Achieving a vanishing SNR-gap to exact lattice decoding at a subexponential complexity

Arun Singh, Petros Elia and Joakim Jaldén

***Abstract*—The work identifies the first lattice decoding solution that achieves, in the general outage-limited MIMO setting and in the high-rate and high-SNR limit, both a vanishing gap to the error-performance of the (DMT optimal) exact solution of preprocessed lattice decoding, as well as a computational complexity that is subexponential in the number of codeword bits. The proposed solution employs lattice reduction (LR)-aided regularized (lattice) sphere decoding and proper timeout policies. These performance and complexity guarantees hold for most MIMO scenarios, all reasonable fading statistics, all channel dimensions and all full-rate lattice codes.**

In sharp contrast to the above very manageable complexity, the complexity of other standard preprocessed lattice decoding solutions is revealed here to be extremely high. Specifically the work is first to quantify the complexity of these lattice (sphere) decoding solutions and to prove the surprising result that the complexity required to achieve a certain rate-reliability performance, is exponential in the lattice dimensionality and in the number of codeword bits, and it in fact matches, in common scenarios, the complexity of ML-based solutions. Through this sharp contrast, the work was able to, for the first time, rigorously demonstrate and quantify the pivotal role of lattice reduction as a special complexity reducing ingredient.

Finally the work analytically refines transceiver DMT analysis which generally fails to address potentially massive gaps between theory and practice. Instead the adopted vanishing gap condition guarantees that the decoder's error curve is arbitrarily close, given a sufficiently high SNR, to the optimal error curve of exact solutions, which is a much stronger condition than DMT optimality which only guarantees an error gap that is subpolynomial in SNR, and can thus be unbounded and generally unacceptable for practical implementations.

I. Introduction

The work applies to the general setting of outage-limited MIMO communications, where MIMO techniques offer significant advantages in terms of increased throughput and reliability, although at a cost of a potentially much higher computational complexity for decoding at the receivers. This high complexity brings to the fore the need for efficient decoders that tradeoff error-performance with complexity in a better manner than computationally expensive decoders like the strictly optimal maximum-likelihood (ML) decoder.

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 Electric R&D Centre Europe project Home-eNodeBS.

A. Singh and P. Elia are with the Mobile Communications Department, EURECOM, Sophia Antipolis, France (email: {singhak, elia}@eurecom.fr)

J. Jaldén is with the ACCESS Linnaeus Center, Signal Processing Lab, KTH Royal Institute of Technology, Stockholm, Sweden (email: jalden@kth.se)

Specifically in terms of ML-based decoding, the use of the brute-force ML decoder, introduces a complexity that scales exponentially with the number of codeword bits. If on the other hand, a small gap to the exact ML performance is acceptable, then different branch-and-bound algorithms such as the sphere decoder (SD) have been known to accept reduced computational resources. Despite the reduced complexity of sphere decoding, recent work in [1] has revealed that, to achieve a vanishing error-gap to optimal ML solutions, even such branch-and-bound algorithms generally require computational resources that, albeit significantly smaller than those required by a brute-force ML decoder, again grow exponentially in the rate and the dimensionality, and remain prohibitive for several MIMO scenarios.

This high complexity required by ML-based decoding solutions, serves as further motivation for exploring other families of decoding methods. A natural alternative is lattice decoding obtained by simply removing the constellation boundaries of the ML-based search, an action that loosely speaking exploits a certain symmetry which in turn may yield faster implementations. It is the case though that even with lattice decoding, the computational complexity can be prohibitive: finding the exact solution to the lattice decoding problem is generally an NP hard problem (cf. [2]). At the same time though, the other extreme of very early terminations of lattice decoding, such as linear solutions, have been known to achieve computational efficiency at the expense though of a very sizable, and often unbounded, gap to the exact solution of the lattice decoding problem.

In this work we explore lattice decoding solutions that, in conjunction with terminating policies, strike the proper balance between this exponential complexity and exponential gap.

A. System model

We consider the general $m \times n$ point-to-point multiple-input multiple-output model given by

$$\mathbf{y} = \sqrt{\rho}\mathbf{H}\mathbf{x} + \mathbf{w} \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^m$, $\mathbf{y} \in \mathbb{R}^n$ and $\mathbf{w} \in \mathbb{R}^n$ respectively denote the transmitted codewords, the received signal vectors, and the additive white Gaussian noise with unit variance, where the parameter ρ takes the role of the signal to noise ratio (SNR), and where the fading matrix $\mathbf{H} \in \mathbb{R}^{n\times m}$ is assumed to be random, with elements drawn from arbitrary statistical distributions. We consider that one use of (1) corresponds to T uses of some underlying "physical" channel. We further assume the transmitted codewords $\mathbf{x}$ to be uniformly distributed

over some codebook $\mathcal{X} \in \mathbb{R}^m$, to be statistically independent of the channel $\mathbf{H}$, and to satisfy the power constraint

$$E\{\|\mathbf{x}\|^2\} \leq T. \tag{2}$$

B. Rate, reliability and complexity in outage-limited MIMO communications

In terms of error performance, we let P_e denote the probability of codeword error, and we consider the rate,

$$R = \frac{1}{T}\log|\mathcal{X}|, \tag{3}$$

in bits per channel use (bpcu), where $|\mathcal{X}|$ denotes the cardinality of $\mathcal{X}$.

Regarding complexity, we let $N_{\max}$ describe the computational resources, in floating point operations (flops) per T channel uses, that the transceiver is endowed with, in the sense that after $N_{\max}$ flops, the transceiver must simply terminate, potentially prematurely and before completion of its task. We note that naturally, $N_{\max}$ is intimately intertwined with the desired P_e and R, and that any attempt to significantly reduce $N_{\max}$ may be at the expense of a substantial degradation in error-performance.

In the high SNR regime, a given encoder $\mathcal{X}_r$ and decoder $\mathcal{D}_r$ are said to achieve a *multiplexing gain* r (cf. [3]) and *diversity gain* $d(r)$ if

$$\lim_{\rho\to\infty}\frac{R(\rho)}{\log\rho} = r, \qquad \text{and} \qquad -\lim_{\rho\to\infty}\frac{\log P_e}{\log\rho} = d(r). \tag{4}$$

In the same high SNR regime, the complexity is here chosen to take the form

$$c(r) := \lim_{\rho\to\infty}\frac{N_{\max}}{\log\rho}, \tag{5}$$

which is henceforth denoted as the *complexity exponent*. Noting that $R = r\log\rho$, we observe that $c(r) > 0$ implies a complexity that is exponential in the rate.

Remark 1: A reasonable question at this point would pertain as to why the computational resources $N_{\max}$ scale with ρ and are dependent on r, to which we note that the complexity of decoding is generally dependent on the density of the codebook, which in turn depends on ρ and R. Furthermore this dependence of the complexity exponent (and by extension of $N_{\max}$) on r, reflects a potential ability to regulate the computational resources depending on the rate. Finally the fact that both P_e and $N_{\max}$ are represented as polynomial functions of ρ, simply stems from the fact that both P_e and $|\mathcal{X}|$ naturally scale as polynomial functions of ρ. Specifically we quickly note that $c(r)$ captures the entire complexity range

$$0 \leq c(r) \leq rT$$

of all reasonable transceivers, with $c(r) = 0$ corresponding to the fastest possible transceiver (requiring a subexponential number of flops per T channel uses), and with $c(r) = rT$ corresponding to the optimal but arguably slowest, full-search uninterrupted ML decoder[1] in the presence of a canonical code with multiplexing gain r, i.e., with $|\mathcal{X}_r| = 2^{RT} = \rho^{rT}$.

If this canonical code though is linear, searching the entire codebook can be avoided by algorithmic solutions like the sphere decoder (SD) which can provide substantial complexity reductions at a potential small loss in error performance. Such solutions take advantage of the linear nature of the code that is defined by a *generator matrix* $\boldsymbol{G}$ and a *shaping region* $\mathcal{R}^{'}$. Specifically for $r \geq 0$, a (sequence of) full-rate linear (lattice) code(s) $\mathcal{X}_r$ is given by $\mathcal{X}_r = \Lambda_r \cap \mathcal{R}^{'}$ where $\Lambda_r \triangleq \rho^{\frac{-rT}{\kappa}}\Lambda$ and $\Lambda \triangleq \{\mathbf{Gs} \mid \mathbf{s} \in \mathbb{Z}^\kappa\}$, where $\mathbb{Z}^\kappa$ denotes the $\kappa = \min\{m, n\}$ dimensional integer lattice, where $\mathcal{R}^{'}$ is a compact convex subset of $\mathbb{R}^\kappa$ that is independent of ρ, and where $\mathbf{G} \in \mathbb{R}^{m\times\kappa}$ is full rank and independent of ρ. For the class of lattice codes considered here, the codewords take the form

$$\mathbf{x} = \rho^{\frac{-rT}{\kappa}}\mathbf{Gs}, \qquad \mathbf{s} \in \mathbb{S}_r^\kappa \triangleq \mathbb{Z}^\kappa \cap \rho^{\frac{rT}{\kappa}}\mathcal{R}, \tag{6}$$

where $\mathcal{R} \subset \mathbb{R}^\kappa$ is a natural bijection of the shaping region $\mathcal{R}^{'}$ that preserves the code, and where $\mathcal{R}$ contains the all zero vector $\mathbf{0}$.

As noted before, despite the reduced complexity of sphere decoding of such lattice codes (as compared to brute-force ML decoding), recent work in [1] has revealed that even such branch-and-bound algorithms generally require computational resources that grow exponentially in the number of codeword bits and the dimensionality. As an indicative example of this high complexity, we note that the work in [1] showed that such SD algorithms, when applied for decoding a large family of high-performing codes including all known full-rate DMT optimal codes, over the $n_{\mathrm{T}} \times n_{\mathrm{R}}$ quasi-static MIMO channel with Rayleigh fading and $n_{\mathrm{R}} \geq n_{\mathrm{T}}$, introduce a complexity exponent[2] of the form

$$c(r) = \frac{T}{n_{\mathrm{T}}}\big(r(n_{\mathrm{T}} - \lfloor r\rfloor - 1) + (n_{\mathrm{T}}\lfloor r\rfloor - r(n_{\mathrm{T}} - 1))^+\big). \tag{7}$$

In the above, $\lfloor r\rfloor$ denotes the largest integer not greater than r. The exponent, which simplifies to $c(r) = \frac{T}{n_{\mathrm{T}}}r(n_{\mathrm{T}} - r)$ for integer values of r, reaches at $r = n_{\mathrm{T}}/2$ (for even values of n_{T}) an overall maximum value of $n_{\mathrm{T}}T/4$ which, for the aforementioned codes is equal to $\kappa/8$, corresponding to complexity in the order of $2^{\frac{1}{8}\kappa\log\rho} = \rho^{\kappa/8} = \sqrt{|\mathcal{X}|}$. At any fixed multiplexing gain, these required computational resources can be seen to be in the order of $2^{RT(\frac{n_{\mathrm{T}}-r}{n_{\mathrm{T}}})}$ flops which reveals a complexity that is exponential in the number of codeword bits, and a corresponding exponential slope of $\frac{n_{\mathrm{T}}-r}{n_{\mathrm{T}}}$.

[1] We here note that strictly speaking, $\mathcal{X}_r, \mathcal{D}_r$ may potentially introduce a complexity exponent larger than rT. In such a case though, $\mathcal{X}_r, \mathcal{D}_r$ may be substituted by a lookup table implementation of $\mathcal{X}_r$ and an unrestricted ML decoder. This encoder-decoder will jointly require resources that are a constant multiple of $|\mathcal{X}_r| \doteq \rho^{rT}$ as it has to construct and visit all possible $|\mathcal{X}_r|$ codewords, at a computational cost of a bounded number of flops per codeword visit. It is noted that the number of flops per visited codeword is naturally independent of ρ.

[2] Although premature at this point, we hasten to note for the expert reader that this complexity indeed holds irrespective of the radius updating policy, irrespective of the decoding ordering, and as we will see later on, holds even in the presence of MMSE preprocessing.

C. Transition to lattice decoding for reducing complexity

As mentioned, this high complexity of ML based (constrained) decoders, motivates consideration of other decoder families, with a natural alternative being the unconstrained (naive) lattice decoder which takes the general form

$$\hat{\mathbf{x}}_L = \arg\min_{\hat{\mathbf{x}}\in\Lambda_r} \|\mathbf{y} - \sqrt{\rho}\mathbf{H}\hat{\mathbf{x}}\|^2. \tag{8}$$

Naturally when $\hat{\mathbf{x}}_L \notin \mathcal{X}_r$, the decoder declares an error.

The use of lattice decoding, and specifically of preprocessed lattice decoding in MIMO communications has received substantial attention from works like [4], [5] and [6], where the latter proved that lattice decoding in the presence of MMSE preprocessing achieves the optimal DMT for specific MIMO channels and statistics, and for DMT-optimal random codes. The use of lattice decoding as an alternative to computationally expensive ML based solutions, was recently further validated on the one hand by the aforementioned work in [1], [7] which revealed the large computational disadvantages of ML based solutions, and on the other hand by the work in [8] which further confirmed the performance advantages of lattice decoding by showing that regularized (MMSE-preprocessed)[3] lattice decoding achieves the optimal DMT performance, for almost all MIMO scenarios and fading statistics, and all non-random lattice codes, irrespective of the codes' ML performance.

It is the case though that the aforementioned extreme complexity of exact lattice decoding solutions, in conjunction with the potentially unbounded error-performance degradation (gap) of very early terminations (as opposed to exact implementations) of lattice decoding, bring to the fore the need for balanced approximations of lattice decoding solutions that better balance the very sizable complexity and gap. Specifically for any simplified variant $\mathcal{D}_r$ of the baseline (exact) MMSE-preprocessed lattice decoder, this gap can, in the high SNR regime, be quantified as

$$g_L(c) \triangleq \lim_{\rho\to\infty} \frac{P_e}{\mathrm{P}\left(\hat{\mathbf{x}} \neq \mathbf{x}\right)} \tag{9}$$

where $\mathrm{P}\left(\hat{\mathbf{x}} \neq \mathbf{x}\right)$ describes the probability of error of the *exact* MMSE-preprocessed lattice decoder, where P_e denotes the probability of error of $\mathcal{D}_r$, and where c (i.e., $c(r)$) is the complexity exponent that describes the (asymptotic rate of increase of the) computational resources required to achieve this performance gap. Generally a smaller computational complexity exponent c implies a larger gap $g_L(c)$. The clear task has remained for some time to construct decoders that optimally traverse this tradeoff between g and c, i.e., that reduce the performance gap to the exact lattice decoding solution, with reasonable computational complexity. Equivalently for $N_{\max}(g)$ denoting the computational resources in flops required to achieve a certain gap g to the baseline exact MMSE-preprocessed lattice decoder, the above task can be described, in the high SNR regime, as trying to minimize

$$\lim_{\rho\to\infty} \frac{\log N_{\max}(g)}{\log \rho}.$$

This will be achieved later on.

