The complexity of sphere decoding perfect codes under a vanishing gap to ML performance

Joakim Jaldén
ACCESS Linnaeus Center - Signal Processing Lab
KTH - Royal Institute of Technology, Stockholm, Sweden
Email: jalden@kth.se

Petros Elia
Mobile Communications Department
EURECOM, Sophia Antipolis, France
Email: elia@eurecom.fr

***Abstract*—We consider the complexity of the sphere decoding (SD) algorithm when decoding a class of full rate space-time block codes that are optimal, over the quasi-static MIMO channel, with respect to the diversity-multiplexing tradeoff (DMT). Towards this we introduce the *SD complexity exponent* which represents the high signal-to-noise ratio (SNR) exponent of the tightest run-time complexity constraints that can be imposed on the SD algorithm while maintaining arbitrarily close to maximum likelihood (ML) performance. Similar to the DMT exposition, our approach naturally captures the dependence of the SD algorithm's computational complexity on the codeword density, code size and channel randomness; providing simple closed form solutions in terms of the system dimensions and the multiplexing gain.**

I. Introduction

The sphere decoding (SD) algorithm [1], [**?**] has arguably become the de-facto method for optimal and near optimal decoding of practically implementable space-time block codes. By choosing the sphere decoder search radius adaptively and allowing for an unbounded worst-case run-time the SD algorithm can provide an exact implementation of the maximum likelihood (ML) decoder at a reduced *average* complexity. However, a more realistic approach to system design than allocating unbounded computational reserves is to impose a hard run-time limitation on the algorithm – declare decoding outages when this limit is not met – and select the run-time limit so that the gap to ML performance is acceptable. This naturally raises the intriguing question of how aggressively one can time-limit the SD algorithm without seriously degrading the decoder performance.

While this question is hard to answer in general, or even ask in a rigorously meaningful way, we show that by following [3] and considering the decoding of a sequence of codes in the high signal-to-noise ratio (SNR) limit, not only can the question be made rigorous: It also admits surprisingly simple explicit answers. Drawing from the diversity-multiplexing tradeoff (DMT) setting, which has also recently been applied to concisely describe the reliability of reduced complexity receivers [4], [5], [6], we introduce the *SD complexity exponent* as a measure of complexity that describes the SNR exponent of the computational reserves required for decoding with the SD algorithm while guaranteeing a vanishing gap to ML performance.

This conference paper represents the shortened version of a larger work [7] that computes the SD complexity exponent for the decoding of a large class of full rate linear dispersion lattice space-time codes. However, due to space constraints and in order to accentuate the main points we shall herein restrict our attention to a class of DMT optimal full rate threaded minimum delay codes that includes the codes in [**?**], [**?**] and the perfect codes in [8], [9]. Extending recent work on the SD algorithm [10] we here also consider the effect of the code, as well as the constellation boundary; something which is essential for achieving arbitrary close to ML performance.

A. System model and space-time codes

We consider the standard Rayleigh fading $n_{\mathrm{T}} \times n_{\mathrm{R}}$ quasi-static point-to-point MIMO channel model with coherence-time T given by

$$\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{X} + \boldsymbol{W} \tag{1}$$

where $\boldsymbol{X} \in \mathbb{C}^{n_{\mathrm{T}} \times T}$, where $\boldsymbol{Y} \in \mathbb{C}^{n_{\mathrm{R}} \times T}$, and where $\boldsymbol{W} \in \mathbb{C}^{n_{\mathrm{R}} \times T}$ denote the transmitted space-time block codeword, the block of received signals, and the additive spatially and temporally white Gaussian noise. The channel gains $\boldsymbol{H} \in \mathbb{C}^{n_{\mathrm{R}} \times n_{\mathrm{T}}}$ are assumed to be i.i.d. circularly symmetric complex Gaussian (i.e., Rayleigh fading) and constant over the duration of the transmission, and we assume that $n_{\mathrm{R}} \geq n_{\mathrm{T}}$. The transmitted codewords $\boldsymbol{X}$ are drawn uniformly from a codebook $\mathcal{X}$ where

$$\mathrm{E}\{\|\boldsymbol{X}\|_{\mathrm{F}}^2\} = \frac{1}{|\mathcal{X}|} \sum_{\boldsymbol{X} \in \mathcal{X}} \|\boldsymbol{X}\|_{\mathrm{F}}^2 = \rho T \,, \tag{2}$$

so that ρ takes on the interpretation of an average SNR.

