Rate-Reliability-Complexity tradeoff for ML and Lattice decoding of full-rate codes

Arun Singh and Petros Elia
Mobile Communications Department
EURECOM, Sophia Antipolis, France
Email: {singhak, elia}@eurecom.fr

Joakim Jaldén
ACCESS Linnaeus Center, Signal Processing Lab
School of Electrical Engineering, KTH, Stockholm, Sweden
Email: jalden@kth.se

***Abstract*—Recent work in [1]–[3] quantified, in the form of a complexity exponent, the computational resources required for ML and lattice sphere decoding to achieve a certain diversity-multiplexing performance. For a specific family of layered lattice designs, and a specific set of decoding orderings, this complexity was shown to be an exponential function in the number of codeword bits, and was shown to meet a universal upper bound on complexity exponents. The same results raised the question of whether complexity reductions away from the universal upper bound are feasible, for example, with a proper choice of decoder (ML vs lattice), or with a proper choice of lattice codes and decoding ordering policies.**

The current work addresses this question by first showing that for almost any full-rate DMT optimal lattice code, there exists no decoding ordering policy that can reduce the complexity exponent of ML or lattice based sphere decoding away from the universal upper bound, i.e., that a randomly picked lattice code (randomly and uniformly drawn from an ensemble of DMT optimal lattice designs) will almost surely be such that no decoding ordering policy can provide exponential complexity reductions away from the universal upper bound. As a byproduct of this, the current work proves the fact that ML and (MMSE-preprocessed) lattice decoding share the same complexity exponent for a very broad setting, which now includes almost any DMT optimal code (again randomly drawn) and all decoding order policies. Under a basic *richness of codes* assumption, this is in fact further extended to hold, with probability one, over all full-rate codes. Under the same assumption, the result allows for a meaningful rate-reliability-complexity tradeoff that holds, almost surely in the random choice of the full-rate lattice design, and which holds irrespective of the decoding ordering policy. This tradeoff can be used to, for example, describe the optimal achievable diversity gain of ML or lattice sphere decoding in the presence of limited computational resources.

I. Introduction

In MIMO systems, the prohibitively large computational costs of maximum likelihood (ML) decoding serve as motivation to consider different branch-and-bound algorithms [4]–[6] which can provide computational savings at the expense of a relatively small error-performance degradation. In the setting of quasi-static MIMO channels, the computational complexity of such decoding algorithms such as the sphere decoder (SD), has random fluctuations which are induced by the randomness in the fading, the noise and the transmitted codeword. In the presence of computational constraints, this fluctuating computational requirements cause error performance degradation due to additional 'outage' events where the instantaneous computational requirements exceed the computational constraints. These constraints, henceforth denoted as $N_{\max}$, describe the amount of computational reserves, in floating point operations (flops) per duration of one codeword, 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. Consequently any attempt to significantly reduce the computational constraints may be at the expense of a substantial degradation in error-performance. This brings to the fore a certain rate-reliability-complexity tradeoff, which we explore.

Recently the work in [1] provided rigorous and simple answers to the broad question of how large computational reserves are required by ML sphere decoders to achieve, for specific codes and decoding ordering policies, a certain diversity-multiplexing performance (DMT, [7]). As in [1], [2], these reserves are quantified in the high signal-to-noise ratio regime, to take the form of a *complexity exponent*

$$c = \lim_{\text{SNR}\to\infty} \frac{\log N_{\max}}{\log \text{SNR}}. \tag{1}$$

The same work in [1] also provided a universal upper bound on the computational resources required to achieve, for increasing SNR, a vanishing gap to the brute force ML error performance. This analysis showed that specific families of lattice codes, and specific decoding ordering policies, in fact meet this universal complexity upper bound. The same conclusions were drawn later on for MMSE-preprocessed lattice decoding, with the work in [2] proving the somewhat surprising fact that, for the above mentioned specific codes and decoding ordering policies, the ML and lattice sphere decoders meet the same universal upper bound, and thus share the same complexity exponent.

The natural question is of course whether there exist codes and decoding ordering policies that do not meet the universal upper bound and thus provide substantial (polynomial in SNR and exponential in the number of codeword bits) reductions in the computational complexity. A step towards answering this was found in [3] which considered a general setting of MIMO scenarios, of fading statistics and lattice codes, to show that there exist code-channel dependent decoding ordering policies for which the complexity meets the universal upper bound.

The current work comes closer to establishing the minimum computational resources required by ML and lattice based SD to achieve the optimal diversity-multiplexing performance (and a vanishing gap to ML), by shedding light on whether all codes and all decoding ordering policies share the same need for resources that match the above mentioned universal upper bound that was shown in [1], [2] to be necessary for specific policies and lattice designs. Towards this our work here proves that this universal upper bound is in fact tight, for both ML and lattice SD, for almost any full-rate DMT optimal lattice code, and then for any decoding ordering policy. To clarify, this means that, if we were to randomly pick a code from the ensemble of DMT optimal lattice designs, then with probability one, this code would be such that there would exist no decoding ordering policy that could result in complexity that is less than that described by the universal upper bound.