D. Contributions

We first show that the computational complexity required by the MMSE-preprocessed (unconstrained) lattice sphere decoder, asymptotically matches the complexity of the (constrained) ML-based (MMSE-preprocessed or not) sphere decoders, and is commonly exponential in the dimensionality and the number of codeword bits. This is established for a large class of codes of arbitrary error-performance, a large class of fading statistics, and specifically for the quasi-static MIMO channel – for example the complexity required for DMT optimal lattice sphere decoding, in the presence of a large family of DMT optimal codes, takes the previously seen simple piecewise linear form in (7). In a parenthetical note, and deviating slightly from the spirit of this paper, we also provide a universal upper bound on the complexity of regularized lattice sphere decoding, which holds irrespective of the lattice code applied and irrespective of the fading statistics. This upper bound again takes the form in (7), matching that in the case of constrained ML-based sphere decoding, thus revealing the surprising fact that there exists no statistical channel behavior that will allow the removal of the bounding region to cause unbounded increases in the complexity of the decoder[4].

With provable evidence of the very high complexity of regularized lattice decoding, we turn to the powerful tool of lattice reduction and seek to understand its effects on computational complexity. While there has existed a general agreement in the community that lattice reduction does reduce complexity, cf. [10], this has not yet been supported analytically in any relevant communication settings. In fact, and quite opposite to common wisdom, it was recently shown that for a fixed-radius[5] sphere decoding implementation of the naive lattice decoder [11], LR does not improve the sphere decoder complexity tail exponent.

What our present work shows is that lattice reduction reduces an ML-like exponentially increasing complexity, to very manageable subexponential values. We specifically proceed to prove that the LR-aided regularized lattice decoder, implemented by a fixed-radius sphere decoder and timeout policies that occasionally abort decoding and declare an error, achieves

$$g_L(\epsilon) = 1, \quad \lim_{\rho\to\infty} \frac{\log N_{\max}(g)}{\log \rho} = 0 \qquad \forall \epsilon > 0, g \geq 1,$$

i.e., achieves a vanishing gap to the exact implementation of regularized lattice decoding and does so with a complexity

[3]We will interchangeably use *MMSE-preprocessed decoder* and *regularized decoder*, with the first term being more commonly used, and with the second implying a more general family of decoders (cf. [8] where the equivalence between the two decoders is discussed.). Even though in the asymptotic setting of interest, the two accept the same results throughout the paper, some extra error-performance gains can be achieved by proper optimization of the regularized decoder (cf. [9]).

[4]In other words, this complexity bound holds even if the channel statistics are such that the channel realizations cause the decoder to always have to solve the hardest possible lattice search problem.

[5]The radius here is considered fixed in the sense that it does not vary with respect to the channel realization and rate.

exponent that vanishes to zero, which in turn implies subexponential complexity in the sense that the complexity scales slower than any conceivable exponential function. It is finally noted that this vanishing gap approach serves the practical purpose of an analytical refinement over basic diversity analysis which generally fails to address potentially massive gaps between theory and practice.

E. Notation

We use $\doteq$ to denote the *exponential equality*, i.e., we write $f(\rho) \doteq \rho^B$ to denote $\lim_{\rho\to\infty} \frac{\log f(\rho)}{\log \rho} = B$, and $\dot{\leq}$, $\dot{\geq}$ are similarly defined. With this notation, we can write $P_e \doteq \rho^{-d(r)}$ (cf. (4)). In this paper we use $\ulcorner \bullet \urcorner$ to denote the smallest integer not smaller than the argument, $\llcorner \bullet \lrcorner$ to denote the largest integer not larger than the argument, $(\bullet)^H$ to denote the conjugate transpose of $(\bullet)$, $(\bullet)^+$ to denote $\max\{0, (\bullet)\}$ and $vec(\bullet)$ to denote the operation whereby the columns of the argument $(\bullet)$ are stacked to form a vector.

II. MMSE-Preprocessed Lattice Sphere Decoding Complexity

We proceed to describe the preprocessed lattice decoder, its sphere decoding implementation, and for a practical setting of interest that includes the quasi-static MIMO channel and common codes, to establish the decoder's computational complexity.

A. Lattice sphere decoding

Combining (1) and (6) yields the equivalent model

$$\mathbf{y} = \mathbf{M}_r \mathbf{s} + \mathbf{w} \tag{10}$$

where

$$\mathbf{M}_r = \rho^{\frac{1}{2}-\frac{rT}{\kappa}} \mathbf{H}\mathbf{G} \in \mathbb{R}^{n\times\kappa} \tag{11}$$

is a function of the multiplexing gain[6] r.

Consequently the corresponding naive lattice decoder in (8) takes the form (see for example [8], also [10])

$$\hat{\mathbf{s}}_L = \arg\min_{\hat{\mathbf{s}}\in\mathbb{Z}^\kappa} \|\mathbf{y} - \mathbf{M}\hat{\mathbf{s}}\|^2 . \tag{12}$$

As a result though of neglecting the boundary region, the above decoder declares additional errors if $\hat{\mathbf{s}}_L \notin \mathbb{S}_r^\kappa$, resulting in possible performance costs. These costs motivated the use of MMSE preprocessing which essentially regularizes the decision metric to penalize vectors outside the boundary constraint $\mathbb{S}_r^\kappa$ (cf. [8]). Specifically the MMSE-preprocessed lattice decoder is obtained by implementing an unconstrained search over the MMSE-preprocessed lattice, and takes the form

$$\hat{\mathbf{s}}_{r-ld} = \arg\min_{\hat{\mathbf{s}}\in\mathbb{Z}^\kappa} \|\mathbf{F}\mathbf{y} - \mathbf{R}\hat{\mathbf{s}}\|^2 , \tag{13}$$

where $\mathbf{F}$ and $\mathbf{R}$ are respectively the MMSE forward and feedback filters such that $\mathbf{F} = \mathbf{R}^{-H}\mathbf{M}^H$, where

$$\mathbf{R}^H\mathbf{R} = \mathbf{M}^H\mathbf{M} + \alpha_r^2\mathbf{I}, \tag{14}$$

[6]For simplicity of notation we will, in most cases, denote $\mathbf{M}_r$ with $\mathbf{M}$.

where $\alpha_r = \rho^{\frac{-rT}{\kappa}}$ and where $\mathbf{R}$ is an upper-triangular matrix (more details can be found in Appendix D). For $\mathbf{r} \triangleq \mathbf{F}\mathbf{y}$, the model transitions from (10) to

$$\begin{aligned} \mathbf{r} &= \mathbf{R}^{-H}\mathbf{M}^H\mathbf{M}\mathbf{s} + \mathbf{R}^{-H}\mathbf{M}^H\mathbf{w} \\ &= \mathbf{R}^{-H}(\mathbf{R}^H\mathbf{R} - \alpha_r^2\mathbf{I})\mathbf{s} + \mathbf{R}^{-H}\mathbf{M}^H\mathbf{w} \\ &= \mathbf{R}\mathbf{s} - {\alpha_r}^2\mathbf{R}^{-H}\mathbf{s} + \mathbf{R}^{-H}\mathbf{M}^H\mathbf{w} \\ &= \mathbf{R}\mathbf{s} + \mathbf{w}' \end{aligned} \tag{15}$$

where

$$\mathbf{w}' = -\alpha_r^2\mathbf{R}^{-H}\mathbf{s} + \mathbf{R}^{-H}\mathbf{M}^H\mathbf{w} \tag{16}$$

is the equivalent noise that includes self-interference (first summand) and colored Gaussian noise. Consequently the corresponding regularized lattice decoder takes the form

$$\hat{\mathbf{s}}_{r-ld} = \arg\min_{\hat{\mathbf{s}}\in\mathbb{Z}^\kappa} \|\mathbf{r} - \mathbf{R}\hat{\mathbf{s}}\|^2 , \tag{17}$$

which is then solved by the sphere decoder which recursively enumerates all lattice vectors $\hat{\mathbf{s}} \in \mathbb{Z}^\kappa$ within a given sphere of radius $\xi > 0$, i.e., which identifies as candidates the vectors $\hat{\mathbf{s}}$ that satisfy

$$\|\mathbf{r} - \mathbf{R}\hat{\mathbf{s}}\|^2 \leq \xi^2. \tag{18}$$

The algorithm specifically uses the upper-triangular nature of $\mathbf{R}$ to recursively identify partial symbol vectors $\hat{\mathbf{s}}_k$, $k = 1, \cdots, \kappa$, for which

$$\|\mathbf{r}_k - \mathbf{R}_k\hat{\mathbf{s}}_k\|^2 \leq \xi^2, \tag{19}$$

where $\hat{\mathbf{s}}_k$ and $\mathbf{r}_k$ respectively denote the last k components of $\hat{\mathbf{s}}$ and $\mathbf{r}$, and where $\mathbf{R}_k$ denotes the $k\times k$ lower-right submatrix of $\mathbf{R}$. Clearly any set of vectors $\hat{\mathbf{s}} \in \mathbb{Z}^\kappa$, with common last k components that fail to satisfy (19), may be excluded from the set of candidate vectors that satisfy (18).

The enumeration of partial symbol vectors $\hat{\mathbf{s}}_k$ is equivalent to the traversal of a regular tree with κ layers – one layer per symbol component of the symbol vectors, such that layer k corresponds to the kth component of the transmitted symbol vector[7] $\mathbf{s}$. There is a one-to-one correspondence between the nodes at layer k and the partial vectors $\hat{\mathbf{s}}_k$. We say that a node is visited by the sphere decoder if and only if the corresponding partial vector $\hat{\mathbf{s}}_k$ satisfies (19), i.e., there is a bijection between the visited nodes at layer k and the set

$$\mathcal{N}_k \triangleq \{\hat{\mathbf{s}}_k \in \mathbb{Z}^k \mid \|\mathbf{r}_k - \mathbf{R}_k\hat{\mathbf{s}}_k\|^2 \leq \xi^2\}. \tag{20}$$

B. Complexity of MMSE-preprocessed lattice sphere decoding

Consequently the total number of visited nodes (in all layers of the tree) is given by

$$N_{SD} = \sum_{k=1}^{\kappa} N_k, \tag{21}$$

where $N_k \triangleq |\mathcal{N}_k|$ is the number of visited nodes at layer k of the search tree. The total number of visited nodes is commonly

[7]We will henceforth refer to the symbol vector $\mathbf{s} \in \mathbb{S}_r^\kappa$ corresponding to the transmitted codeword $\mathbf{x} = \rho^{\frac{-rT}{\kappa}}\mathbf{G}\mathbf{s}$ (cf. (6)) , simply as the *transmitted symbol vector*.

taken as a measure of the sphere decoder complexity. It is easy to show that in the scale of interest the SD complexity exponent $c(r)$ would not change if instead of considering the number of visited nodes, we considered the number of flops spent by the decoder[8].

Naturally the total number of visited nodes is a function of the search radius ξ. We here use a fixed radius, which may result in a non-zero probability that the transmitted symbol vector $\mathbf{s}$ is not in $\mathcal{N}_\kappa$. Consequently we must choose a radius that strikes the proper balance between decreasing the aforementioned probability and at the same time sufficiently decreasing the size of $\mathcal{N}_\kappa$. Towards this we note that for the transmitted symbol vector $\mathbf{s}$, the metric in (17) satisfies

$$\|\mathbf{r}-\mathbf{R}\mathbf{s}\|^2 = \|\mathbf{w}^{'}\|^2,$$

which means that if $\|\mathbf{w}^{'}\| > \xi$, then the transmitted symbol vector is excluded from the search, resulting in a decoding error. As Lemma 2 will later argue taking into consideration the self-interference and non-Gaussianity of $\mathbf{w}^{'}$, we can set $\xi = \sqrt{z\log\rho}$, for some $z > d(r)$ such that

$$\mathrm{P}\left(\|\mathbf{w}^{'}\|^2 > \xi^2\right) \doteq \rho^{-d(r)},$$

which implies a vanishing probability of excluding the transmitted information vector from the search, and a vanishing degradation of error performance.

We here note that the MMSE-preprocessed lattice sphere decoder differs from its ML-based equivalent in two aspects: the presence of MMSE preprocessing and the absence of a bounding region to constrain the search. These two aspects are generally perceived to have an opposite effect on the complexity. On the one hand, MMSE preprocessing, which we recall from (20) to introduce unpruned sets

$$\mathcal{N}_k \triangleq \{\hat{\mathbf{s}}_k \in \mathbb{Z}^k \mid \|\mathbf{r}_k - \mathbf{R}_k\hat{\mathbf{s}}_k\|^2 \leq \xi^2\}, \quad k=1,\cdots,\kappa,$$

is associated to reduced complexity in lattice-based SD solutions (cf. [11]) due to the resulting penalization of faraway lattice points (cf. [8]). On the other hand, the absence of boundary constraints can be associated to increased complexity as it introduces an unbounded number of candidate vectors. We proceed to show that in terms of the complexity exponent, under common MIMO scenarios and codes, these two aspects exactly cancel each other out, and that consequently MMSE-preprocessed lattice sphere decoding introduces a complexity exponent that matches that of ML-based sphere decoding (cf. [1]), which it self is shown here to also match the complexity exponent of ML-based SD in the presence of MMSE preprocessing[9].

Before proceeding we note that this analysis is specific to sphere decoding, and that it does not account for any other ML based solutions that could, under some (arguably rare) circumstances, be more efficient. A classical example of such rare circumstances would be a MIMO scenario, or equivalently a set of fade statistics, that always generate diagonal channel matrices. Another example would be having codes drawn from orthogonal designs which introduce very small decoding complexity, but which are provably shown to be highly suboptimal except for very few unique cases like the $n_{\mathrm{T}}=2, n_{\mathrm{R}}=1$ quasi-static case [12]. In light of this, in this section only, we mainly focus on the widely considered $n_T \times n_R$ ($n_R \geq n_T$) i.i.d. and quasi-static MIMO setting and on the large but specific family of full-rate ($\kappa = 2\min\{n_{\mathrm{T}}, n_{\mathrm{R}}\}T = 2n_{\mathrm{T}}T$) threaded codes (cf. [13]–[16]), which includes all known DMT optimal codes as well as uncoded transmission (V-BLAST).

We proceed with the main Theorem of the section, which applies under natural detection ordering (cf. [1], [5]), and under the assumption of i.i.d. regular fading statistics[10].

Theorem 1: The complexity exponent for MMSE-preprocessed lattice sphere decoding any full-rate threaded code over the quasi-static MIMO channel with i.i.d. regular fading statistics, is equal to the complexity exponent of ML-based SD with or without MMSE preprocessing.

Proof: See Appendix A. ∎

We clarify that even though all three decoders are DMT optimal, the above result incorporates more than just DMT optimal decoding, in the sense that any timeout policy will tradeoff $d(r)$ with $c(r)$ identically for ML-based and lattice-based sphere decoding. In other words the three decoders share the same $d(r)$ and $c(r)$ capabilities, irrespective of the timeout policy.

Furthermore, considering different SD detection orderings (cf. [5]), the following extends the range of codes for which the ML-based and lattice-based SD share a similar complexity. The proof follows from the proof of Theorem 1 in Appendix A, and from Theorem 4 in [1].

Corollary 1a: Given any full-rate code of arbitrary DMT performance, there is always at least one non-random fixed permutation of the columns of $\boldsymbol{G}$, for which the complexity exponent of the MMSE-preprocessed lattice sphere decoder matches that of the ML based sphere decoder.

The following focuses on a specific example of practical interest.