By vectorizing the model in (1) it follows that

$$\boldsymbol{y} = (\boldsymbol{I}_T \otimes \boldsymbol{H})\boldsymbol{x} + \boldsymbol{w} \tag{3}$$

where $\boldsymbol{y} = \mathrm{vec}(\boldsymbol{Y})$, where $\boldsymbol{x} = \mathrm{vec}(\boldsymbol{X})$, where $\boldsymbol{w} = \mathrm{vec}(\boldsymbol{W})$, and where $\mathrm{vec}(\cdot)$ denotes the operation whereby the columns of the argument are stacked to form a vector. For the class of (complex) lattice codes considered here, the codewords take the form

$$\boldsymbol{x} = \theta \boldsymbol{G}\boldsymbol{s} \tag{4}$$

where θ regulates the transmit power, where $\boldsymbol{G} \in \mathbb{C}^{\kappa \times \kappa}$ is the full rank generator matrix of the code, and where $\boldsymbol{s}$ belongs to the Gaussian integers lattice $\mathbb{Z}[i]^{\kappa}$. We make the restriction to DMT optimal full rate threaded minimum delay codes, where

$n_{\mathrm{T}} = T = n$ and $\kappa = n^2$, and where the data symbols are drawn from the QAM-like constellation

$$\mathbb{S}_\eta \triangleq \{ s \,|\, \Re(s), \Im(s) \in \mathbb{Z} \cap [-\eta, \eta] \,\} \subset \mathbb{Z}[i]^\kappa \tag{5}$$

where η regulates the size of the constellation, and thereby also the rate of the code. We omit the shift (translation) of the Gaussian integer lattice that is typically present in QAM constellations as such shifts have no effect on our main results. Choosing[1] $\eta \doteq \rho^{\frac{r}{2n}}$ yields a sequence of codes with *multiplexing gain*

$$r \triangleq \lim_{\rho\to\infty} \frac{1}{n} \frac{\log|\mathcal{X}|}{\log\rho}\,, \tag{6}$$

and the power constraint in (2) mandates that $\theta^2 \doteq \rho^{1-\frac{r}{n}}$. At a given multiplexing gain r the probability of an ML decoding error is $P_e \doteq \rho^{-d(r)}$ where $d(r)$ is the *diversity gain*. When the sequence of codes generated by $\boldsymbol{G}$ is DMT optimal, as is assumed herein, $d(r)$ is maximized and equal to the outage exponent [3] of the channel in (1). In addition to the assumptions listed above, we also need to assume that the threaded code-structure of the codes in [8], [9] applies in order to complete the analysis, see [7] for details. We will also hereafter for brevity use the word *code* both when refereing to a particular code and the sequence of codes generated by a single generator matrix $\boldsymbol{G}$.

B. The Sphere Decoding Algorithm

Combining (3) and (4) yields the equivalent data model

$$\boldsymbol{y} = \boldsymbol{M}\boldsymbol{s} + \boldsymbol{w} \tag{7}$$

where the combined code-channel generator matrix $\boldsymbol{M}$ is

$$\boldsymbol{M} \triangleq \theta(\boldsymbol{I}_T \otimes \boldsymbol{H})\boldsymbol{G} \in \mathbb{C}^{n_{\mathrm{R}}T\times\kappa}\,. \tag{8}$$

Let $\boldsymbol{QR} = \boldsymbol{M}$ be the thin QR factorization of $\boldsymbol{M}$, where $\boldsymbol{Q}^{\mathrm{H}}\boldsymbol{Q} = \boldsymbol{I}$ and $\boldsymbol{R}$ is upper triangular. The coherent ML decoder for $\boldsymbol{s}$ can then be expressed as

$$\hat{\boldsymbol{s}}_{\mathrm{ML}} = \arg\min_{\hat{\boldsymbol{s}}\in\mathbb{S}_\eta^\kappa} \|\boldsymbol{r} - \boldsymbol{R}\hat{\boldsymbol{s}}\|^2\,, \tag{9}$$

where $\boldsymbol{r} = \boldsymbol{Q}^{\mathrm{H}}\boldsymbol{y}$. The sphere decoding algorithm solves (9) by enumerating all codeword hypotheses $\hat{\boldsymbol{s}}$ that satisfy

$$\|\boldsymbol{r} - \boldsymbol{R}\hat{\boldsymbol{s}}\|^2 \leq \xi^2 \tag{10}$$

for a given *search radius* $\xi > 0$. The algorithm works by identifying partial symbol vectors $\hat{\boldsymbol{s}}_k$, where $\hat{\boldsymbol{s}}_k$ contains the last k elements of $\hat{\boldsymbol{s}}$, for which

$$\|\boldsymbol{r} - \boldsymbol{R}\hat{\boldsymbol{s}}\|^2 \geq \|\boldsymbol{r}_k - \boldsymbol{R}_k\hat{\boldsymbol{s}}_k\|^2 > \xi^2\,, \tag{11}$$

where $\boldsymbol{r}_k$ denotes the last k components of $\boldsymbol{r}$, where $\boldsymbol{R}_k$ is the lower right corner of $\boldsymbol{R}$, and where the first inequality follows as $\boldsymbol{R}$ is upper triangular. Upon identifying these $\hat{\boldsymbol{s}}_k$, the algorithm proceeds to simultaneously reject all vectors $\hat{\boldsymbol{s}}$ that share the same $\hat{\boldsymbol{s}}_k$. The sphere decoding algorithm can be viewed as a branch and bound algorithm over a regular tree of height κ with $|\mathbb{S}_\eta|$ branches extending from each node, where each $\hat{\boldsymbol{s}}_k$ corresponds to a node at layer k, where the root node is at layer $k = 0$, and where (11) is used to prune sub-trees from the search. The set of remaining, unpruned, nodes at level k is given by

$$\mathcal{N}_k \triangleq \{\, \hat{\boldsymbol{s}}_k \in \mathbb{S}_\eta^k \mid \|\boldsymbol{r}_k - \boldsymbol{R}_k\hat{\boldsymbol{s}}_k\|^2 \leq \xi^2 \,\}\,. \tag{12}$$