As a consequence to this, the work also shows that ML and (MMSE-preprocessed) lattice SD have matching complexities with probability one (in the choice of DMT optimal codes) and irrespective of the decoding ordering policy. Under a basic *richness of codes* assumption defined in Section III, this is in fact further extended to hold, with probability one, over all full-rate codes.

Under the same assumption, the above breakthroughs allow for a more meaningful rate-reliability-complexity tradeoff that holds for both ML and lattice SD, almost surely in the random choice of the full-rate code, and which holds irrespective of the decoding ordering policy. Whereas in the past, such tradeoff could be derived under the strict restrictions on the decoding ordering, now this tradeoff is more meaningful and can provide insights that hold, with probability one in the choice of full-rate code, irrespective of the channel-dependent or channel-independent decoding ordering policy. One could use this tradeoff to, for example, provide insights on the achievable DMT performance of powerful encoders and decoders that are limited to use computational resources corresponding to less computational expensive but also less powerful transceivers.

A. System model

The general $n_T \times n_R$ ($n_T \leq n_R$) point-to-point quasi-static MIMO channel is given by

$$\mathbf{Y}_C = \sqrt{\rho}\mathbf{H}_C\mathbf{X}_C + \mathbf{W}_C, \tag{2}$$

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, where the fading matrix $\mathbf{H}_C \in \mathbb{C}^{n_R\times n_T}$ is assumed to be random, with elements drawn from an i.i.d. Rayleigh fading statistics, and where ρ denotes the signal to noise ratio. After vectorization the real valued representation of (2) takes the form

$$\mathbf{y} = \sqrt{\rho}\mathbf{H}\mathbf{x} + \mathbf{w}, \tag{3}$$

where $\mathbf{H} = \mathbf{I}_T \otimes \mathbf{H}_R$ with $\mathbf{H}_R = \begin{bmatrix} Re\{\mathbf{H}_C\} & -Im\{\mathbf{H}_C\} \\ Im\mathbf{H}_C\} & Re\{\mathbf{H}_C\} \end{bmatrix}$, $\mathbf{x} = (\mathbf{x}_1^{\mathrm{T}}, \cdots, \mathbf{x}_T^{\mathrm{T}})^{\mathrm{T}}$ 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 $\mathbf{x}$. For a rate R that scales with SNR as a function of the multiplexing gain $r = R/\log\rho \geq 0$, we consider a (sequence of) full-rate linear (lattice) code(s) $\mathcal{X}_r$ given by $\mathcal{X}_r = \Lambda_r \cap \mathcal{R}'$ to consist of the elements of the lattice Λ_r that lie inside the *shaping region* $\mathcal{R}'$ which is a compact convex subset of $\mathbb{R}^\kappa$ that is independent of ρ. Specifically $\Lambda_r \triangleq \rho^{\frac{-rT}{\kappa}}\Lambda$ and $\Lambda \triangleq \{\mathbf{G}\mathbf{s} \mid \mathbf{s} \in \mathbb{Z}^\kappa\}$, where $\mathbb{Z}^\kappa$ denotes the κ dimensional integer lattice, and where the *lattice generator matrix* $\mathbf{G} \in \mathbb{R}^{\kappa\times\kappa}$ is full rank and independent of ρ. We clarify that the code is drawn randomly from the ensemble of DMT optimal lattice designs, in the sense that the coefficients in $\mathbf{G}$ are independently and randomly chosen from a continuous distribution over the real numbers, and are then held fixed. We also clarify that in our setting, the use of DMT optimal full-rate codes means that $\kappa = 2n_T T$. After vectorization the codewords take the form

$$\mathbf{x} = \rho^{\frac{-rT}{\kappa}}\mathbf{G}\mathbf{s}, \quad \mathbf{s} \in \mathbb{S}_r^\kappa \triangleq \mathbb{Z}^\kappa \cap \rho^{\frac{rT}{\kappa}}\mathcal{R}, \tag{4}$$

where $\mathcal{R} \subset \mathbb{R}^\kappa$ is a natural bijection of the shaping region $\mathcal{R}'$ that preserves the code, and contains the all zero vector $\mathbf{0}$. For simplicity we consider $\mathcal{R} \triangleq [-1,\ 1]^\kappa$ to be a hypercube in $\mathbb{R}^\kappa$ although this can be partially relaxed. Combining (3) and (4) yields the equivalent system model

$$\mathbf{y} = \mathbf{M}\mathbf{s} + \mathbf{w}, \tag{5a}$$

$$\text{where} \quad \mathbf{M} \triangleq \rho^{\frac{1}{2}-\frac{rT}{\kappa}}\mathbf{H}\mathbf{G} \in \mathbb{R}^{2n_R T\times\kappa}. \tag{5b}$$