Corollary 1b: The complexity exponent for DMT optimal MMSE-preprocessed lattice sphere decoding of minimum delay ($T = n_T$) DMT optimal threaded codes over the quasi-static MIMO channel with i.i.d. regular fading statistics, takes the following form

$$c_{r-ld}(r) = r(n_T - \lfloor r \rfloor - 1) + (n_T \lfloor r \rfloor - r(n_T-1))^+, \tag{22}$$

which simplifies to

$$c_{r-ld}(r) = r(n_T - r) \tag{23}$$

for integer values of r.

[8]To see this, we consider that the cost of visiting a node, is independent of ρ. Once at a visited node, this same bounded cost includes the cost of establishing which children-nodes not to visit in the next layer.

[9]We clarify that ML-based SD in the presence of MMSE preprocessing, corresponds to unpruned sets $\mathcal{N}_k \cap \mathbb{S}_r^k$ where $\mathbb{S}_r^k$ is the k-dimensional set resulting from the natural reduction of $\mathbb{S}_r^\kappa$ from (6).

[10]The i.i.d. regular fading statistics satisfy the general set of conditions as described in [17], where a) the near-zero behavior of the fading coefficients h is bounded in probability as $c_1|h|^t \leq p(h) \leq c_2|h|^t$ for some positive and finite c_1, c_2 and t, where b) the tail behavior of h is bounded in probability as $p(h) \leq c_2 e^{-b|h|^\beta}$ for some positive and finite c_2, b and β, and where c) $p(h)$ is upper bounded by a constant K.

Proof: See Appendix B. ■

Further evidence that connects the complexity behavior of MMSE-preprocessed lattice-based SD, with that of its ML-based counterpart, now comes in the form of a non-trivial universal bound that is shared by the two methods. This is particularly relevant because unconstrained lattice decoding could conceivably require unbounded computational resources given the unbounded number of candidate lattice points. Specifically the following universal upper bound on the complexity of regularized lattice-based SD, matches the upper bound in [1] for the ML case, and it holds irrespective of the full-rate lattice code applied and irrespective of the fading statistics. The generality with respect to the fading statistics is important because it guarantees that no set of fading statistics, even those that always generate infinitely dense lattices, can cause an unbounded increase in the complexity due to removal of the boundary constraints.

Corollary 1c: Irrespective of the fading statistics and of the full-rate lattice code applied, the complexity exponents of MMSE-preprocessed lattice SD and of ML-based SD, are upper bounded by

$$\overline{c}(r) = \frac{T}{n_T}\left(r(n_T - \lfloor r \rfloor - 1) + (n_T \lfloor r \rfloor - r(n_T - 1))^+\right) \tag{24}$$

which simplifies to

$$\overline{c}(r) = \frac{T}{n_T} r(n_T - r) \tag{25}$$

for integer r.

Proof: See Appendix B. ■

The above results revealed the very high, ML-like complexity of MMSE-preprocessed lattice decoding. Coming back to the main focus of the paper, and after reverting to the most general setting of MIMO scenarios, statistics and full-rate lattice codes, we proceed to show how proper utilization of lattice sphere decoding and LR techniques can indeed reduce the complexity exponent to zero, at an error-performance cost that vanishes in the high SNR limit.

III. LR-aided Regularized Lattice Sphere Decoding Complexity

Lattice reduction techniques have been typically used in the MIMO setting to improve the error performance of suboptimal decoders (cf. [18], [19], see also [20], [21]). In the current setting the LR algorithm, which is employed at the receiver after the action of MMSE preprocessing, modifies the search of the MMSE-preprocessed lattice decoder, from

$$\hat{\mathbf{s}}_{rld} = \arg\min_{\hat{\mathbf{s}}\in\mathbb{Z}^\kappa} \|\mathbf{r} - \mathbf{R}\hat{\mathbf{s}}\|^2$$

(cf. (17)), to the new

$$\tilde{\mathbf{s}}_{lr-rld} = \arg\min_{\hat{\mathbf{s}}\in\mathbb{Z}^\kappa} \|\mathbf{r} - \mathbf{R}\mathbf{T}\hat{\mathbf{s}}\|^2, \tag{26}$$

by accepting as input the MMSE-preprocessed lattice generator matrix $\mathbf{R}$, and producing as output the matrix $\mathbf{T} \in \mathbb{Z}^{\kappa\times\kappa}$ which is unimodular meaning that it has integer coefficients and unit-norm determinant, and which is designed so that $\mathbf{RT}$ is (loosely speaking) more orthogonal than $\mathbf{R}$. As a result of this unimodularity, we have that $\mathbf{T}^{-1}\mathbb{Z}^\kappa = \mathbb{Z}^\kappa$, and consequently the new search in (26) corresponds to yet another *lattice* decoder, referred to as the LR-aided MMSE-preprocessed lattice decoder, which operates over a generally better conditioned channel matrix $\mathbf{RT}$.

Finally with sphere decoding in mind, the LR algorithm is followed by the QR decomposition[11] of the new lattice-reduced MMSE-preprocessed matrix $\mathbf{RT}$, resulting in a new upper-triangular model

$$\tilde{\mathbf{r}} = \tilde{\mathbf{R}}\tilde{\mathbf{s}} + \mathbf{w}'' \tag{27}$$

and in the new LR-aided MMSE-preprocessed lattice search, which accepts the application of the sphere decoder, and which takes the form

$$\tilde{\mathbf{s}}_{lr-rld} = \arg\min_{\hat{\mathbf{s}}\in\mathbb{Z}^\kappa} \left\|\tilde{\mathbf{r}} - \tilde{\mathbf{R}}\hat{\mathbf{s}}\right\|^2, \tag{28}$$

where $\tilde{\mathbf{Q}}\tilde{\mathbf{R}} = \mathbf{RT}$ corresponds to the QR-decomposition of $\mathbf{RT}$, where $\tilde{\mathbf{R}}$ is upper-triangular, where $\tilde{\mathbf{r}} \triangleq \tilde{\mathbf{Q}}^H\mathbf{r}$, $\tilde{\mathbf{s}} = \mathbf{T}^{-1}\mathbf{s}$, and where $\mathbf{w}'' = \tilde{\mathbf{Q}}^H\mathbf{w}'$.

At the very end,

$$\hat{\mathbf{s}}_{lr-rld} = \mathbf{T}\tilde{\mathbf{s}}_{lr-rld}, \tag{29}$$

allows for calculation of the estimate of the transmitted symbol vector $\mathbf{s}$ in (10).

We note here that this (exact) solution of the LR-aided MMSE-preprocessed lattice decoder defined by (28), (29), is identical to the exact solution of the MMSE-preprocessed lattice decoder given by (17), because

$$\begin{aligned}
\min_{\hat{\mathbf{s}}\in\mathbb{Z}^\kappa} \|\mathbf{r} - \mathbf{R}\hat{\mathbf{s}}\|^2 &= \min_{\hat{\mathbf{s}}\in\mathbb{Z}^\kappa} \left\|\mathbf{r} - \mathbf{R}\mathbf{T}\mathbf{T}^{-1}\hat{\mathbf{s}}\right\|^2 \\
&\overset{(a)}{=} \min_{\hat{\mathbf{s}}\in\mathbb{Z}^\kappa} \left\|\mathbf{r} - \tilde{\mathbf{Q}}\tilde{\mathbf{R}}\mathbf{T}^{-1}\hat{\mathbf{s}}\right\|^2 \\
&\overset{(b)}{=} \min_{\hat{\mathbf{s}}\in\mathbb{Z}^\kappa} \left\|\tilde{\mathbf{r}} - \tilde{\mathbf{R}}\mathbf{T}^{-1}\hat{\mathbf{s}}\right\|^2 \\
&= \min_{\hat{\mathbf{s}}\in\mathbf{T}^{-1}\mathbb{Z}^\kappa} \left\|\tilde{\mathbf{r}} - \tilde{\mathbf{R}}\hat{\mathbf{s}}\right\|^2 \\
&\overset{(c)}{=} \min_{\hat{\mathbf{s}}\in\mathbb{Z}^\kappa} \left\|\tilde{\mathbf{r}} - \tilde{\mathbf{R}}\hat{\mathbf{s}}\right\|^2,
\end{aligned} \tag{30}$$

where (a) follows from the fact that $\tilde{\mathbf{Q}}\tilde{\mathbf{R}} = \mathbf{RT}$, (b) follows from the rotational invariance of the Euclidean norm, and (c) follows from the fact that $\mathbf{T}^{-1}\mathbb{Z}^\kappa = \mathbb{Z}^\kappa$.

While though the two lattice decoding solutions (with and without LR) provide identical error performance in the setting of exact implementations, we proceed to show that, in terms of complexity, lattice reduction techniques, and specifically a proper utilization of the LLL algorithm [22], can provide dramatic improvements.

A. Complexity of the LR-aided regularized lattice sphere decoder

We are here interested in establishing the complexity of the LR-aided regularized lattice sphere decoder. Given that

[11] A more proper statement would be that the QR decomposition is performed by the LR algorithm it self.

the costs of implementing MMSE preprocessing and of implementing the linear transformation in (29) are negligible in the scale of interest[12], we limit our focus on establishing the cost of lattice reduction, and then the cost of the SD implementation of the search in (28). Starting with the SD complexity, as in (20), we identify the corresponding unpruned set at layer k to be

$$\mathcal{N}_k \triangleq \{\hat{\mathbf{s}}_k \in \mathbb{Z}^k \mid \|\tilde{\mathbf{r}}_k - \tilde{\mathbf{R}}_k \hat{\mathbf{s}}_k\|^2 \leq \xi^2\}, \tag{31}$$

and in bounding the size of the above, we first focus on understanding the statistical behavior of the $k \times k$ lower-right submatrices $\tilde{\mathbf{R}}_k$ of matrix $\tilde{\mathbf{R}}$ ($k = 1, \cdots, \kappa$), where we recall that $\tilde{\mathbf{R}}$ is the upper triangular code-channel matrix, after MMSE preprocessing and LLL lattice reduction. Towards this, and for $d_L(r-\epsilon)$ denoting the diversity gain of the exact implementation of the regularized lattice decoder at multiplexing gain $r-\epsilon$, we have the following lemma on the smallest singular value of $\tilde{\mathbf{R}}_k$. The proof appears in Appendix C.

Lemma 1: The smallest singular value $\sigma_{min}(\tilde{\mathbf{R}}_k)$ of submatrix $\tilde{\mathbf{R}}_k$, $k = 1, \cdots, \kappa$, satisfies

$$\mathrm{P}\left(\sigma_{min}(\tilde{\mathbf{R}}_k) \dot{<} \rho^{\frac{-\epsilon T}{\kappa}}\right) \dot{\leq} \rho^{-d_L(r-\epsilon)}, \text{ for all } r \geq \epsilon > 0. \tag{32}$$

To bound the cardinality N_k of $\mathcal{N}_k$ (cf. (31)), and eventually the total number $N_{SD} = \sum_{k=1}^{\kappa} N_k$ of lattice points visited by the SD, we proceed along the lines of the work in [1], making the proper modifications to account for MMSE preprocessing, for the removal of the bounding region, and for lattice reduction.

Towards this we see that, after removing the boundary constraint, Lemma 1 in [1] tells us that

$$N_k \triangleq |\mathcal{N}_k| \leq \prod_{i=1}^{k} \left[\sqrt{k} + \frac{2\xi}{\sigma_i(\tilde{\mathbf{R}}_k)}\right],$$

where

$$\sigma_{min}(\tilde{\mathbf{R}}_k) = \sigma_1(\tilde{\mathbf{R}}_k) \leq \cdots \leq \sigma_k(\tilde{\mathbf{R}}_k)$$

are the singular values of $\tilde{\mathbf{R}}_k$. Consequently we have that

$$N_k \quad \leq \quad \left[\sqrt{k} + \frac{2\xi}{\sigma_{min}(\tilde{\mathbf{R}}_k)}\right]^k. \tag{33}$$

As a result, for any $\tilde{\mathbf{R}}_k$ such that

$$\sigma_{min}(\tilde{\mathbf{R}}_k) \dot{\geq} \rho^{\frac{-\epsilon T}{\kappa}}, \tag{34}$$

and given that $\xi = \sqrt{z \log \rho}$ for some finite z, then

$$N_k \dot{\leq} \left(\sqrt{k} + \frac{2\sqrt{z \log \rho}}{\rho^{\frac{-\epsilon T}{\kappa}}}\right)^k \doteq \rho^{\frac{\epsilon T k}{\kappa}}, \tag{35}$$

which guarantees that the total number of visited lattice points is upper bounded as

$$N_{SD} = \sum_{k=1}^{\kappa} N_k \dot{\leq} \sum_{k=1}^{\kappa} \rho^{\frac{\epsilon T k}{\kappa}} \doteq \rho^{\epsilon T}. \tag{36}$$

Consequently, directly from Lemma 1, we have that

$$\mathrm{P}\left(N_{SD} \dot{\geq} \rho^{\epsilon T}\right) \dot{\leq} \rho^{-d_L(r-\epsilon)}. \tag{37}$$

A similar approach deals with the complexity of the LLL algorithm, which is known (cf. [23]) to be generally unbounded. Specifically drawing from [8, Lemma 2], under the natural assumption of power-limited channels[13] (cf. [8]), under the natural assumption that $d_L(r-\epsilon) > d_L(r)$ for all $\epsilon > 0$, and for N_{LR} denoting the number of flops spent by the LLL algorithm, one can readily conclude that

$$\mathrm{P}\left(N_{LR} \geq \gamma \log \rho\right) \dot{\leq} \rho^{-d_L(r-\epsilon)}, \tag{38}$$

for any $\gamma > \frac{1}{2}(d_L(r-\epsilon))$. Consequently the overall complexity

$$N \doteq N_{SD} + N_{LR},$$

in flops, for the LR-aided MMSE preprocessed lattice sphere decoder, satisfies the following

$$\begin{aligned} \mathrm{P}\left(N \dot{\geq} \rho^{\epsilon T}\right) &\doteq \mathrm{P}\left(\{N_{SD} \dot{\geq} \rho^{\epsilon T}\} \cup \{N_{LR} \dot{\geq} \rho^{\epsilon T}\}\right) \\ &\dot{\leq} \rho^{-d_L(r-\epsilon)}. \end{aligned} \tag{39}$$

Now going back to (5), and having in mind appropriate timeout policies that bound $N_{\max}$ while at the same time specifically guarantee a *vanishing error performance gap* to the exact solution of regularized lattice decoding, we can see that the complexity exponent $c(r)$ takes the equivalent form recently introduced (for the ML case) in [1]

$$c(r) = \inf\{x \mid -\lim_{\rho \to \infty} \frac{\log \mathrm{P}\left(N \geq \rho^x\right)}{\log \rho} > d_L(r)\}. \tag{40}$$

To see this we quickly note that for $N_{\max} = \rho^x$ where $x = c(r) - \delta$ for any $\delta > 0$, it is the case that (cf. (9)) $\lim_{\rho \to \infty} \frac{\mathrm{P}(N \geq \rho^x)}{\mathrm{P}(\hat{\mathbf{x}}_L \neq \mathbf{x})} \to \infty$.