It is common (cf. [10]) to use this total number of unpruned (or visited) nodes

$$N = \sum_{k=1}^{\kappa} |\mathcal{N}_k| \tag{13}$$

as a measure of the complexity of the algorithm. It can also be shown that the total number of floating point operations (flops) required by the algorithm deviates from N by at most by some dimension-dependent multiplicative constants, that are independent of ρ. It is also worth noting that whenever $\mathcal{N}_\kappa \neq \emptyset$, the sphere decoder recovers the ML decision.

We assume that the sphere radius ξ is a deterministic function of ρ that satisfies $\xi \doteq \rho^0$. This assumption may easily be motivated by noting that $\boldsymbol{r} - \boldsymbol{R}\boldsymbol{s} = \boldsymbol{Q}^{\mathrm{H}}\boldsymbol{w}$ and $\mathrm{P}(\|\boldsymbol{Q}^{\mathrm{H}}\boldsymbol{w}\|^2 > z\log\rho) \doteq \rho^{-z}$ for $z > 0$, i.e., by choosing $z > d(r)$ the probability of excluding the transmitted codeword from $\mathcal{N}_\kappa$ is made arbitrarily small (and vanishing) in relation to the probability of decoding error while satisfying $\xi = z\log\rho \,\dot{\leq}\, \rho^0$. At the same time, for any fixed ξ independent of ρ, the probability that $\boldsymbol{s}$ is not found is independent of ρ, which is clearly undesirable when implementing DMT optimal decoding, and which implies that ξ should satisfy $\xi \,\dot{\geq}\, \rho^0$. Finally, it should be stressed that while we for simplicity consider a non-random search radius, it is in fact shown in [7] that the SD complexity exponent defined next cannot be reduced by adaptive radius updates, e.g., those used in the Schnorr-Euchner SD implementation [**?**].

II. The SD complexity exponent

To see what may be a reasonable scale of interest for measuring complexity it is illustrative to note that at a multiplexing gain of r the codebook has a cardinality of $|\mathcal{X}| \doteq \rho^{rn}$, and the SD algorithm needs to find at least $1 \doteq \rho^0$ codeword. This motivates us to measure complexity in terms of a power of the SNR, i.e., ρ^x for $0 \leq x \leq rn$, and arrive at the following definition for the SD complexity exponent.

Definition 1: The *SD complexity exponent* $c(r)$ is given by

$$c(r) \;\triangleq\; \inf\{x \,|\, \Psi(x) > d(r)\} \tag{14}$$

where $d(r)$ is the diversity gain of the code at multiplexing gain r, and where

$$\Psi(x) \triangleq -\lim_{\rho\to\infty} \frac{\log\mathrm{P}\left(N \geq \rho^x\right)}{\log\rho} \tag{15}$$

for $x \geq 0$ and for N being the complexity as given by (13).

In order to see the operational significance of the SD complexity exponent, note that for any $x > c(r)$ we have by

[1] As in [3] we use the $\doteq$ notation to denote equality in the exponent where $f(\rho) \doteq \rho^a$ denotes $\lim_{\rho\to\infty} \log f(\rho)/\log\rho = a$. The symbols $\dot{\leq}$, $\dot{\geq}$, $\dot{<}$ and $\dot{>}$ are similarly defined.

definition that $\mathrm{P}\left(N \geq \rho^{x}\right) \dot{<} \rho^{-d(r)}$, i.e., the probability that the SD complexity exceeds ρ^x for $x > c(r)$ vanishes strictly faster at high SNR than the minimum probability of decoding error. This implies that if we were to put a run-time limit of ρ^x on the sphere decoder – and declare a decoding outage whenever this limit is not met – it would cause a vanishing degradation of the overall probability of error which would at high SNR be completely dominated by ML errors. Further, as the sphere decoder search radius ξ can be selected such that the probability of $\mathcal{N}_\kappa = \emptyset$ can also be made arbitrarily small in relation to the ML error probability, it follows that it is possible to implement a decoder based on the SD algorithm coupled with a time-out policy that guarantees a worst-case complexity of ρ^x for any $x > c(r)$, and still obtain a vanishing SNR gap to the ML decoder. This is however not possible for any $x < c(r)$. Thus, up to the high SNR exponent, $c(r)$ quantifies the smallest computational reserves that must be designed for in order to achieve both maximum diversity decoding with the SD algorithm as well as a vanishing gap to the ML decoder. Fortunately, $c(r)$ can also be given explicitly as is shown next.

