,

$$\Pr\left\{N \geq L^{(\gamma+\delta)m}\right\} \to 0 \qquad \text{as} \qquad m \to \infty. \tag{99}$$

In other words, the probability that the algorithm is of substantially larger computational complexity than what is predicted by the expected value tend to zero as m grows large.

VIII. Examples

Theorem 3 of the previous section was used to compute the rate γ for various choices of L and ρ. The result can be seen in Figure 2. The curves show, not surprisingly, that the higher the SNR the lower the decoding complexity. This is a direct consequence of the choice of search radius as $r = \sqrt{\sigma^2 m}$ which decreases with increasing SNR. An intuitive explanation for this is that for high SNR the received points are tightly clustered and thus a smaller radius is required to ensure the same small probability of failure. Figure 2 also shows that, for this problem at high SNR, an additional 6 dB roughly leads to a reduction in γ by a factor 2. That is, by adding 6 dB to the SNR problems of twice the size are made computationally feasible. Experiments show that this reduction in γ, in general, is dependent on the particular structure of the problem.

As noted in the introduction, [10] introduces an expression for computing the exact expected complexity, $C(m)$, for the problem in Section VII. This expression gets increasingly cumbersome to compute when increasing the problem size m. Nonetheless, it can still be used, for moderate m, to compare the asymptotic results of this paper with the exact value of $C(m)$. That is, the asymptotic rate, γ, can be approximated by

$$\hat{\gamma} = m^{-1} \log_L C(m) \tag{100}$$

for some large m. The agreement between this $\hat{\gamma}$ and γ is shown in Figure 3 for $m = 80$. The correspondence between $L^{\gamma m}$ and $C(m)$ is illustrated by Figure 4 for the case of $\rho = 6$ dB and $L = 2$. It can be seen from Figure 4 that the asymptotic expression $L^{\gamma m}$ is applicable as an approximation of the true complexity, $C(m)$, for problems of quite moderate size. However, as indicated by Figure 3, larger m will be required for this to be true at high SNR.

From Figure 2 it can also be seen that the reduction in γ is smaller the larger the constellation size is. An intuitive explanation for this is that while the SNR is largely dominated

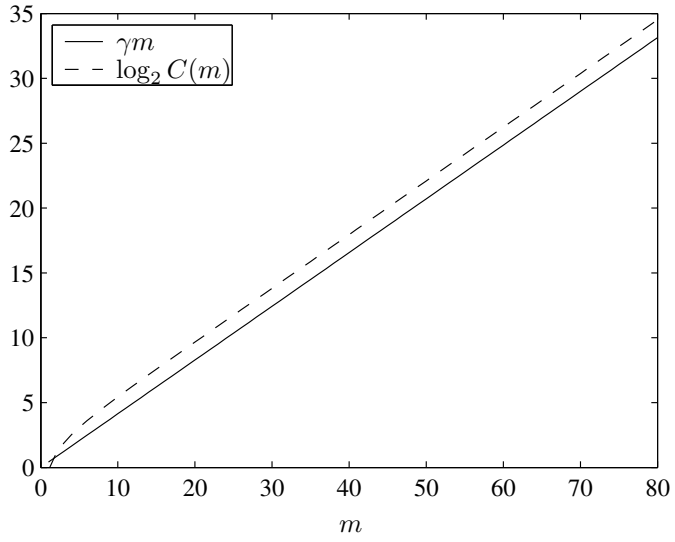

Fig. 4. Exponent, γm, versus $\log_L C(m)$ for $L = 2$ and $\rho = 6$ dB as a function of m.

by the constellation points at the edge of the constellation, the complexity is dominated by instances of $\mathbf{s}$ and $\bar{\mathbf{s}}$ that are close. With increasing constellation size the constellation will be more dense if the SNR is fixed.