Finally applying (39) we see that for any positive $\epsilon_1 < \epsilon$, it is the case that

$$c(r) = \inf\{\epsilon \mid -\lim_{\rho \to \infty} \frac{\log \mathrm{P}\left(N \geq \rho^{\epsilon T + \epsilon_1}\right)}{\log \rho} > d_L(r)\} \tag{41}$$

which vanishes arbitrarily close to zero, resulting in a zero complexity exponent.

What remains is to consider the error-performance gap in the presence the LR-aided regularized lattice SD with a timeout policy that interrupts at $N_{\max} = \rho^x$ for any vanishingly small $x > 0$.

[12] Even though the work here focuses on decoding, we can also quickly state the obvious fact that the cost of constructing the codewords is also negligible in the scale of interest because it again only involves a finite-dimensional linear transformation (cf. (6)).

[13] This is a moderate assumption that asks that $\mathrm{E}\left\{\|\boldsymbol{H}\|_{\mathrm{F}}^2\right\} \dot{\leq} \rho$. We note that this holds true for any telecommunications setting.

B. Gap to the exact solution of MMSE-preprocessed lattice decoding

We here prove that the LR-aided regularized lattice sphere decoder and the associated time-out policies that guarantee a vanishing complexity exponent, also guarantee a vanishing gap to the error performance of the exact lattice decoding implementation. This result is motivated by potentially exponential gaps in the performance of other DMT optimal decoders (cf. [8]), where these gaps may grow exponentially up to $2^{\frac{\kappa}{2}}$ (cf. [24]) or may potentially be unbounded [25].

Towards establishing this gap, we recall that the exact MMSE-preprocessed lattice decoder in (13) makes errors when $\hat{\mathbf{s}}_{r-ld} \neq \mathbf{s}$. On the other hand the LLL-reduced MMSE-preprocessed lattice sphere decoder with run-time constraints, in addition to making the same errors ($\hat{\mathbf{s}}_{r-lr-ld} \neq \mathbf{s}$), also makes errors when the run-time limit of ρ^x flops becomes active, i.e., when $N \geq \rho^x$, as well as when a small search radius causes $\mathcal{N}_\kappa = \emptyset$. Consequently the corresponding performance gap to the exact regularized decoder, takes the form

$$g_L(x) = \lim_{\rho\to\infty} \frac{\mathrm{P}\left(\{\hat{\mathbf{s}}_{r-lr-ld} \neq \mathbf{s}\} \cup \{N \geq \rho^x\} \cup \{\mathcal{N}_\kappa = \emptyset\}\right)}{\mathrm{P}\left(\hat{\mathbf{s}}_{r-ld} \neq \mathbf{s}\right)}.$$

To bound the above gap, we apply the union bound and the fact that

$$\mathrm{P}\left(\mathcal{N}_\kappa = \emptyset\right) \leq \mathrm{P}\left(\|\mathbf{w}''\| > \xi\right)$$

to get that

$$\begin{aligned} g_L(x) \quad \leq \quad & \lim_{\rho\to\infty} \frac{\mathrm{P}\left(\hat{\mathbf{s}}_{r-lr-ld} \neq \mathbf{s}\right)}{\mathrm{P}\left(\hat{\mathbf{s}}_{r-ld} \neq \mathbf{s}\right)} + \lim_{\rho\to\infty} \frac{\mathrm{P}\left(N \geq \rho^x\right)}{\mathrm{P}\left(\hat{\mathbf{s}}_{r-ld} \neq \mathbf{s}\right)} \\ & + \lim_{\rho\to\infty} \frac{\mathrm{P}\left(\|\mathbf{w}''\| > \xi\right)}{\mathrm{P}\left(\hat{\mathbf{s}}_{r-ld} \neq \mathbf{s}\right)}. \end{aligned} \tag{42}$$

Furthermore from (30) we observe that

$$\mathrm{P}\left(\hat{\mathbf{s}}_{r-lr-ld} \neq \mathbf{s}\right) = \mathrm{P}\left(\hat{\mathbf{s}}_{r-ld} \neq \mathbf{s}\right), \tag{43}$$

and from (39) we recall that

$$\mathrm{P}\left(N \dot{\geq} \rho^{\epsilon T}\right) \dot{\leq} \rho^{-d_L(r-\epsilon)}$$

which implies that for any $x > 0$ it holds that

$$\lim_{\rho\to\infty} \frac{\mathrm{P}\left(N \geq \rho^x\right)}{\mathrm{P}\left(\hat{\mathbf{s}}_{r-ld} \neq \mathbf{s}\right)} = 0. \tag{44}$$

Finally the last term in (42) relates to the search radius ξ, and to the behavior of the noise $\mathbf{w}''$ which was shown in (16), (27) to take the form

$$\mathbf{w}'' = \tilde{\mathbf{Q}}^H \left(-\alpha_r^2 \mathbf{R}^{-H}\mathbf{s} + \mathbf{R}^{-H}\mathbf{M}^H\mathbf{w}\right). \tag{45}$$

The following lemma, whose proof is found in Appendix D, accounts for the fact that $\mathbf{w}''$ includes self-interference and colored noise, to bound the last term in (42).

Lemma 2: There exist a finite $z > d_L(r)$ for which a search radius $\xi = \sqrt{z \log \rho}$ guarantees that

$$\lim_{\rho\to\infty} \frac{\mathrm{P}\left(\|\mathbf{w}''\| > \xi\right)}{\mathrm{P}\left(\hat{\mathbf{s}}_{r-ld} \neq \mathbf{s}\right)} = 0. \tag{46}$$

Consequently combining (43), (44) and (46) gives that $g_L(x) = 1, \quad \forall x > 0$. The following directly holds.

Theorem 2: LR-aided MMSE-preprocessed lattice sphere decoding with a computational constraint activated at ρ^x flops, allows for a vanishing gap to the exact solution of MMSE-preprocessed lattice decoding, for any $x > 0$. Equivalently the same LR-aided decoder guarantees that

$$g_L(\epsilon) = 1 \quad \text{and} \quad \lim_{\rho\to\infty} \frac{\log N_{\max}(g)}{\log \rho} = 0 \qquad \forall \epsilon > 0, g \geq 1,$$

for all fading statistics, all MIMO scenarios, and all full-rate lattice codes.

IV. Conclusions

The work identified the first lattice decoding solution that achieves, in the most general outage-limited MIMO setting and the high rate and high SNR limit, both a vanishing gap to the error-performance of the (DMT optimal) exact solution of preprocessed lattice decoding, as well as a computational complexity that is subexponential in the number of codeword bits. The proposed solution employs lattice reduction (LR)-aided regularized lattice sphere decoding and proper timeout policies. As it turns out, lattice reduction is a special ingredient that allows for complexity reductions; a role that was rigorously demonstrated here for the first time, by proving that without lattice reduction, for most common codes, the complexity cost for asymptotically optimal regularized lattice sphere decoding is exponential in the number of codeword bits, and in many cases it in fact matches the complexity cost of ML sphere decoding.

In light of the fact that, prior to this work, a vanishing error performance gap was generally attributed only to near-full lattice searches that have exponential complexity, in conjunction with the fact that subexponential complexity was generally attributed to early-terminated (linear) solutions which have though a performance gap that can be up to exponential in dimension and/or rate, the work constitutes the first proof that subexponential complexity need not come at the cost of exponential reductions in lattice decoding error performance.

Appendix A
Proof for Theorem 1 and Corollary 1a

In the following we begin by providing an upper bound on the complexity exponent of MMSE-preprocessed (unconstrained) lattice sphere decoding, where this bound holds for the general quasi-static MIMO channel, for all fading statistics and for any full-rate lattice code. We will then proceed to provide a lower bound on the complexity exponent of the same decoder, where this bound, under the extra assumptions of regular i.i.d. fading statistics and of layered codes, will in fact match the above mentioned upper bound to prove the theorem and the associated corollaries. Before proceeding with the bounds, we describe the $n_T \times n_R$ ($n_R \geq n_T$) quasi-static point-to-point MIMO channel, and its corresponding association to the general MIMO channel model in (10) and metric in (17).

The aforementioned quasi-static channel model takes the form

$$\mathbf{Y}_C = \sqrt{\rho}\mathbf{H}_C\mathbf{X}_C + \mathbf{W}_C, \tag{47}$$

where $\mathbf{X}_C \in \mathbb{C}^{n_T \times T}$, $\mathbf{Y}_C \in \mathbb{C}^{n_R \times T}$ and $\mathbf{W}_C \in \mathbb{C}^{n_R \times T}$ represent the transmitted, received and noise signals over a period of T time slots, and where $\mathbf{H}_C \in \mathbb{C}^{n_R \times n_T}$ represents the matrix of fade coefficients. The real-valued representation of (47) can be written as

$$\mathbf{y} \quad = \quad \sqrt{\rho}(\mathbf{I}_T \otimes \mathbf{H}_R)\mathbf{x} + \mathbf{w} \tag{48}$$

where $\mathbf{H}_R = \begin{bmatrix} Re\{\mathbf{H}_C\} & -Im\{\mathbf{H}_C\} \\ Im\mathbf{H}_C\} & Re\{\mathbf{H}_C\} \end{bmatrix}$, $\mathbf{x} = (\mathbf{x}_1^{\mathrm{T}}, \cdots, \mathbf{x}_T^{\mathrm{T}})^{\mathrm{T}}$ with $\mathbf{x}_t = [Re\{\mathbf{X}_{t,C}\}^{\mathrm{T}}, Im\{\mathbf{X}_{t,C}\}^{\mathrm{T}}]^{\mathrm{T}}$ for $t = 1, \cdots, T$, where $\mathbf{X}_{t,C}$ is t-th column of $\mathbf{X}_C$, $\mathbf{y}$ and where $\mathbf{w}$ are defined similar to $\mathbf{x}$. The system model in (48) is of the familiar form

$$\mathbf{y} = \sqrt{\rho}\mathbf{H}\mathbf{x} + \mathbf{w} \tag{49}$$

as in (1) with $m = 2n_T T$, $n = 2n_R T$, and where

$$\mathbf{H} = \mathbf{I}_T \otimes \mathbf{H}_R. \tag{50}$$

As before the vectorized codewords $\mathbf{x}$, associated to the full-rate code, take the form

$$\mathbf{x} = \rho^{\frac{-rT}{\kappa}}\mathbf{G}\mathbf{s},\ \mathbf{s} \in \mathbb{Z}^{\kappa} \cap \rho^{\frac{rT}{\kappa}}\mathcal{R}, \tag{51}$$

where $\kappa = 2\min\{n_{\mathrm{T}}, n_{\mathrm{R}}\}T = 2n_{\mathrm{T}}T = m$, which allows us to rewrite the model as

$$\mathbf{y} = \mathbf{M}\mathbf{s} + \mathbf{w}, \tag{52}$$

for

$$\mathbf{M} = \rho^{\frac{1}{2}-\frac{rT}{\kappa}}\mathbf{H}\mathbf{G} = \rho^{\frac{1}{2}-\frac{rT}{\kappa}}(\mathbf{I}_T \otimes \mathbf{H}_R)\mathbf{G}. \tag{53}$$

Finally the corresponding coherent MMSE-preprocessed lattice decoder for the transmitted symbol vector $\mathbf{s}$, can be expressed to be (cf. (17))

$$\hat{\mathbf{s}}_{r-ld} = \arg\min_{\hat{\mathbf{s}} \in \mathbb{Z}^{\kappa}} \|\mathbf{r} - \mathbf{R}\hat{\mathbf{s}}\|^2, \tag{54}$$

where $\mathbf{r} = \mathbf{Q}_1^H\mathbf{y}$ and $\mathbf{R} \in \mathbb{C}^{\kappa \times \kappa}$ is the upper-triangular matrix, where furthermore both $\mathbf{Q}_1$ and $\mathbf{R}$ result from the thin QR decomposition of the $(n+\kappa) \times \kappa$ dimensional preprocessed channel matrix

$$\mathbf{M}^{reg} \triangleq \begin{bmatrix} \mathbf{M} \\ \alpha_r \mathbf{I} \end{bmatrix} = \mathbf{Q}\mathbf{R} = \begin{bmatrix} \mathbf{Q}_1 \\ \mathbf{Q}_2 \end{bmatrix} \mathbf{R} \tag{55}$$

and where as before $\alpha_r = \rho^{\frac{-rT}{\kappa}}$.

A. Upper bound on complexity of regularized lattice SD

In establishing the upper bound, we consider Lemma 1 in [1], which we properly modify to account for MMSE preprocessing and for the removal of the constellation boundaries, and get that the number N_k of nodes visited at layer k by the MMSE-preprocessed lattice sphere decoder, is upper bounded as

$$N_k = |\mathcal{N}_k| \leq \prod_{i=1}^{k} \left[\sqrt{2k} + \frac{2\xi}{\sigma_i(\mathbf{R}_k)}\right], \tag{56}$$

where $\sigma_i(\mathbf{R}_k)$, $i = 1, \cdots, k$ denote the singular values of $\mathbf{R}_k$ in increasing order.

Towards lower bounding $\sigma_i(\mathbf{R}_k)$, we note that

$$\sigma_i(\mathbf{R}_k) \geq \sigma_i(\mathbf{R}) = \sigma_i(\mathbf{M}^{reg}) = \sqrt{\alpha_r^2 + \sigma_i(\mathbf{M}^H\mathbf{M})}, \tag{57}$$

where the first inequality makes use of the interlacing property of singular values of sub-matrices [26]. Furthermore for

$$\mu_j \triangleq -\frac{\log \sigma_j(\mathbf{H}_C^H\mathbf{H}_C)}{\log \rho},\ j = 1, \cdots, n_T \tag{58}$$

and $\mu_1 \geq \cdots \geq \mu_{n_T}$, we see that $\sigma_j(\mathbf{H}_C) = \rho^{-\frac{1}{2}\mu_j}$, and from (53) that

$$\begin{aligned} \sigma_i(\mathbf{M}) &\geq \rho^{\frac{1}{2}-\frac{rT}{\kappa}}\sigma_{\min}(\mathbf{G})\sigma_{(i)}(\mathbf{I}_T \otimes \mathbf{H}_R)) \\ &\doteq \rho^{\frac{1}{2}-\frac{rT}{\kappa}}\sigma_{l_{2T}(i)}(\mathbf{H}_C) \\ &= \rho^{\frac{-rT}{\kappa}+\frac{1}{2}(1-\mu_{l_{2T}(i)})}, \end{aligned} \tag{59}$$

where $l_T(i) \triangleq \left\lceil \frac{i}{T} \right\rceil$, and where the asymptotic equality is due to the fact that $\sigma_{\min}(\mathbf{G}) \doteq \rho^0$. Substituting from (59) in (57) we now have that

$$\sigma_i(\mathbf{R}_k) \dot{\geq} \rho^{\frac{-rT}{\kappa}+\frac{1}{2}(1-\mu_{l_{2T}(i)})^+},\quad i = 1, \cdots, \kappa. \tag{60}$$

Corresponding to (56) we see that

$$\left[\sqrt{2k} + \frac{2\xi}{\sigma_i(\mathbf{R}_k)}\right] \dot{\leq} \rho^{\left(\frac{rT}{\kappa}-\frac{1}{2}(1-\mu_{l_{2T}(i)})^+\right)^+},$$

for any $i = 1, \cdots, 2n_T T$, and from (56) we have that

$$N_k(\boldsymbol{\mu}) \dot{\leq} \rho^{\sum_{i=1}^{k}\left(\frac{rT}{\kappa}-\frac{1}{2}(1-\mu_{l_{2T}(i)})^+\right)^+}, \tag{61}$$

where $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_{n_T})$. It follows that

$$\begin{aligned} N_{SD}(\boldsymbol{\mu}) = \sum_{k=1}^{\kappa} N_k(\boldsymbol{\mu}) &\dot{\leq} \sum_{k=1}^{\kappa} \rho^{\sum_{i=1}^{k}\left(\frac{rT}{\kappa}-\frac{1}{2}(1-\mu_{l_{2T}(i)})^+\right)^+} \\ &\doteq \rho^{\sum_{i=1}^{\kappa}\left(\frac{rT}{\kappa}-\frac{1}{2}(1-\mu_{l_{2T}(i)})^+\right)^+} \\ &\doteq \rho^{T\sum_{j=1}^{n_T}\left(\frac{r}{n_T}-(1-\mu_j)^+\right)^+}, \end{aligned} \tag{62}$$

where the last asymptotic equality is due to the multiplicity of the singular values.