Theorem 1: The SD complexity exponent for decoding any DMT optimal $n \times n$ threaded full rate code is

$$c(r) = r(n - \lfloor r \rfloor - 1) + \big(n\lfloor r \rfloor - r(n-1)\big)^{+} \tag{16}$$

for $0 \leq r \leq n$, where $\lfloor r \rfloor$ denotes the largest integer lower than or equal to r and where $(\cdot)^+ = \max(\cdot, 0)$. For integer values of r, i.e., when $r = k$, the expression in (16) becomes

$$c(k) = k(n-k)\,. \tag{17}$$

The proof of Theorem 1 is given in full in [7]. However, in order to illustrate the main concepts behind the proof we provide a partial proof in Appendix A and B, establishing the upper bound $c(r) \leq r(n - \lfloor r \rfloor - 1) + (n\lfloor r \rfloor - r(n-1))^+$. We hasten to add that while the partial proof offered in this paper makes no assumptions on $\boldsymbol{G}$ other than that it is square and full rank, the threaded structure of the codes in [**?**], [**?**], [8], [9] is explicitly used in the construction of the lower bound on $c(r)$ in [7]. However, this also implies that (16) provides an upper bound for the SD complexity exponent when decoding any DMT optimal full rate minimum delay code, regardless of its structure.

The SD complexity exponent $c(r)$ in (16) is shown in Fig. 1 for $n = 2, \ldots, 6$. Here, a rather surprising characteristic of $c(r)$ may be observed. At low multiplexing gains, $c(r)$ tends to increase with r while the opposite is true for high multiplexing gains. Consequently, the complexity is the highest at intermediate multiplexing gains and not at high multiplexing gains as may have been expected. The explanation to this somewhat counterintuitive result is that at high multiplexing gains, also the probability of error is higher and therefore the SD algorithm can tolerate a larger number of decoding outages without significant performance degradation. Another way to see this is that at high data-rates, only well conditioned channels support these rates and consequently, the SD algorithm only has to deal with well conditioned instances of (9) in order to provide close to ML performance, thus benefiting from the fact that good channels are likely to introduce reduced complexity. Another observation to be made is that $c(r)$ is independent of the number of receive antennas n_{R} as long as $n_{\mathrm{R}} \geq n_{\mathrm{T}}$. Finally, for any r we have that $c(r) < rn$ where rn is the SNR exponent of $|\mathcal{X}|$, implying that SD requires strictly less complexity reserves than the full search.

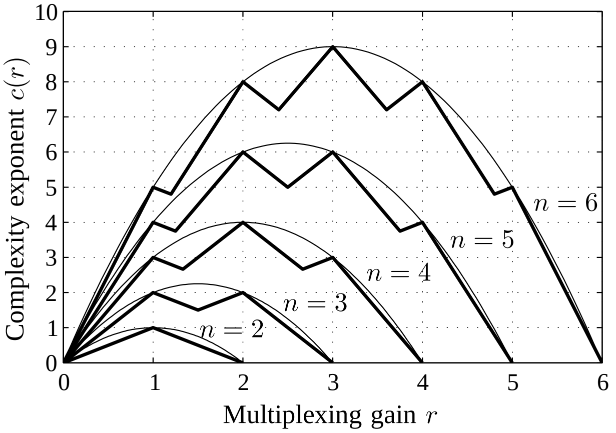

Fig. 1. The SD complexity exponent $c(r)$ in (16) for the decoding of threaded minimum delay DMT optimal codes with $n_{\mathrm{T}} = T = n$ for $n = 2, \ldots, 6$. The complexity exponent is illustrated by the bold lines. The thin lines show the quadratic function $r(n-r)$, cf. (17).

To get an intuitive understanding of the result of Theorem 1, and in particular (17), it is illustrative to consider a heuristic argument involving low rank channel matrices $\boldsymbol{H}$. As noted in [3], the typical outages at integer multiplexing gains $r = k$ are caused by $\boldsymbol{H}$ that are close to the set of rank k matrices, i.e., that have $n-k$ small singular values. If we for the purpose of the illustration assume that $\boldsymbol{H}$ has rank k, it follows that $\boldsymbol{I}_T \otimes \boldsymbol{H}$, and $\boldsymbol{M}$, have rank nk as $n = T$, and consequently a null-space of dimension $n(n-k)$. This implies that the $n(n-k) \times n(n-k)$ lower right block of $\boldsymbol{R}$, $\boldsymbol{R}_{n(n-k)}$, is[2] identically equal to zero, and the sphere decoder pruning criteria become totally ineffective up to and including layer $n(n-k)$. As the size of $\mathbb{S}_\eta$ is $|\mathbb{S}_\eta| \doteq \rho^{\frac{k}{n}}$ for $r = k$, the number of nodes searched at layer $n(n-k)$ of the SD search tree is therefore $|\mathbb{S}_\eta^{n(n-k)}| \doteq \rho^{k(n-k)}$ (cf. (17)). In order to ensure close to optimal performance, the sphere decoder must be able to decode for $\boldsymbol{H}$ where $n-k$ singular values are close to zero. However, channels with even more singular values close to zero occur with a probability that is small in relation to the probability of ML decoder errors (or channel outages), and can thus be safely ignored by the decoder without significant degradation in performance.