IX. Conclusions

In this paper it has been shown, by obtaining a lower bound, that the expected complexity of sphere decoding applied to a large class of problems is exponential in the number of symbols jointly detected. This is mainly a consequence of the fact that to ensure a certain probability of finding a point within the sphere, the radius of the sphere must grow with the problem size. Proving the existence of such a lower bound turns out to be much easier than obtaining the exponential function that best, or asymptotically, describes the complexity. It is however, for a simplified problem structure, possible to compute the exact asymptotic rate of increase for the complexity using results from large deviation theory. By using the derived expressions the effect of SNR on the complexity could be studied more closely. It turns out, not unexpectedly, that the complexity is significantly reduced by an increased SNR.

A main point of this paper is that while the complexity of sphere decoding is exponential this does not necessarily mean that the algorithm is inefficient. As has been demonstrated the rate of the exponential function depends on the SNR of the problem and for high SNR it is quite small. Therefore sphere decoding can outperform polynomial time algorithms for many practical problems and is worth consideration. There will however not be any SNR for which the complexity can, for all problem sizes, be bounded by a polynomial function.

Acknowledgements

We would like to thank our colleagues Cristoff Martin and Prof. Mikael Skoglund for all their help and support while writing this paper.

References

[1] T. Kailath and H. V. Poor, "Detection of stochastic processes," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, Oct. 1998.
[2] S. Verdú, *Multiuser Detection*. Cambridge, U.K.: Cambridge Univ. Press, 1998.
[3] J. G. Proakis, *Digital Communications*, 3rd ed. McGraw-Hill, 1995.
[4] B. Hassibi and B. M. Hochwald, "High-rate codes that are linear in space and time," *IEEE Trans. Inform. Theory*, vol. 48, no. 7, pp. 1804–1824, June 2002.
[5] E. Viterbo and J. Boutros, "A universal lattice code decoder for fading channels," *IEEE Trans. Inform. Theory*, vol. 45, no. 5, pp. 1639–1642, July 1999.
[6] O. Damen, A. Chkeif, and J.-C. Belfiore, "Lattice code decoder for space-time codes," *IEEE Comm. Let.*, vol. 4, no. 5, pp. 161–163, May 2000.
[7] H. Vikalo and B. Hassibi, "Maximum-likelihood sequence detection of multiple antenna systems over dispersive channels via sphere decoding," *EURASIP Journal on Applied Signal Processing*, vol. 2002, no. 5, pp. 525–531, May 2002.
[8] U. Fincke and M. Pohst, "Improved methods for calculating vectors of short length in lattice, including a complexity analysis," *Mathematics of Computation*, vol. 44, no. 170, pp. 463–471, Apr. 1985.
[9] E. Agrell, T. Eriksson, A. Vardy, and K. Zeger, "Closest point search in lattices," *IEEE Trans. Inform. Theory*, vol. 48, no. 8, pp. 2201–2214, Aug. 2002.
[10] B. Hassibi and H. Vikalo, "On the expected complexity of integer least-squares problems," in *Proc. IEEE ICASSP'02*, vol. 2, May 2002, pp. 1497–1500.
[11] ——, "Maximum-likelihood decoding and integer least-squares: The expected complexity," in *Multiantenna Channels: Capacity, Coding and Signal Processing*, J. Foschini and S. Verdu, Eds. Amer. Math. Soc., 2003.
[12] G. L. Nemhauser and L. A. Wolsey, *Integer and Combinatorial Optimization*. Wilie-Interscience, 1988.
[13] A. V. Aho, J. E. Hopcroft, and J. D. Ullman, *The Design and Analysis of Computer Algorithms*. Addison-Wesley Pub Co, 1974.
[14] S. Verdú, "Computational complexity of multiuser detection," *Algorithmica*, vol. 4, pp. 303–312, 1989.
[15] O. Damen, K. Abed-Meraim, and M. S. Lemdani, "Further results on the sphere decoder," in *Proc. ISIT'01*, June 2001, p. 333.
[16] S. Ross, *A First Course in Probability*, 2nd ed. Macmillan Publishing Company, 1984.
[17] A. Shwarts and A. Weiss, *Large Deviations for Performance Analysis*. Chapman and Hall, 1995.