Now consider the set

$$\mathcal{T}(x) \triangleq \left\{ \boldsymbol{\mu} \mid T\sum_{j=1}^{n_T}\left(\frac{r}{n_T} - (1-\mu_j)^+\right)^+ \geq x \right\}, \tag{63}$$

and note that for any $y < x$, then (62) and $\boldsymbol{\mu} \notin \mathcal{T}(y)$ jointly imply that $N_{SD} < \rho^x$, which in turn implies that $\mathrm{P}\left(\boldsymbol{\mu} \notin \mathcal{T}(y)\right) \leq \mathrm{P}\left(N_{SD} < \rho^x\right)$ and consequently that

$$-\lim_{\rho\to\infty}\frac{\log \mathrm{P}\left(N_{SD} \geq \rho^x\right)}{\log \rho} \geq -\lim_{\rho\to\infty}\frac{\log \mathrm{P}\left(\boldsymbol{\mu} \in \mathcal{T}(y)\right)}{\log \rho}. \tag{64}$$

In evaluating the right hand side of (64) we note that $\mathcal{T}(y)$ is a closed set and thus, applying the large deviation principle (cf. [27]), we have that

$$-\lim_{\rho\to\infty}\frac{\log \mathrm{P}\left(\boldsymbol{\mu} \in \mathcal{T}(y)\right)}{\log \rho} \geq \inf_{\boldsymbol{\mu} \in \mathcal{T}(y)} I(\boldsymbol{\mu}) \tag{65}$$

for some rate function $I(\boldsymbol{\mu})$. Consequently from (64) and (65), it follows that

$$-\lim_{\rho\to\infty}\frac{\log \mathrm{P}\left(N_{SD}\geq\rho^{x}\right)}{\log\rho}\geq\inf_{\boldsymbol{\mu}\in\mathcal{T}(y)}I(\boldsymbol{\mu}). \tag{66}$$

This lower bound specified in (66) holds for any $y<x$. Consequently to get the tightest possible bound, we need to find $\sup_{y<x}\inf_{\boldsymbol{\mu}\in\mathcal{T}(y)}I(\boldsymbol{\mu})$. As $\inf_{\boldsymbol{\mu}\in\mathcal{T}(y)}I(\boldsymbol{\mu})$ is non-decreasing and left-continuous in y, it follows that

$$\sup_{y<x}\inf_{\boldsymbol{\mu}\in\mathcal{T}(y)}I(\boldsymbol{\mu})=\inf_{\boldsymbol{\mu}\in\mathcal{T}(x)}I(\boldsymbol{\mu}).$$

Consequently

$$-\lim_{\rho\to\infty}\frac{\log \mathrm{P}\left(N_{SD}\geq\rho^{x}\right)}{\log\rho}\geq\inf_{\boldsymbol{\mu}\in\mathcal{T}(x)}I(\boldsymbol{\mu}), \tag{67}$$

which in conjunction with (40) gives that

$$\begin{aligned}c_{r-ld}(r)\leq\overline{c}_{r-ld}(r)&\triangleq\inf\{x|\inf_{\boldsymbol{\mu}\in\mathcal{T}(x)}I(\boldsymbol{\mu})>d_{L}(r)\}\\&=\sup\{x|\inf_{\boldsymbol{\mu}\in\mathcal{T}(x)}I(\boldsymbol{\mu})\leq d_{L}(r)\}\\&=\max\{x|\inf_{\boldsymbol{\mu}\in\mathcal{T}(x)}I(\boldsymbol{\mu})\leq d_{L}(r)\}\end{aligned} \tag{68}$$

where the above follows from the aforementioned fact that $-\lim_{\rho\to\infty}\frac{\log \mathrm{P}(N_{SD}\geq\rho^{x})}{\log\rho}$ (and by extension also $\inf_{\boldsymbol{\mu}\in\mathcal{T}(x)}I(\boldsymbol{\mu})$) is continuous and nondecreasing in x, and from the fact that $\mathcal{T}(x)$ is a closed set. Consequently $\overline{c}_{r-ld}(r)$ takes the form

$$\overline{c}_{r-ld}(r)\triangleq\max_{\boldsymbol{\mu}}\ x \tag{69a}$$

$$\text{s.t.}\ \ T\sum_{j=1}^{n_T}\left(\frac{r}{n_T}-(1-\mu_j)^{+}\right)^{+}\geq x, \tag{69b}$$

$$I(\boldsymbol{\mu})\leq d_{L}(r), \tag{69c}$$

$$\mu_1\geq\cdots\geq\mu_{n_T}\geq 0. \tag{69d}$$

Furthermore since $\mathcal{T}(x)$ is a closed set, the maximum x in (69) must be such that (69b) is satisfied with equality, in which case $\overline{c}_{r-ld}(r)$ can be obtained as the solution to a constrained maximization problem according to

$$\overline{c}_{r-ld}(r)\triangleq\max_{\boldsymbol{\mu}}\ T\sum_{j=1}^{n_T}\left(\frac{r}{n_T}-(1-\mu_j)^{+}\right)^{+} \tag{70a}$$

$$\text{s.t.}\ \ I(\boldsymbol{\mu})\leq d_{L}(r), \tag{70b}$$

$$\mu_1\geq\cdots\geq\mu_{n_T}\geq 0. \tag{70c}$$

Equivalently for $\boldsymbol{\mu}^{*}=(\mu_1^{*},\cdots,\mu_{n_T}^{*})$ being one of the maximizing vectors[14], i.e., such that $\boldsymbol{\mu}^{*}\in\mathcal{T}(x)$ and $I(\boldsymbol{\mu}^{*})=d_{L}(r)$, then $\overline{c}_{r-ld}(r)$ takes the form

$$\overline{c}_{r-ld}(r)=T\sum_{j=1}^{n_T}\left(\frac{r}{n_T}-(1-\mu_j^{*})^{+}\right)^{+}. \tag{71}$$

As we will now show, the above bound is also shared by the ML-based sphere decoder, with or without MMSE preprocessing, irrespective of the full-rate code and the fading statistics. Directly from [1, Theorem 2], and taking into consideration that MMSE-preprocessed lattice decoding is DMT optimal for any code [8], we recall that the equivalent upper bound for the ML-based sphere decoder, without MMSE preprocessing, takes the form

$$\overline{c}_{ml}(r)\triangleq\max_{\boldsymbol{\mu}}\ T\sum_{j=1}^{n_T}\min\left(\frac{r}{n_T}-1+\mu_j,\frac{r}{n_T}\right)^{+} \tag{72a}$$

$$\text{s.t.}\ \ I(\boldsymbol{\mu})\leq d_{L}(r), \tag{72b}$$

$$\mu_1\geq\cdots\geq\mu_{n_T}\geq 0. \tag{72c}$$

Comparing (70) and (72) we are able to conclude that both the objective functions (70a) and (72a) as well as both pairs of constraints are identical. To see this, we first note that for $0\leq\mu_j\leq 1$, then

$$\begin{aligned}\min\left(\frac{r}{n_T}-1+\mu_j,\frac{r}{n_T}\right)^{+}&=\left(\frac{r}{n_T}-1+\mu_j\right)^{+},\\\left(\frac{r}{n_T}-(1-\mu_j)^{+}\right)^{+}&=\left(\frac{r}{n_T}-1+\mu_j\right)^{+},\end{aligned}$$

and furthermore we note that for $\mu_j>1$, then

$$\min\left(\frac{r}{n_T}-1+\mu_j,\frac{r}{n_T}\right)^{+}=\left(\frac{r}{n_T}-(1-\mu_j)^{+}\right)^{+}=\frac{r}{n_T},$$

which proves that $\overline{c}_{ml}(r)$ and $\overline{c}_{r-ld}(r)$ are identical.

In considering the case of MMSE-preprocessed ML SD, it is easy to see that the summands in the objective function in (72a) will be modified to take the form $\min\left(\frac{r}{n_T}-(1-\mu_j)^{+},\frac{r}{n_T}\right)^{+}$ which can be seen to match (70a) for all $\mu_j\geq 0$, which in turn concludes the proof that the upper bound $\overline{c}_{r-ld}(r)$ for MMSE-preprocessed lattice SD is also shared by the ML-based sphere decoder, with or without MMSE preprocessing, irrespective of the full-rate code, and for all fade statistics represented by monotonic rate functions.

B. Lower bound on complexity of regularized lattice SD

We will here, under the extra assumptions of regular i.i.d. fading statistics and of layered codes with natural decoding order, provide a lower bound that matches the upper bound in (71). The same bound and tightness will also apply to any full-rate code, under the assumption of a fixed, worst case decoding ordering.

The goal here is to show that at layer $k=2qT$, for some $q\in[1,n_{\mathrm{T}}]$, the sphere decoder visits close to $\rho^{\overline{c}_{r-ld}(r)}$ nodes with a probability that is large compared to the probability of decoding error $\mathrm{P}\left(\mathbf{s}_{L}\neq\mathbf{s}\right)\doteq\rho^{-d_{L}(r)}$, which from the expression of the complexity exponent (40), will prove that $c_{r-ld}(r)=\overline{c}_{r-ld}(r)$.

Going back to (71), we let q be the largest integer for which

$$\frac{r}{n_T}-(1-\mu_q^{*})^{+}>0, \tag{73}$$

in which case (71) takes the form

$$\overline{c}_{r-ld}(r)=T\sum_{j=1}^{q}\frac{r}{n_T}-(1-\mu_j^{*})^{+}. \tag{74}$$

We recall from (58) that $\mu_j=-\frac{\log\sigma_j(\mathbf{H}_C^{H}\mathbf{H}_C)}{\log\rho}$, $j=1,\cdots,n_T$, and that $\boldsymbol{\mu}^{*}\in\mathcal{T}(x)$ satisfies $I(\boldsymbol{\mu}^{*})=d_{L}(r)$ and

[14]In general, (70) does not have a unique optimal point because $(a)^{+}$ is constant in a for $a\leq 0$.

maximizes (70a). We also note that without loss of generality we can assume that $q \geq 1$ as otherwise $\overline{c}_{r-ld}(r) = 0$ (cf. (71)). Consequently it is the case that $\mu_j^* > 0$ for $j = 1, \cdots, q$. Furthermore given the monotonicity of the rate function $I(\boldsymbol{\mu})$, and the fact that the objective function in (70) does not increase in μ_j beyond $\mu_j = 1$, we may also assume without loss of generality that $\mu_j^* \leq 1$ for $j = 1, \cdots, n_T$.

As in [1] we proceed to define two events Ω_1 and Ω_2 which we will prove to be jointly sufficient so that, at layer $k = 2qT$, the sphere decoder visits close to $\rho^{\overline{c}_{r-ld}(r)}$ nodes. These are given by

$$\begin{aligned}\Omega_1 \triangleq \{&\mu_j^* - 2\delta < \mu_j < \mu_j^* - \delta, j = 1, \cdots, q \\ &0 < \mu_j < \delta, j = q+1, \cdots, n_T\},\end{aligned} \tag{75}$$

for a given small $\delta > 0$, and

$$\Omega_2 \triangleq \{\sigma_1\left((\mathbf{I}_T \otimes \mathbf{V}_p^H)\mathbf{G}_{|p}\right) \geq u\}, \tag{76}$$

for some given $u > 0$, where for $p \triangleq n_T - q$ then $\mathbf{G}_{|p}$ denotes the first $2pT$ columns of $\mathbf{G}$, and where $\mathbf{V}_p$ denotes the last $2p$ columns of $\mathbf{V}$ obtained by applying the singular value decomposition on $\mathbf{H}_R$, i.e., $\mathbf{H}_R = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^H$, where

$$\boldsymbol{\Sigma} \triangleq \mathrm{diag}\{\sigma_1(\mathbf{H}_R), \cdots, \sigma_{2n_T}(\mathbf{H}_R)\}$$

with $\sigma_1(\mathbf{H}_R) \leq \cdots \leq \sigma_{2n_T}(\mathbf{H}_R)$ and $\mathbf{V}\mathbf{V}^H = \mathbf{I}$. Hence, $\mathbf{V}_p^H$ corresponds to the $2p$ largest singular values of $\mathbf{H}_R$.

Note also that by choosing δ sufficiently small, and using the fact that $\mu_i^* > 0$ for $i = 1, \cdots, q$, we may without loss of generality assume that Ω_1 implies that $\mu_j > 0$ for all $j = 1, \cdots, n_T$.

Modifying the approach in [1, Theorem 1] to account for MMSE preprocessing and unconstrained decoding, the lower bound on the number of nodes visited at layer k by the sphere decoder, is given by

$$N_k \geq \prod_{i=1}^{k} \left[\frac{2\xi}{\sqrt{k}\sigma_i(\mathbf{R}_k)} - \sqrt{k}\right]^+. \tag{77}$$

In the following, and up until (83), we will work towards upper bounding $\sigma_i(\mathbf{R}_k)$ so that we can then lower bound N_k.