III. Conclusion

We have introduced the SD complexity exponent as a measure of SD complexity when the algorithm is applied to decode DMT optimal threaded space-time codes. The SD

[2] This also requires that the first nk columns of $\boldsymbol{R}$ are linearly independent. In fact, the rigorous treatment of this technical detail is largely responsible for much of the difficulty in establishing the lower bounds on $c(r)$ in [7].

complexity exponent naturally incorporates factors such as codeword density, codebook size, SNR, and channel fading into a single scalar quantity $c(r)$ that is expressed as a function of the multiplexing gain r of the code. To date, $c(r)$ also asymptotically represents the smallest known complexity required for arbitrarily close to optimal decoding of, e.g., the DMT optimal codes in [8], [9]. The simplicity of the expressions obtained for $c(r)$ allows for quickly assessing the rate, reliability, and complexity characteristics of communication over the quasi-static MIMO channel using such coding and decoding schemes, which should prove useful both for insight into the algorithm as well as for system design.

IV. Acknowledgments

The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement no. 228044, from the Swedish Foundation for Strategic Research (SSF) / grant ICA08-0046, from FP7/2007-2013 grant agreement no. 257616 (CONECT), and from the Mitsubishi RD project Home-eNodeBS.

Appendix A
Partial proof of Theorem 1

To obtain a bound on the number of nodes visited by the SD algorithm at layer k, i.e., $N_k \triangleq |\mathcal{N}_k|$ with $\mathcal{N}_k$ defined in (12), we consider the following lemma, proven in Appendix B.

Lemma 1: Let $\mathcal{E} \subset \mathbb{R}^m$ be the ellipsoidal set given by

$$\mathcal{E} \triangleq \{\boldsymbol{d} \in \mathbb{R}^m \mid \|\boldsymbol{c} - \boldsymbol{D}\boldsymbol{d}\|^2 \le \xi^2\} \tag{18}$$

where $\boldsymbol{D} \in \mathbb{R}^{m\times m}$ and $\boldsymbol{c} \in \mathbb{R}^m$. Let $\mathcal{B} \subset \mathbb{R}^m$ be the hypercube given by

$$\mathcal{B} \triangleq \{\boldsymbol{d} \in \mathbb{R}^m \mid \|\boldsymbol{d}\|_\infty \le \eta\}\,, \tag{19}$$

where $\|\boldsymbol{d}\|_\infty \triangleq \max(|d_1|, \ldots, |d_m|)$. Then, the number of integer points contained in the intersection of $\mathcal{E}$ and $\mathcal{B}$ is upper bounded as

$$|\mathcal{E} \cap \mathcal{B} \cap \mathbb{Z}^m| \le \prod_{i=1}^{m} \left[\sqrt{m} + \min\left(\frac{2\xi}{\sigma_i(\boldsymbol{D})}\,,\, 2\sqrt{m}\eta\right)\right]\,, \tag{20}$$

where $\sigma_m(\boldsymbol{D}) \ge \ldots \ge \sigma_1(\boldsymbol{D})$ denotes the ordered singular values of $\boldsymbol{D}$.

In order to apply Lemma 1 to obtain the size of $\mathcal{N}_k$, note that $\hat{\boldsymbol{s}}_k \in \mathcal{N}_k$ if and only if $\|\underline{\boldsymbol{r}}_k - \underline{\boldsymbol{R}}_k\hat{\underline{\boldsymbol{s}}}_k\|^2 \le \xi^2$, $\|\hat{\underline{\boldsymbol{s}}}_k\|_\infty \le \eta$, and $\hat{\underline{\boldsymbol{s}}}_k \in \mathbb{Z}^{2k}$, where

$$\underline{\boldsymbol{r}}_k = \begin{bmatrix}\Re(\boldsymbol{r}_k)\\ \Im(\boldsymbol{r}_k)\end{bmatrix} \in \mathbb{R}^{2k},\ \underline{\boldsymbol{R}}_k = \begin{bmatrix}\Re(\boldsymbol{R}_k) & -\Im(\boldsymbol{R}_k)\\ \Im(\boldsymbol{R}_k) & \Re(\boldsymbol{R}_k)\end{bmatrix} \in \mathbb{R}^{2k\times 2k},$$

and where

$$\hat{\underline{\boldsymbol{s}}}_k = \begin{bmatrix}\Re(\hat{\boldsymbol{s}}_k)\\ \Im(\hat{\boldsymbol{s}}_k)\end{bmatrix} \in \mathbb{Z}^{2k}\,.$$

Consequently, applying Lemma 1 with $m = 2k$, with $\boldsymbol{d} = \underline{\boldsymbol{r}}_k$, and with $\boldsymbol{D} = \underline{\boldsymbol{R}}_k$ yields

$$N_k \triangleq |\mathcal{N}_k| \le \prod_{i=1}^{2k} \left[\sqrt{2k} + \min\left(\frac{2\xi}{\sigma_i(\underline{\boldsymbol{R}}_k)}\,, 2\sqrt{2k}\eta\right)\right]\,. \tag{21}$$