Towards this let

$$\mathbf{M}_{|p}^{reg} \triangleq \begin{bmatrix} \rho^{\frac{1}{2}-\frac{rT}{\kappa}}\mathbf{H}\mathbf{G}_{|p} \\ \alpha_r \mathbf{I}_{|p} \end{bmatrix} \in \mathbb{R}^{2(n_R+n_T)T \times 2pT}$$

contain the first $2pT$ columns of $\mathbf{M}^{reg}$ from (55), and note that

$$(\mathbf{M}_{|p}^{reg})^H \mathbf{M}_{|p}^{reg} = \rho^{1-\frac{2rT}{\kappa}} \mathbf{G}_{|p}^H \mathbf{H}^H \mathbf{H} \mathbf{G}_{|p} + \alpha_r^2 \mathbf{I}\,,$$

and that from (50) we get

$$(\mathbf{M}_{|p}^{reg})^H \mathbf{M}_{|p}^{reg} = \rho^{1-\frac{2rT}{\kappa}} \mathbf{G}_{|p}^H (\mathbf{I}_T \otimes \mathbf{H}_R^H \mathbf{H}_R) \mathbf{G}_{|p} + \alpha_r^2 \mathbf{I}.$$

Since

$$\begin{aligned}\mathbf{H}_R^H \mathbf{H}_R &= \mathbf{V}(\mathrm{diag}\{\sigma_1(\mathbf{H}_R^H\mathbf{H}_R), \cdots, \sigma_{2n_T}(\mathbf{H}_R^H\mathbf{H}_R)\})\mathbf{V}^H \\ &= \mathbf{V}(\mathrm{diag}\{\sigma_1(\mathbf{H}_R^H\mathbf{H}_R), \cdots, \sigma_{2n_T}(\mathbf{H}_R^H\mathbf{H}_R)\} \\ &\quad - \sigma_{(2q+1)}(\mathbf{H}_R^H\mathbf{H}_R)\mathrm{diag}\{\underbrace{0, \cdots, 0}_{2q}, \underbrace{1, \cdots, 1}_{2p}\})\mathbf{V}^H \\ &\quad + \sigma_{(2q+1)}(\mathbf{H}_R^H\mathbf{H}_R)\mathbf{V}(\mathrm{diag}\{\underbrace{0, \cdots, 0}_{2q}, \underbrace{1, \cdots, 1}_{2p}\})\mathbf{V}^H,\end{aligned}$$

we have that

$$\begin{aligned}\mathbf{H}_R^H \mathbf{H}_R &\succeq \sigma_{(2q+1)}(\mathbf{H}_R^H\mathbf{H}_R)\mathbf{V}(\mathrm{diag}\{\underbrace{0, \cdots, 0}_{2q}, \underbrace{1, \cdots, 1}_{2p}\})\mathbf{V}^H \\ &= \sigma_{(2q+1)}(\mathbf{H}_R^H\mathbf{H}_R)\mathbf{V}(\mathrm{diag}\{\underbrace{0, \cdots, 0}_{2q}, \underbrace{1, \cdots, 1}_{2p}\}) \\ &\qquad (\mathrm{diag}\{\underbrace{0, \cdots, 0}_{2q}, \underbrace{1, \cdots, 1}_{2p}\})\mathbf{V}^H \\ &= \sigma_{(2q+1)}(\mathbf{H}_R^H\mathbf{H}_R)\mathbf{V}_p\mathbf{V}_p^H\end{aligned}$$

where the last equality follows from the fact that $\mathbf{V}_p$ contains the last $2p$ columns of $\mathbf{V}$ and where $\mathbf{A} \succeq \mathbf{B}$ denotes that $\mathbf{A} - \mathbf{B}$ is positive-semidefinite. Since $\sigma_i(\mathbf{H}^H\mathbf{H}) \in \mathbb{R}$ and since the Kronecker product induces singular value multiplicity, it follows that

$$\begin{aligned}&(\mathbf{M}_{|p}^{reg})^H \mathbf{M}_{|p}^{reg} \\ \succeq\ &\rho^{1-\frac{2rT}{\kappa}} \sigma_{(2q+1)}(\mathbf{H}_R^H\mathbf{H}_R)\mathbf{G}_{|p}^H(\mathbf{I}_T \otimes \mathbf{V}_p\mathbf{V}_p^H)\mathbf{G}_{|p} + \alpha_r^2\mathbf{I}.\end{aligned}$$

With respect to the smallest singular value of $(\mathbf{M}_{|p}^{reg})^H \mathbf{M}_{|p}^{reg}$ we have

$$\begin{aligned}\sigma_1((\mathbf{M}_{|p}^{reg})^H \mathbf{M}_{|p}^{reg}) \geq\ &\rho^{1-\frac{2rT}{\kappa}} \sigma_{(2q+1)}(\mathbf{H}_R^H\mathbf{H}_R) \cdot \\ &\sigma_1\left(\mathbf{G}_{|p}^H(\mathbf{I}_T \otimes \mathbf{V}_p\mathbf{V}_p^H)\mathbf{G}_{|p}\right) + \alpha_r^2\end{aligned}$$

and consequently, given that $\mathbf{H}_R \in \Omega_2$, we have that

$$\begin{aligned}\sigma_1(\mathbf{M}_{|p}^{reg}) &\geq \rho^{-\frac{rT}{\kappa}} \sqrt{u^2\rho\sigma_{l_2(2q+1)}(\mathbf{H}_C^H\mathbf{H}_C) + 1} \\ &\doteq \rho^{-\frac{rT}{\kappa}} \rho^{\frac{1}{2}(1-\mu_{q+1})^+} \\ &\geq \rho^{-\frac{rT}{\kappa}+\frac{1}{2}(1-\delta)^+},\end{aligned} \tag{78}$$

where the first inequality follows from (76), the exponential equality follows from (58) and from the fact that $u > 0$ is fixed and independent of ρ, and the last inequality follows from (75).

From (53) we have that

$$\begin{aligned}\sigma_i(\mathbf{M}^{reg}) &\leq \rho^{\frac{-rT}{\kappa}} \sqrt{(1 + \rho(\sigma_\kappa(\mathbf{G})\sigma_{l_{2T}(i)}(\mathbf{H}_C))^2)} \\ &\doteq \rho^{\frac{-rT}{\kappa}+\frac{1}{2}(1-\mu_{l_{2T}(i)})^+}, \quad i = 1, \cdots, 2n_TT,\end{aligned} \tag{79}$$

where the asymptotic equality follows from the fact that $\sigma_\kappa(\mathbf{G})$ is fixed and independent of ρ. Furthermore (75) gives that for $i = 1, \cdots, 2qT$ then

$$\sigma_i(\mathbf{M}^{reg}) \dot{\leq} \rho^{-\frac{rT}{\kappa}+\delta+\frac{1}{2}(1-\mu^*_{l_{2T}(i)})^+}, \tag{80}$$

where we have made use of the fact that $\mu_j^* \leq 1$ for $j = 1, \cdots, n_T$.

Given that $\mu_j^* > 0$ for $j = 1, \cdots, q$, then for sufficiently small δ and for $i = 1, \cdots, 2qT$, we have that

$$-\frac{rT}{\kappa} + \frac{1}{2}(1-\delta)^+ \geq -\frac{rT}{\kappa} + \delta + \frac{1}{2}(1-\mu^*_{l_{2T}(i)})^+,$$

which means that for sufficiently small δ, a comparison of (78) and (80) yields

$$\sigma_i(\mathbf{M}^{reg}) < \sigma_1(\mathbf{M}_{|p}^{reg}),$$

for $i = 1, \cdots, 2qT$. The above inequality allows us to apply Lemma 3 in [1], which in turn gives that

$$\sigma_i(\mathbf{R}_k) \leq \left[\frac{\sigma_\kappa(\mathbf{M}^{reg})}{\sigma_1(\mathbf{M}^{reg}_{|p})} + 1\right] \sigma_i(\mathbf{M}^{reg}), \tag{81}$$

for $i = 1, \cdots, 2qT$.

Setting $i = \kappa$ in (79) upper bounds the maximum singular value of $\mathbf{M}^{reg}$ as

$$\sigma_\kappa(\mathbf{M}^{reg}) \dot{\leq} \rho^{-\frac{rT}{\kappa}+\frac{1}{2}(1-\mu_{n_T})^+} \leq \rho^{\frac{1}{2}-\frac{rT}{\kappa}}, \tag{82}$$

where the last inequality is due to the fact that $\mu_j \geq 0$. Consequently combining (82) and (78) gives that

$$\left[\frac{\sigma_\kappa(\mathbf{M}^{reg})}{\sigma_1(\mathbf{M}^{reg}_{|p})} + 1\right] \dot{\leq} \rho^{\frac{1}{2}\delta},$$

which together with (80) and (81) gives that

$$\sigma_i(\mathbf{R}_k) \dot{\leq} \rho^{-\frac{rT}{\kappa}+\frac{3}{2}\delta+\frac{1}{2}(1-\mu^*_{l_{2T}(i)})^+}, \quad i = 1, \cdots, 2qT. \tag{83}$$

Consequently, going back to (77), we have that

$$\left[\frac{2\xi}{\sqrt{k}\sigma_i(\mathbf{R}_k)} - \sqrt{k}\right]^+ \dot{\geq} \rho^{\left(\frac{rT}{\kappa}-\frac{3}{2}\delta-\frac{1}{2}(1-\mu^*_{l_{2T}(i)})^+\right)} > 0 \tag{84}$$

and furthermore for $i = 1, \cdots, 2qT$, we have that $\frac{rT}{\kappa} - \frac{3}{2}\delta - \frac{1}{2}(1-\mu^*_{l_{2T}(i)})^+ > 0$ directly from definition of q and for sufficiently small δ. As a result, for $k \leq 2qT$ we have that

$$N_k \dot{\geq} \prod_{i=1}^{k} \rho^{\left(\frac{rT}{\kappa}-\frac{3}{2}\delta-\frac{1}{2}(1-\mu^*_{l_{2T}(i)})\right)} \tag{85}$$

$$= \rho^{\sum_{i=1}^{k}\left(\frac{rT}{\kappa}-\frac{1}{2}(1-\mu^*_{l_{2T}(i)})^+\right)-\frac{3}{2}k\delta}, \tag{86}$$

and setting $k = 2qT$ we have that

$$N_{2qT} \dot{\geq} \rho^{\left(\sum_{i=1}^{2qT}\left(\frac{rT}{\kappa}-\frac{1}{2}(1-\mu^*_{l_{2T}(i)})^+\right)-3qT\delta\right)} \tag{87}$$

$$= \rho^{\left(T\sum_{j=1}^{q}\left(\frac{rT}{\kappa}-(1-\mu^*_j)^+\right)-3qT\delta\right)} \tag{88}$$

$$= \rho^{(\overline{c}_{r-ld}(r)-3qT\delta)}, \tag{89}$$

where the last equality follows from (74). Consequently

$$N_{SD} \geq N_{2qT} \dot{\geq} \rho^{\overline{c}_{r-ld}(r)-3qT\delta},$$

for small $\delta > 0$. Given that δ can be chosen arbitrarily small, and given that events Ω_1 and Ω_2 occur, then the number of nodes visited by the SD at layer $2qT$ is arbitrarily close to the upper bound of $\rho^{\overline{c}_{r-ld}(r)}$.

Now to show that $c_{r-ld}(r) \geq \overline{c}_{r-ld}(r) - 3qT\delta$, we just have to prove that $-\lim_{\rho\to\infty} \frac{\mathrm{P}\left(N_{SD} \dot{\geq} \rho^{\overline{c}_{r-ld}(r)-3qT\delta}\right)}{\log\rho} < d_L(r)$. Toward this we note that as (75) and (76) imply that $N_{SD} \dot{\geq} \rho^{\overline{c}_{r-ld}(r)-3qT\delta}$, it follows that

$$\mathrm{P}\left(N_{SD} \dot{\geq} \rho^{\overline{c}_{r-ld}(r)-3qT\delta}\right) \geq \mathrm{P}(\Omega_1 \cap \Omega_2) = \mathrm{P}(\Omega_1)\,\mathrm{P}(\Omega_2)$$

where the equality follows from the i.i.d. assumption on the entries in $\mathbf{H}_C$, which makes the singular values of $\mathbf{H}_C^H\mathbf{H}_C$ independent of the singular vectors of $\mathbf{H}_C^H\mathbf{H}_C$ [28], [29], and which in turn also implies independence of the singular values of $\mathbf{H}_C^H\mathbf{H}_C$ (event Ω_1) from the singular vectors of $\mathbf{H}_R^H\mathbf{H}_R$ (event Ω_2).

We now turn to [1, Lemma 2] and recall that for the layered codes assumed here, as well as for any full-rate design and some non-random fixed decoding ordering (corresponding to a permutation of the columns of $\boldsymbol{G}$), there exists a unitary matrix $\mathbf{V}_p^{'}$ such that $\mathrm{rank}\Big((\mathbf{I}_T \otimes (\mathbf{V}_p^{'})^H)\mathbf{G}_{|p}\Big) = 2pT$ i.e., that

$$\sigma_1\left((\mathbf{I}_T \otimes (\mathbf{V}_p^{'})^H)\mathbf{G}_{|p}\right) > 0.$$

However, by continuity of singular values [26] it follows for sufficiently small $u > 0$ (cf.(76)) that $\mathrm{P}(\Omega_2) > 0$, which implies[15] that $\mathrm{P}(\Omega_2) \doteq \rho^0$ as Ω_2 is independent of ρ. This in turn implies that

$$\mathrm{P}\left(N_{SD} \dot{\geq} \rho^{\overline{c}_{r-ld}(r)-3qT\delta}\right) \dot{\geq} \mathrm{P}(\Omega_1). \tag{90}$$

With Ω_1 being an open set, we have that

$$\begin{aligned} -\lim_{\rho\to\infty}\frac{\mathrm{P}(\Omega_1)}{\log\rho} &\leq \inf_{\boldsymbol{\mu}\in\Omega_1} I(\boldsymbol{\mu}), \\ &= \sum_{j=1}^{q}(|n_T-n_R|+2j-1)(\mu_j^*-2\delta), \\ &= d_L(r) - 2(|n_T-n_R|+q)q\delta, \\ &< d_L(r), \end{aligned} \tag{91}$$

where the above follows from the monotonicity of the rate function

$$I(\boldsymbol{\mu}) = \sum_{j=1}^{n_T}(|n_T-n_R|+2j-1)\mu_i + \frac{n_R n_T t}{2}\mu_{n_T},$$

evaluated at

$$\{\mu_1^*-2\delta\cdots,\mu_q^*-2\delta,0,\cdots,0\} = \arg\inf_{\boldsymbol{\mu}\in\Omega_1} I(\boldsymbol{\mu}),$$

and[16] also follows from the fact that, by definition, $I(\boldsymbol{\mu}^*) = d_L(r)$.

Consequently from (90) we have that

$$-\lim_{\rho\to\infty}\frac{\mathrm{P}\left(N_{SD} \dot{\geq} \rho^{\overline{c}_{r-ld}(r)-3qT\delta}\right)}{\log\rho} < d_L(r), \tag{92}$$

and directly from the definition of the complexity exponent, we have that $c_{r-ld}(r) \geq \overline{c}_{r-ld}(r) - 3qT\delta$. As the bound holds for arbitrarily small $\delta > 0$, it follows that $c_{r-ld}(r) = \overline{c}_{r-ld}(r)$. Directly from [1, Theorem 4] which analyzes the ML-based complexity exponent $c_{ml}(r)$, together with the fact that the ML-based sphere decoder, with or without MMSE preprocessing, shares the same upper bound $\overline{c}_{r-ld}(r)$ as the MMSE-preprocessed lattice decoder, gives that $c_{ml}(r) = \overline{c}_{r-ld}(r)$, which in turns implies that

$$c_{r-ld}(r) = c_{ml}(r).$$

This establishes Theorem 1 and Corollary 1a. □

[15] In light of the fact that event $\mathbf{V}_p^{'}$ has zero measure, what the continuity of eigenvalues guarantees is that we can construct a neighborhood of matrices around $\mathbf{V}_p^{'}$ which are full rank, and which have a non zero measure. We also note that the matrices $\mathbf{V}_p^{'}$ can be created recursively, starting from a single matrix $\mathbf{V}_{n_T}^{'}$.

[16] Recall that parameter t was previously introduced as a parameter that regulates the near zero behavior of the random variable.