The singular values of $\underline{\boldsymbol{R}}_k$ are the same as those of $\boldsymbol{R}_k$ albeit with a multiplicity of 2, i.e., $\sigma_i(\underline{\boldsymbol{R}}_k) = \sigma_{\iota_2(i)}(\boldsymbol{R}_k)$ where $\iota_m(i) \triangleq \lceil i/m \rceil$, and where $\lceil x \rceil$ denotes the smallest integer larger than or equal to x. By the interlacing property of singular values of sub-matrices [12] we have that $\sigma_i(\boldsymbol{R}_k) \ge \sigma_i(\boldsymbol{R})$ for $i = 1, \ldots, k$ and therefore $\sigma_i(\underline{\boldsymbol{R}}_k) \ge \sigma_{\iota_2(i)}(\boldsymbol{R})$ for $i = 1, \ldots, 2k$. The singular values of $\boldsymbol{R}$ are the same as those of $\boldsymbol{M}$, as $\boldsymbol{Q}\boldsymbol{R} = \boldsymbol{M}$ with $\boldsymbol{Q}$ being unitary. The singular values of $\boldsymbol{M}$ satisfy $\sigma_i(\boldsymbol{M}) \ge \theta\gamma\sigma_i(\boldsymbol{I}_T \otimes \boldsymbol{H})$ where $\gamma \triangleq \sigma_1(\boldsymbol{G})$. Further, the singular values of $\boldsymbol{I}_T \otimes \boldsymbol{H}$ are the same as those of $\boldsymbol{H}$ albeit with a multiplicity of $n = T$ [12]. This implies that $\sigma_i(\underline{\boldsymbol{R}}_k) \ge \theta\gamma\sigma_{\iota_{2n}(i)}(\boldsymbol{H})$ for $i = 1, \ldots, 2k$, and by (21) that

$$N_k \le \prod_{i=1}^{2k} \left[\sqrt{2k} + \min\left(\frac{2\xi}{\theta\gamma\sigma_{\iota_{2n}(i)}(\boldsymbol{H})}\,, 2\sqrt{2k}\eta\right)\right]\,, \tag{22}$$

which upper bounds the number of nodes visited at layer k in terms of the singular values of $\boldsymbol{H}$.

Next, and following [3], we introduce the notion of *singularity levels* $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_n]$ defined by

$$\alpha_i \triangleq -\frac{\log \sigma_i(\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H})}{\log \rho} \quad \Leftrightarrow \quad \sigma_i(\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H}) = \rho^{-\alpha_i} \tag{23}$$

where $\alpha_1 \ge \ldots \ge \alpha_n$, or equivalently $\sigma_i(\boldsymbol{H}) = \rho^{-\frac{1}{2}\alpha_i}$ for $i = 1, \ldots, n$. Applied to (22) this yields

$$\sqrt{2k} + \min\left(\frac{2\xi}{\theta\gamma\sigma_{\iota_{2n}(i)}(\boldsymbol{H})}\,, 2\sqrt{2k}\eta\right) \doteq \rho^{\nu_i(\boldsymbol{\alpha})}$$

where

$$\nu_i(\boldsymbol{\alpha}) \triangleq \min\left(\frac{r}{2n} - \frac{1}{2} + \frac{1}{2}\alpha_{\iota_{2n}(i)}\,, \frac{r}{2n}\right)^{+}\,, \tag{24}$$

as $\gamma \doteq \rho^0, \theta \doteq \rho^{\frac{1}{2} - \frac{r}{2n}}$, $\xi \doteq \rho^0$, and $\eta \doteq \rho^{\frac{r}{2n}}$. This implies that

$$N_k \mathrel{\dot{\le}} \prod_{i=1}^{2k} \rho^{\nu_i(\alpha)} = \rho^{\sum_{i=1}^{2k} \nu_i(\boldsymbol{\alpha})}\,. \tag{25}$$

However, as $\nu_i(\boldsymbol{\alpha}) \ge 0$ for $i = 1, \ldots, 2\kappa$ it follows that

$$N = \sum_{k=1}^{\kappa} N_k \mathrel{\dot{\le}} \sum_{k=1}^{\kappa} \rho^{\sum_{i=1}^{2k} \nu_i(\boldsymbol{\alpha})} \mathrel{\dot{\le}} \rho^{\upsilon(\boldsymbol{\alpha})} \tag{26}$$

where

$$\upsilon(\boldsymbol{\alpha}) \triangleq \sum_{i=1}^{2\kappa} \nu_i(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \min\big(r - n(1-\alpha_i)\,, r\big)^{+}\,, \tag{27}$$

i.e., $\upsilon(\boldsymbol{\alpha})$ provides an asymptotic upper bound on N in terms of the singularity levels of $\boldsymbol{H}$.