APPENDIX B
PROOF FOR COROLLARIES 1B AND 1C

Section A-A shows that $\overline{c}_{r-ld}(r)$ can be obtained as the solution to the constrained maximization problem

$$\overline{c}_{r-ld}(r) \triangleq \max_{\boldsymbol{\mu}} T\sum_{j=1}^{n_T}\left(\frac{r}{n_T}-(1-\mu_j)^+\right)^+$$
$$\text{s.t.} \quad I(\boldsymbol{\mu}) \leq d_L(r), \tag{93a}$$
$$\mu_1 \geq \cdots \geq \mu_{n_T} \geq 0. \tag{93b}$$

In some cases though, further knowledge of the error performance of the encoder and decoder, can result in an explicit characterization of the complexity exponent. Take for instance the case of DMT optimal encoding [15], [16] and DMT optimal MMSE-preprocessed lattice decoding [6], [8], where the constraint $I(\boldsymbol{\mu}) \leq d_L(r)$ in (93a) reverts to the constraint $\sum_{j=1}^{n_T}(1-\mu_j)^+ \geq r$ (cf. [8]), which may be recognized to correspond to the no-outage region (cf. [3]). In this case $\overline{c}_{r-ld}(r)$ can then be explicitly obtained from the optimization problem

$$\overline{c}_{r-ld}(r) = \max_{\boldsymbol{\mu}} T\sum_{j=1}^{n_T}\left(\frac{r}{n_T}-(1-\mu_j)^+\right)^+ \tag{94a}$$
$$\text{s.t.} \ \sum_{j=1}^{n_T}(1-\mu_j)^+ \geq r \tag{94b}$$
$$\mu_1 \geq .... \geq \mu_{n_T} \geq 0, \tag{94c}$$

which can be solved in a straightforward manner to give that

$$\overline{c}_{r-ld}(r) = \frac{T}{n_T}\left(r(n_T-\lfloor r\rfloor-1)+(n_T\lfloor r\rfloor - r(n_T-1))^+\right),$$

describing the upper bound on the complexity exponent for MMSE-preprocessed lattice sphere decoding of DMT optimal full-rate codes, which for minimum delay ($n_T = T$) DMT optimal full-rate codes takes the form

$$\overline{c}_{r-ld}(r) = r(n_T-\lfloor r\rfloor-1)+(n_T\lfloor r\rfloor - r(n_T-1))^+, \tag{95}$$

and which further simplifies to

$$\overline{c}_{r-ld}(r) = r(n_T-r),$$

for integer multiplexing gains $r = 0, 1, \cdots, n_T$. In conjunction with the lower bound in Section A-B, under the conditions layered codes in Corollary 1b, we have that $c_{r-ld}(r) = \overline{c}_{r-ld}(r)$, which proves Corollary 1b. □

Moving on to the universal upper bound, we can see from (70) that, regardless of the fading statistics and the corresponding $I(\boldsymbol{\mu})$, the exponent $\overline{c}_{r-ld}(r)$ is non-decreasing in $d_L(r)$ and is hence maximized when $d_L(r)$ is itself maximized, i.e., it is maximized in the presence of DMT optimal encoding and decoding. Combined with the fact that the corresponding maximization problem in (94) does not depend on the fading distribution, other than the natural fact that its tail must vanish exponentially fast, results in the fact that, for any full-rate code and statistical characterization of the channel, the complexity of MMSE-preprocessed lattice SD is universally upper bounded as (cf. [1])

$$\frac{T}{n_T}\left(r(n_T-\lfloor r\rfloor-1)+(n_T\lfloor r\rfloor - r(n_T-1))^+\right). \tag{96}$$

This proves Corollary 1c. □

APPENDIX C
PROOF FOR LEMMA 1

For $\mathbf{R}_r^H\mathbf{R}_r = \mathbf{M}_r^H\mathbf{M}_r + \alpha_r^2\mathbf{I}$ (cf. (14))[17], it follows by the bounded orthogonality defect of LLL reduced bases that there is a constant $K_\kappa > 0$ independent of $\mathbf{R}_r$ and ρ, for which (cf. [22] and the proof in [30])

$$\sigma_{max}(\tilde{\mathbf{R}}_r^{-1}) \leq \frac{K_\kappa}{\lambda(\mathbf{R}_r)} \tag{97}$$

where

$$\lambda(\mathbf{R}_r) \triangleq \min_{\mathbf{c}\in\mathbb{Z}^\kappa\setminus\mathbf{0}} \|\mathbf{R}_r\mathbf{c}\| \tag{98}$$

denotes the shortest vector in the lattice generated by $\mathbf{R}_r$. As a result we have that

$$\sigma_{min}(\tilde{\mathbf{R}}_r) \geq \frac{\lambda(\mathbf{R}_r)}{K_\kappa}. \tag{99}$$

Looking to lower bound $\sigma_{min}(\tilde{\mathbf{R}}_r)$, we seek a bound on $\lambda(\mathbf{R}_r)$. Towards this let $r' = r-\gamma$ for some $r \geq \gamma > 0$, in which case for $\mathbf{s}$ being the transmitted symbol vector, and for any $\hat{\mathbf{s}} \in \mathbb{Z}^\kappa$ such that $\hat{\mathbf{s}} \neq \mathbf{s}$, it follows that

$$\begin{aligned}\|\mathbf{r}-\mathbf{R}_{r'}\hat{\mathbf{s}}\| &= \|(\mathbf{r}-\mathbf{R}_{r'}\mathbf{s})+\mathbf{R}_{r'}(\mathbf{s}-\hat{\mathbf{s}})\| \\ &\leq \|(\mathbf{r}-\mathbf{R}_{r'}\mathbf{s})\|+\|\mathbf{R}_{r'}(\mathbf{s}-\hat{\mathbf{s}})\|\end{aligned} \tag{100}$$

and

$$\begin{aligned}\|\mathbf{R}_{r'}(\mathbf{s}-\hat{\mathbf{s}})\| &\geq \|\mathbf{r}-\mathbf{R}_{r'}\hat{\mathbf{s}}\|-\|(\mathbf{r}-\mathbf{R}_{r'}\mathbf{s})\| \\ &= \|\mathbf{r}-\mathbf{R}_{r'}\hat{\mathbf{s}}\|-\|\mathbf{w}\|.\end{aligned} \tag{101}$$

From (101) it is clear that to find a lower bound on $\lambda(\mathbf{R}_{r'})$, we need to lower bound $\|\mathbf{r}-\mathbf{R}_{r'}\hat{\mathbf{s}}\|$ for all $\hat{\mathbf{s}} \in \mathbb{Z}^\kappa$ and upper bound $\|\mathbf{w}\|$. Let us, for now, assume that $\|\mathbf{w}\|^2 \leq \rho^b$. To lower bound $\|\mathbf{r}-\mathbf{R}_{r'}\hat{\mathbf{s}}\|$, we draw from the equivalence of MMSE preprocessing and the regularized metric (cf. equation (45) in [8]), and rewrite

$$\|\mathbf{r}-\mathbf{R}_{r'}\hat{\mathbf{s}}\|^2 = \|\mathbf{y}-\mathbf{M}_{r'}\hat{\mathbf{s}}\|^2 + \alpha_{r'}^2\|\hat{\mathbf{s}}\|^2 - c, \tag{102}$$

where $c \triangleq \mathbf{y}^H[\mathbf{I}-\mathbf{M}_{r'}^H(\mathbf{M}_{r'}^H\mathbf{M}_{r'}+\alpha_{r'}^2\mathbf{I})^{-1}\mathbf{M}_{r'}]\mathbf{y} \geq 0$. We now note that for $\hat{\mathbf{s}} = \mathbf{s}$ then $\|\mathbf{y}-\mathbf{M}_{r'}\mathbf{s}\|^2 + \alpha_{r'}^2\|\mathbf{s}\|^2 \doteq \rho^b$, and since the left hand side of (102) cannot be negative, and furthermore given that c is independent of $\hat{\mathbf{s}}$, we conclude that $c \doteq \rho^b$.

We will now proceed to lower bound $\|\mathbf{y}-\mathbf{M}_{r'}\hat{\mathbf{s}}\|^2 + \alpha_{r'}^2\|\hat{\mathbf{s}}\|^2$ and then use (102) to lower bound $\|\mathbf{r}-\mathbf{R}_{r'}\hat{\mathbf{s}}\|$. Towards lower bounding $\|\mathbf{y}-\mathbf{M}_{r'}\hat{\mathbf{s}}\|^2 + \alpha_{r'}^2\|\hat{\mathbf{s}}\|^2$ we draw from Theorem 1 in [8] and we let $\mathcal{B}$ be the spherical region given by

$$\mathcal{B} \triangleq \{d \in \mathbb{R}^\kappa | \, \|\mathbf{d}\|^2 \leq \Gamma^2\}$$

[17] Note the transition to the notation reflecting the dependence of $\mathbf{R}$ on r.

where the radius $\Gamma > 0$ is independent of ρ and is chosen so that $\mathbf{d}_1 + \mathbf{d}_2 \in \mathcal{R}$ for any $\mathbf{d}_1, \mathbf{d}_2 \in \mathcal{B}$. The existence of the set $\mathcal{B}$ follows by the assumption that $\mathbf{0}$ is contained in the interior of $\mathcal{R}$. Now let

$$\nu_{r'} \triangleq \min_{\mathbf{d} \in \rho^{\frac{r'T}{\kappa}} \mathcal{B} \cap \mathbb{Z}^\kappa : \mathbf{d} \neq \mathbf{0}} \frac{1}{4} \left\| \mathbf{M}_{r'} \mathbf{d} \right\|^2,$$

and for given $\gamma > \zeta > 0$ choose $b > 0$ such that

$$\frac{2\zeta T}{\kappa} > b > 0.$$

This may clearly be done for arbitrary $\zeta > 0$. We will in the following temporarily assume that $\nu_{r'+\zeta} \geq 1$ and prove that, together with $\left\| \mathbf{w} \right\|^2 \leq \rho^b$, the two conditions are sufficient for $\lambda(\tilde{\mathbf{R}}_{r'}) \dot{\geq} \rho^{\frac{\zeta T}{\kappa}}$ to hold.

In order to bound the metric for $\hat{\mathbf{s}} \in \mathbb{Z}^\kappa$ where $\hat{\mathbf{s}} \neq \mathbf{s}$, we note that $\nu_{r'+\zeta} \geq 1$ implies that $\forall \mathbf{d} \in \rho^{\frac{(r'+\zeta)T}{\kappa}} \mathcal{B} \cap \mathbb{Z}^\kappa, \mathbf{d} \neq \mathbf{0}$ it is the case that

$$\begin{aligned} \frac{1}{4} \left\| \mathbf{M}_{r'+\zeta} \mathbf{d} \right\|^2 &\geq 1 \\ \frac{1}{4} \left\| \rho^{\frac{1}{2} - \frac{(r'+\zeta)T}{\kappa}} \mathbf{H} \mathbf{G} \mathbf{d} \right\|^2 &\overset{(a)}{\geq} 1 \\ \frac{1}{4} \left\| \rho^{\frac{1}{2} - \frac{r'T}{\kappa}} \mathbf{H} \mathbf{G} \mathbf{d} \right\|^2 &\geq \rho^{\frac{2\zeta T}{\kappa}} \end{aligned}$$

where (a) follows from the fact that $\mathbf{M}_r = \rho^{\frac{1}{2} - \frac{rT}{\kappa}} \mathbf{H} \mathbf{G}$. Consequently

$$\frac{1}{4} \left\| \mathbf{M}_{r'} \mathbf{d} \right\|^2 \geq \rho^{\frac{2\zeta T}{\kappa}}, \ \forall \mathbf{d} \in \rho^{\frac{(r'+\zeta)T}{\kappa}} \mathcal{B} \cap \mathbb{Z}^\kappa, \mathbf{d} \neq \mathbf{0}. \quad (103)$$

As $\mathcal{R}$ is bounded, and as $\zeta > 0$, it holds that $\mathcal{R} \subset \frac{1}{2} \rho^{\frac{\zeta T}{\kappa}} \mathcal{B}$ for all $\rho \geq \rho_1$, for a sufficiently large ρ_1. This implies that $\mathbf{s} \in \frac{1}{2} \rho^{\frac{(r'+\zeta)T}{\kappa}} \mathcal{B}$ for $\rho \geq \rho_1$ since $\mathbf{s} \in \rho^{\frac{r'T}{\kappa}} \mathcal{R}$.

For $\mathbf{s}, \mathbf{d} \in \frac{1}{2} \rho^{\frac{(r'+\zeta)T}{\kappa}} \mathcal{B} \cap \mathbb{Z}^\kappa$, there exists an $\hat{\mathbf{s}} \in \rho^{\frac{(r'+\zeta)T}{\kappa}} \mathcal{B} \cap \mathbb{Z}^\kappa$, $\hat{\mathbf{s}} \neq \mathbf{s}$, such that $\hat{\mathbf{s}} = \mathbf{d} + \mathbf{s}$. Hence for any $\hat{\mathbf{s}} \in \rho^{\frac{(r'+\zeta)T}{\kappa}} \mathcal{B} \cap \mathbb{Z}^\kappa$, we have from (103) that

$$\frac{1}{4} \left\| \mathbf{M}_{r'} (\hat{\mathbf{s}} - \mathbf{s}) \right\|^2 = \frac{1}{4} \left\| \mathbf{M}_{r'} \mathbf{d} \right\|^2 \geq \rho^{\frac{2\zeta T}{\kappa}}. \quad (104)$$

As $\left\| \mathbf{w} \right\|^2 \leq \rho^b$, it follows that $\frac{1}{4} \left\| \mathbf{M}_{r'} \mathbf{d} \right\|^2 \geq \left\| \mathbf{w} \right\|^2$ for large ρ, and that

$$\left\| \mathbf{y} - \mathbf{M}_{r'} \hat{\mathbf{s}} \right\|^2 = \left\| \mathbf{M}_{r'} (\mathbf{s} - \hat{\mathbf{s}}) + \mathbf{w} \right\|^2 \dot{\geq} \rho^{\frac{2\zeta T}{\kappa}}. \quad (105)$$

Consequently

$$\left\| \mathbf{y} - \mathbf{M}_{r'} \hat{\mathbf{s}} \right\|^2 + \alpha_{r'}^2 \left\| \hat{\mathbf{s}} \right\|^2 \dot{\geq} \rho^{\frac{2\zeta T}{\kappa}}. \quad (106)$$

On the other hand if $\hat{\mathbf{s}} \notin \rho^{\frac{(r'+\zeta)T}{\kappa}} \mathcal{B}$, then by definition of $\mathcal{B}$ we have that $\alpha_{r'}^2 \left\| \hat{\mathbf{s}} \right\|^2 \geq \frac{1}{4} \Gamma^2 \rho^{\frac{2\zeta T}{\kappa}}$, and consequently that

$$\left\| \mathbf{y} - \mathbf{M}_{r'} \hat{\mathbf{s}} \right\|^2 + \alpha_{r'}^2 \left\| \hat{\mathbf{s}} \right\|^2 \geq \frac{1}{4} \Gamma^2 \rho^{\frac{2\zeta T}{\kappa}}. \quad (107)$$

From (106) and (107) we then conclude that

$$\left\| \mathbf{y} - \mathbf{M}_{r'} \hat{\mathbf{s}} \right\|^2 + \alpha_{r'}^2 \left\| \hat{\mathbf{s}} \right\|^2 \dot{\geq} \rho^{\frac{2\zeta T}{\kappa}}. \quad (108)$$

Given (106) and (108), for any $\hat{\mathbf{s}} \in \mathbb{Z}^\kappa$ such that $\hat{\mathbf{s}} \neq \mathbf{s}$, it is the case that $\left\| \mathbf{y} - \mathbf{M}_{r'} \hat{\mathbf{s}} \right\|^2 + \alpha_{r'}^2 \left\| \hat{\mathbf{s}} \right\|^2 \dot{\geq} \rho^{\frac{2\zeta T}{\kappa}}$, which combined with $c \dot{\leq} \rho^b$ allows for (102) to give that

$$\left\| \mathbf{r} - \mathbf{R}_{r'} \hat{\mathbf{s}} \right\|^2 \dot{\geq} \rho^{\frac{2\zeta T}{\kappa}}. \quad (109)$$

Applying (98) and (101), we have

$$\begin{aligned} \lambda(\mathbf{R}_{r'}) &\geq \left\| \mathbf{r} - \mathbf{R}_{r'} \hat{\mathbf{s}} \right\| - \left\| \mathbf{w} \right\| \\ &\dot{\geq} \rho^{\frac{\zeta T}{\kappa}} - \rho^{\frac{b}{2}} \\ &\doteq \rho^{\frac{\zeta T}{\kappa}} \end{aligned} \quad (110)$$

where the exponential inequality follows from (109). Furthermore we know that

$$\lambda(\mathbf{R}_r) = \rho^{\frac{-\gamma T}{\kappa}} \lambda(\mathbf{R}_{r'}) \dot{\geq} \rho^{\frac{-\epsilon T}{\kappa}} \quad (111)$$

where $\epsilon = \gamma - \zeta$, $r \geq \epsilon > 0$, and from (99) and (111) it follows that $\sigma_{min}(\tilde{\mathbf{R}}_r) \dot{\geq} \rho^{\frac{-\epsilon T}{\kappa}}$.