For a DMT optimal code at multiplexing gain r, *typical errors* are caused by singularity levels in the outage set $\mathcal{A}(r) \triangleq \{\boldsymbol{\alpha} \mid \sum_{i=1}^{n}(1-\alpha_i)^+ < r\}$ [3]. In particular, it holds under the i.i.d. Rayleigh fading assumption that $\mathrm{P}\left(\boldsymbol{\alpha} \in \mathcal{A}(r)\right) \doteq \rho^{-d(r)}$, and that $\mathrm{P}\left(\boldsymbol{\alpha} \notin \mathbb{R}_+^n\right) \doteq \rho^{-\infty}$. If we let $\bar{c}(r) \triangleq \sup_{\boldsymbol{\alpha} \in \bar{\mathcal{A}}'(r)} \upsilon(\boldsymbol{\alpha})$ where

$$\bar{\mathcal{A}}'(r) \triangleq \mathbb{R}_+^n \backslash \mathcal{A}(r) = \mathbb{R}_+^n \cap \left\{\boldsymbol{\alpha} \,\middle|\, \sum_{i=1}^{n}(1-\alpha_i)^+ \ge r\right\}$$

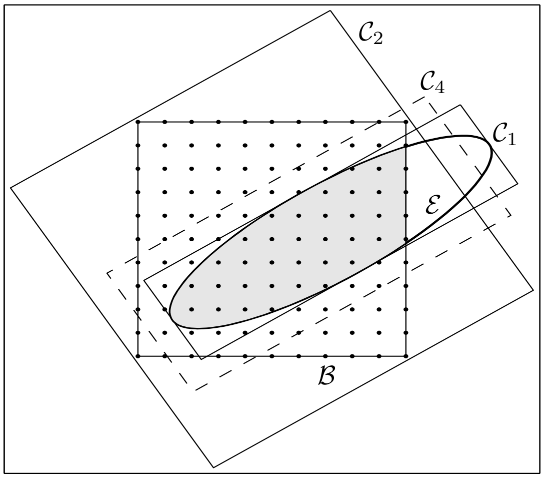

Fig. 2. Illustration of the proof of (20) in Lemma 1 in the case of $n = 2$. The lemma provides an upper bound on the number of integer points within the shaded area, corresponding to the intersection of the ellipsoid and the constellation boundary.

we have that $\mathrm{P}\left(N \geq \rho^{\bar{c}(r)}\right) \doteq \mathrm{P}\left(\boldsymbol{\alpha} \in \mathcal{A}(r)\right) \doteq \rho^{-d(r)}$ as $\boldsymbol{\alpha} \in \bar{\mathcal{A}}'(r)$ implies that $N \leq \rho^{\bar{c}(r)}$ by (26). In words, $\bar{c}(r)$ is the worst-case upper bound in (26) over the set of $\boldsymbol{\alpha}$ that are not in outage and occur with a probability that do not vanish exponentially fast. By a similar argument we can strengthen this statement to $\mathrm{P}\left(N \geq \rho^{x}\right) \dot{<} \rho^{-d(r)}$ for any $x > \bar{c}(r)$ by starting with $\mathcal{A}(r-\epsilon)$ and noting that $\mathrm{P}\left(\boldsymbol{\alpha} \in \mathcal{A}(r-\epsilon)\right) \doteq \rho^{-d(r-\epsilon)} \dot{<} \rho^{-d(r)}$ as $d(r-\epsilon) > d(r)$ for any $\epsilon > 0$. Consequently, $c(r) \leq \bar{c}(r)$ is a valid upper bound on the complexity exponent.

In order to evaluate $\bar{c}(r)$ we note that the sum on the left hand side of (27) is symmetric in α_i and each term is equal to zero for small a_i, increases linearly with α_i for intermediated vales, and then saturates at a value of r for $\alpha_i \geq 1$. Using these observations it is straightforward to see that an $\boldsymbol{\alpha}$ that maximizes $v(\boldsymbol{\alpha})$ over $\bar{\mathcal{A}}'(r)$, labeled $\boldsymbol{\alpha}^\star = [\alpha_1^\star, \ldots, \alpha_n^\star]$ satisfies $\alpha_i^\star = 1$ for $i = 1, \ldots, n-k-1$ where $k = \lfloor r \rfloor$, $\alpha_{n-k}^\star = k + 1 - r$, and $\alpha_i^\star = 0$ for $i = n-k+1, \ldots, n$. Note that this is also, not surprisingly, the same $\boldsymbol{\alpha}$ that gives the *typical outages* in [3]. Evaluating $v(\boldsymbol{\alpha}^\star)$ gives

$$v(\boldsymbol{\alpha}^\star) = \bar{c}(r) = r(n - \lfloor r \rfloor - 1) + \big(n\lfloor r \rfloor - r(n-1)\big)^+$$

and establishes one side of the equality in (16) as $c(r) \leq \bar{c}(r)$. The expression in (17) for integer values of r is obtained by setting $k = r = \lfloor r \rfloor$ in (16). The proof that $c(r) \geq \bar{c}(r)$ is similar in spirit and provided in [7]. □

Appendix B
Proof of Lemma 1

The aim is to provide an upper bound on the number of integer points contained in $\mathcal{E} \cap \mathcal{B} \cap \mathbb{Z}^m$ which is graphically illustrated by the shaded region in Fig. 2. To this end, note that the length of the ith semi-axis of $\mathcal{E}$, is $e_i \triangleq 2\xi/\sigma_i(\boldsymbol{D})$. Let $\mathcal{C}_1$ be the smallest orthotope (box), aligned with and containing $\mathcal{E}$, i.e., $\mathcal{C}_1$ is an orthotope with side lengths e_i (see Fig. 2). Let $\mathcal{C}_2$ be a hypercube with side-length $2\sqrt{m}\eta$, centered at the origin and aligned with $\mathcal{C}_1$ (see Fig. 2). As the diagonal of $\mathcal{B}$ is $2\sqrt{m}\eta$ it follows that $\mathcal{B} \subset \mathcal{C}_2$, regardless of the orientation of $\mathcal{C}_2$. Let $\mathcal{C}_3$ be given by $\mathcal{C}_3 = \mathcal{C}_1 \cap \mathcal{C}_2$ and note that $\mathcal{E} \cap \mathcal{B} \subset \mathcal{C}_3$ as $\mathcal{E} \subset \mathcal{C}_1$ and $\mathcal{B} \subset \mathcal{C}_2$. As $\mathcal{C}_1$ and $\mathcal{C}_2$ are aligned, it follows that $\mathcal{C}_3$ is also an orthotope. Let $l_1, \ldots, l_m$ denote the side-lengths of $\mathcal{C}_3$ and note that $l_i \leq \min(e_i\,,\, 2\sqrt{m}\eta)$.

The number of integer lattice points in a set is under uniform random perturbations of the lattice equal to its volume [13], i.e., for $\mathcal{C} \subset \mathbb{R}^n$ it holds that

$$\mathrm{Vol}(\mathcal{C}) = \int_{\mathcal{U}} |\mathbb{Z}^n \cap \mathcal{C} + \boldsymbol{u}| d\boldsymbol{u} \qquad (28)$$

where $\mathcal{U} \triangleq \big[-\frac{1}{2}, \frac{1}{2}\big]^n$ denotes the unit cube in $\mathbb{R}^n$. This is referred to as the mean value theorem in [13]. Let $\mathcal{C}_4$ be the orthotope, aligned with and centered around $\mathcal{C}_3$, with side lenghts $l_i + \sqrt{m}$ (see Fig. 2). By construction, it follows that $\mathcal{C}_3 \subset \mathcal{C}_4 + \boldsymbol{u}$ for *all* $\boldsymbol{u} \in \mathcal{U}$. It therefore follows by (28) that $|\mathcal{C}_3 \cap \mathbb{Z}^m| \leq \mathrm{Vol}(\mathcal{C}_4) = \prod_{i=1}^n [\sqrt{m} + l_i]$. where $l_i \leq \min\left(2\xi/\sigma_i(\boldsymbol{D})\,,\, 2\sqrt{m}\eta\right)$. As $\mathcal{E} \cap \mathcal{B} \subset \mathcal{C}_3$ the upper bound in (20) follows. □

References

[1] E. Viterbo and J. Boutros, "A universal lattice code decoder for fading channels," *IEEE Trans. Inform. Theory*, vol. 45, no. 5, pp. 1639–1642, Jul. 1999.

[2] M. O. Damen, H. E. Gamal, and G. Caire, "On maximum-likelihood detection and the search for the closest lattice point," *IEEE Transactions on Information Theory*, vol. 49, no. 10, pp. 2389–2401, Oct. 2003.

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

[4] K. Kumar, G. Caire, and A. Moustakas, "Asymptotic performance of linear receivers in mimo fading channels," *IEEE Transactions on Information Theory*, vol. 55, no. 10, pp. 4398–4418, October 2009.

[5] M. Taherzadeh and A. K. Khandani, "On the limitations of the naive lattice decoding," *IEEE Transactions on Information Theory*, vol. 56, no. 10, pp. 4820–4826, October 2010.

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

[7] ——, "Sphere decoding complexity exponent for full rate codes over the quasi-static mimo channel," Jan. 2011, submitted to the IEEE Trans. on Information Theory. Available as arXiv:???

[8] F. Oggier, G. Rekaya, J.-C. Belfiore, and E. Viterbo, "Perfect space-time block codes," *IEEE Trans. Inform. Theory*, vol. 52, no. 9, pp. 3885–3902, Sep. 2006.

[9] P. Elia, B. A. Sethuraman, and P. Vijay Kumar, "Perfect space-time codes for any number of transmit antennas," *IEEE Trans. Inform. Theory*, vol. 53, no. 11, pp. 3853–3868, Nov. 2007.

[10] D. Seethaler, J. Jalden, C. Studer, and H. Bolcskei, "Tail behavior of sphere-decoding complexity in random lattices," in *Proc. IEEE International Symposium on Information Theory, ISIT*, Jun. 2009, pp. 729 – 733.

[11] W. Abediseid and M. Damen, "Lattice sequential decoder for coded MIMO channel: Performance and complexity analysis," Jan. 2011, submitted to the IEEE Trans. on Information Theory. Available as arXiv:1101.0339v1.

[12] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*. Cambridge University Press, 1991.

[13] P. Gritzmann and J. M. Wills, "Lattice points," in *Handbook of Convex Geometry*, P. M. Gruber and J. M. Wills, Eds. North-Holland, 1993, vol. B, ch. 3.2.