We now note that the above implies that for $\nu_{r'+\zeta} \geq 1$ and $\left\| \mathbf{w} \right\|^2 \leq \rho^b$ then $\sigma_{min}(\tilde{\mathbf{R}}_r) \dot{\geq} \rho^{\frac{-\epsilon T}{\kappa}}$, and thus applying the union bound yields

$$\begin{aligned} \mathrm{P}\left(\sigma_{min}(\tilde{\mathbf{R}}_r) \dot{<} \rho^{\frac{-\epsilon T}{\kappa}} \right) &= \mathrm{P}\left((\nu_{r'+\zeta} < 1) \cup (\left\| \mathbf{w} \right\|^2 > \rho^b) \right) \\ &\leq \mathrm{P}\left(\nu_{r'+\zeta} < 1 \right) + \mathrm{P}\left(\left\| \mathbf{w} \right\|^2 > \rho^b \right). \end{aligned}$$

We know from the exponential tail of the Gaussian distribution that $\mathrm{P}\left(\left\| \mathbf{w} \right\|^2 > \rho^b \right) \doteq \rho^{-\infty}$ and from Lemma 1 in [8] that $\mathrm{P}\left(\nu_{r'+\zeta} < 1 \right) \dot{\leq} \rho^{-d_{ML}(r'+\zeta)}$. Hence

$$\mathrm{P}\left(\sigma_{min}(\tilde{\mathbf{R}}_r) \dot{<} \rho^{\frac{-\epsilon T}{\kappa}} \right) \dot{\leq} \rho^{-d_{ML}(r-\epsilon)}$$

for all $r \geq \epsilon > 0$.

The association with the singular values

$$\sigma_1(\tilde{\mathbf{R}}_{r,k}) \leq \cdots \leq \sigma_k(\tilde{\mathbf{R}}_{r,k})$$

is made using the interlacing property of singular values of sub-matrices, which gives that

$$\sigma_i(\tilde{\mathbf{R}}_{r,k}) \geq \sigma_i(\tilde{\mathbf{R}}_r), \ i \leq k = 1, \cdots, \kappa, \quad (112)$$

and for $k = 1, \cdots, \kappa$, that

$$\mathrm{P}\left(\sigma_{min}(\tilde{\mathbf{R}}_{r,k}) \dot{<} \rho^{\frac{-\epsilon T}{\kappa}} \right) \dot{\leq} \rho^{-d_{ML}(r-\epsilon)}.$$

Finally from the DMT optimality of the exact implementation of the regularized lattice decoder [6], [8], we have that

$$\mathrm{P}\left(\sigma_{min}(\tilde{\mathbf{R}}_{r,k}) \dot{<} \rho^{\frac{-\epsilon T}{\kappa}} \right) \dot{\leq} \rho^{-d_L(r-\epsilon)}.$$

This proves Lemma 1.□

Appendix D
Proof for Lemma 2

For a search radius that grows as $\xi = \sqrt{z \log \rho} \doteq \rho^0$, we first prove that

$$\mathrm{P}\left(\left\| \mathbf{w}'' \right\|^2 > \xi^2 \right) \dot{\leq} \rho^{-z'}$$

for $z > z' > d_L(r)$. Towards establishing the properties of the equivalent noise $\mathbf{w}''$ (cf. (45)), we consider an equivalent representation of the MMSE-preprocessed lattice decoder and let (cf. [31])

$$\mathbf{Q}\mathbf{R} = \begin{bmatrix} \mathbf{Q}_1 \\ \mathbf{Q}_2 \end{bmatrix} \mathbf{R} = \begin{bmatrix} \mathbf{M} \\ \alpha_r \mathbf{I} \end{bmatrix} \in \mathbb{R}^{(n+\kappa) \times \kappa} \quad (113)$$

be the thin QR factorization of the modified channel matrix, where $\mathbf{Q}_1 = \mathbf{R}^{-1} \mathbf{M} \in \mathbb{R}^{n \times \kappa}$, $\mathbf{Q}_2 = \alpha_r \mathbf{R}^{-1} \in \mathbb{R}^{\kappa \times \kappa}$ and

where $\mathbf{R}^H\mathbf{R} = \mathbf{M}^H\mathbf{M} + \alpha_r^2\mathbf{I}$. It then follows that for $\mathbf{F} = \mathbf{Q}_1^H$, the MMSE-preprocessed lattice decoder is equivalent to lattice decoding in the presence of channel $\mathbf{R}$ and noise

$$\begin{aligned}\mathbf{w}^{'} &= -\alpha_r^2\mathbf{R}^{-H}\mathbf{s} + \mathbf{R}^{-H}\mathbf{M}^H\mathbf{w} \\ &= -\alpha_r\mathbf{Q}_2^H\mathbf{s} + \mathbf{Q}_1^H\mathbf{w}. \end{aligned} \quad (114)$$

Consequently we calculate

$$\begin{aligned}& \mathrm{P}\left(\|\mathbf{w}^{'}\| > \xi\right) \\ \leq\ & \mathrm{P}\left(\|-\alpha_r\mathbf{Q}_2^H\mathbf{s}\| + \|\mathbf{Q}_1^H\mathbf{w}\| > \xi\right) \\ \overset{(a)}{=}\ & \mathrm{P}\left(\|-\alpha_r\mathbf{Q}^H\begin{bmatrix}\mathbf{0}\\ \mathbf{s}\end{bmatrix}\| + \|\mathbf{Q}^H\begin{bmatrix}\mathbf{w}\\ \mathbf{0}\end{bmatrix}\| > \xi\right) \\ \leq\ & \mathrm{P}\left(\kappa\big(\|\mathbf{w}\| + \sup_{\mathbf{s}\in\mathbb{S}_r^\kappa}\|-\alpha_r\mathbf{s}\|\big) > \xi\right) \\ \overset{(b)}{=}\ & \mathrm{P}\left(\kappa\|\mathbf{w}\| + \kappa K > \xi\right) \\ =\ & \mathrm{P}\left(\kappa\|\mathbf{w}\| > (z\log\rho)^{\frac{1}{2}} - \kappa K\right) \\ \overset{(c)}{\leq}\ & \mathrm{P}\left(\kappa\|\mathbf{w}\| > (z_1\log\rho)^{\frac{1}{2}}\right) \\ =\ & \mathrm{P}\left(\|\mathbf{w}\|^2 > \frac{z_1}{\kappa^2}\log\rho\right) \\ \overset{(d)}{=}\ & \mathrm{P}\left(\|\mathbf{w}\|^2 > z_2\log\rho\right) \\ \doteq\ & \rho^{-z_2} \end{aligned} \quad (115)$$

where (a) follows from the MMSE-preprocessed equivalent channel representation (cf. (113)), and where the inequalities in (b), (c) and (d) follow for some fixed K that upper bounds $\sup_{\mathbf{s}\in\mathbb{S}_r^\kappa}\|-\alpha_r\mathbf{s}\|$, and for some arbitrary z_1, z_2 satisfying $z > z_1 > z_2 > 0$ independent of ρ. Consequently

$$\mathrm{P}\left(\|\mathbf{w}^{''}\| > \xi\right) = \mathrm{P}\left(\|\tilde{\mathbf{Q}}^H\mathbf{w}^{'}\| > \xi\right) \dot{\leq} \rho^{-z'}$$

for some $0 < z' < z_2$, and as a result

$$\lim_{\rho\to\infty}\frac{\mathrm{P}\left(\|\mathbf{w}^{''}\| > \xi\right)}{\mathrm{P}\left(\hat{\mathbf{s}}_{r-ld} \neq \mathbf{s}\right)} = \lim_{\rho\to\infty}\rho^{(d_L(r)-z')} = 0,$$

where the last equality follows after choosing the search radius such that $z > z' > d_L(r)$. This proves Lemma 2. □

REFERENCES

[1] J. Jaldén and P. Elia, "Sphere decoding complexity exponent for decoding full rate codes over the quasi-static mimo channel," Feb. 2011, submitted to *IEEE Trans. Inform. Theory*, available on arXiv:1102.1265 [cs.IT].

[2] D. Micciancio, "The hardness of the closest vector problem with preprocessing," *IEEE Trans. Inf. Theory*, vol. 47, no. 3, pp. 1212 –1215, Mar. 2001.

[3] 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.

[4] E. Agrell, T. Eriksson, A. Vardy, and K. Zeger, "Closest point search in lattices," *IEEE Trans. Inf. Theory*, vol. 48, no. 8, pp. 2201–2214, Aug. 2002.

[5] A. D. Murugan, H. E. Gamal, M. O. Damen, and G. Caire, "A unified framework for tree search decoding: rediscovering the sequential decoder," *IEEE Trans. Inf. Theory*, vol. 52, no. 3, pp. 933– 953, Mar. 2006.

[6] H. El Gamal, G. Caire, and M. O. Damen, "Lattice coding and decoding achieve the optimal diversity-multiplexing tradeoff of MIMO channels," *IEEE Trans. Inf. Theory*, vol. 50, no. 6, pp. 968–985, Jun. 2004.

[7] J. Jaldén and P. Elia, "The complexity of sphere decoding perfect codes under a vanishing gap to ml performance," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, Jul. 2011.

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

[9] J. Pan and W.-K. Ma, "A lagrangian dual relaxation approach to ML MIMO detection: reinterpreting regularized lattice decoding," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011.

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

[11] D. Seethaler, J. Jaldén, C. Studer, and H. Bölcskei, "Tail behavior of sphere-decoding complexity in random lattices," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, Jun. 2009, pp. 729 – 733.

[12] J. Radon, *Lineare Scharen Orthogonaler Matrizen*. Abh. Math. Sem. Hamburg I, 1923.

[13] H. El Gamal and M. O. Damen, "Universal space-time coding," *IEEE Trans. Inf. Theory*, vol. 49, no. 5, pp. 1097–1119, May 2003.

[14] B. A. Sethuraman, B. Sundar Rajan, and V. Shashidhar, "Full-diversity, high-rate, space-time block codes from division algebras," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2596–2616, Oct. 2003.

[15] P. Elia, K. R. Kumar, S. A. Pawar, P. Vijay Kumar, and H.-F. Lu, "Explicit space-time codes achieving the diversity-multiplexing gain tradeoff," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 3869–3884, Sep. 2006.

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

[17] L. Zhao, W. Mo, Y. Ma, and Z. Wang, "Diversity and multiplexing tradeoff in general fading channels," *IEEE Trans. Inf. Theory*, vol. 53, no. 4, pp. 1547–1557, Apr. 2007.

[18] H. Yao and G. W. Wornell, "Lattice-reduction-aided detectors for MIMO communication systems," in *Proc. IEEE Global Conf. Communications (GLOBECOM)*, Taipei, Taiwan, Nov. 2002.

[19] C. Windpassinger and R. F. H. Fischer, "Low-complexity near-maximum-likelihood detection and precoding for MIMO systems using lattice reduction," in *Proc. IEEE Information Theory Workshop (ITW)*, Paris, France, Mar. 2003.

[20] D. Wübben, D. Seethaler, J. Jaldén, and G. Matz, "Lattice reduction: A survey with with applications to wireless communication," *IEEE Signal Processing Magazine*, vol. 28, no. 3, pp. 70 – 91, Jun. 2011.

[21] C. Ling, "On the proximity factors of lattice reduction-aided decoding," *IEEE Trans. Signal Process.*, vol. 59, no. 6, pp. 2795 – 2808, 2011.

[22] A. K. Lenstra, H. W. Lenstra, and L. Lovász, "Factoring polynomials with rational coefficients," *Matematische Annalen*, vol. 261, no. 4, pp. 1432–1807, Dec. 1982.

[23] J. Jaldén, D. Seethaler, and G. Matz, "Worst- and average-case complexity of LLL lattice reduction in MIMO wireless systems," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, Nevada, USA, Apr. 2008.

[24] L. Babai, "On Lovász' lattice reduction and the nearest lattice point problem," *Combinatorica*, vol. 6, no. 1, pp. 1–13, 1986.

[25] S. Loyka and G. Levin, "Finite-snr diversity-multiplexing tradeoff via asymptotic analysis of large mimo systems," *IEEE Trans. Inf. Theory*, vol. 56, no. 10, pp. 4781 – 4792, Oct. 2010.

[26] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 1985.

[27] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed. Springer-Verlag New York Inc., 1998.

[28] V. Raghavan, A. Sayeed, and N. Boston, "When is limited feedback for transmit beamforming beneficial?" Sep. 2005, pp. 1544 –1548.

[29] A. Tulino and S. Verdú, "Random matrix theory and wireless communications," *Foundations and Trends in Communications and Information Theory*, vol. 1, June 2004.

[30] M. Taherzadeh, A. Mobasher, and A. K. Khandani, "LLL reduction achieves the receive diversity in MIMO decoding," *IEEE Trans. Inform. Theory*, vol. 53, no. 12, pp. 4801–4805, Dec. 2007.

[31] D. Wübben, R. Bohnke, V. Kuhn, and K.-D. Kammeyer, "Near-maximum-likelihood detection of MIMO systems using MMSE-based lattice reduction," in *Proc. IEEE Int. Conf. Communications (ICC)*, Paris, France, Jun. 2004.