B. Sphere Decoder

Let $\mathbf{QR} = \mathbf{M}$ be the thin QR factorization of the code-channel matrix $\mathbf{M}$ and $\mathbf{r} \triangleq \mathbf{Q}^H\mathbf{y}$, then (5a) yields $\mathbf{r} = \mathbf{R}\mathbf{s} + \mathbf{Q}^H\mathbf{w}$ and the ML decoder for this system takes the form

$$\hat{\mathbf{s}}_{ML} = \arg\min_{\hat{\mathbf{s}}\in\mathbb{S}_r^\kappa} \|\mathbf{r} - \mathbf{R}\hat{\mathbf{s}}\|^2. \tag{6}$$

We use SD to implement the decoder in (6), which recursively enumerates all candidate vectors $\hat{\mathbf{s}} \in \mathbb{S}_r^\kappa$ within a given sphere of radius $\xi > 0$. 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

$$\|\mathbf{r}_k - \mathbf{R}_k\hat{\mathbf{s}}_k\|^2 \leq \xi^2, \tag{7}$$

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 $\mathbf{R}$.

We note that the error performance and the total number of visited nodes is a function of the search radius ξ. As in [1], we use a fixed search radius $\xi = \sqrt{z\log\rho}$ for some $z > d(r)$ such that

$$\mathrm{P}\left(\|\mathbf{Q}^H\mathbf{w}\|^2 > \xi^2\right) \ \dot{<}\ \rho^{-d(r)}, \tag{8}$$

which implies a vanishing probability of excluding the transmitted information vector from the search. We use $\doteq$ to denote the *exponential equality*, i.e., we write $f(\rho) \doteq \rho^B$ to denote $\lim_{\rho\to\infty} \dfrac{\log f(\rho)}{\log\rho} = B$, and $\dot{\leq}, \dot{<}$, and $\dot{\geq}, \dot{>}$ are defined similarly.

a) Decoding order policies:: We note that permuting the columns of generator matrix $\mathbf{G}$ by replacing it with $\mathbf{G}\mathbf{\Pi}$ for some permutation matrix $\mathbf{\Pi} \in \mathbb{R}^{\kappa\times\kappa}$, changes the order in which the symbols in $\mathbf{s}$ are enumerated by the sphere decoder, without changing the codebook. Such permutations have been known to play a role in reducing complexity by reducing the number of nodes visited by the sphere decoding algorithm. Such policies can correspond to a *fixed decoding order* where $\mathbf{\Pi}$ does not change with the channel realization, or to a *dynamic decoding order* where $\mathbf{\Pi}$ can vary with the channel realization. We will here consider the most general case where $\mathbf{\Pi}$ may vary with the channel.

C. Rate-reliability-complexity tradeoff in outage-limited MIMO communications

In the high SNR regime, a given encoder $\mathcal{X}_r$ and decoder $\mathcal{D}_r$ are said to achieve a *multiplexing gain* r and *diversity gain* $d_{\mathcal{D}}(r)$ if (cf. [7])

$$\lim_{\rho\to\infty} \frac{R(\rho)}{\log\rho} = r, \qquad \text{and} \qquad -\lim_{\rho\to\infty} \frac{\log P_e}{\log\rho} = d_{\mathcal{D}}(r) \tag{9}$$

where P_e denotes the probability of codeword error with a ML-based sphere decoder $\mathcal{D}_r$ employing time-out policies.

As noted, for $N_{\max}$ describing the computational reserves in flops, the complexity exponent takes the form

$$c(r) := \lim_{\rho\to\infty} \frac{\log N_{\max}}{\log \rho}. \tag{10}$$

In terms of error-performance gaps, we first consider the gap of $\mathcal{D}_r$ to ML, i.e, the gap between the error performance P_e of $\mathcal{D}_r$ to the optimal error performance $\mathrm{P}\left(\hat{\mathbf{s}}_{ML} \neq \mathbf{s}\right)$ of the brute force ML decoder. Given a certain computational constraint $N_{\max} \doteq \rho^c$ for $\mathcal{D}_r$, this gap is quantified in the high SNR regime to be

$$g(c) \triangleq \lim_{\rho\to\infty} \frac{P_e}{\mathrm{P}\left(\hat{\mathbf{s}}_{ML} \neq \mathbf{s}\right)}. \tag{11}$$

A *vanishing gap* $g(c) = 1$ means that with $N_{\max} \doteq \rho^c$ flops, $\mathcal{D}_r$ can asymptotically have near identical error performance as the optimal ML decoder.

Similarly, when considering the MMSE-preprocessed lattice sphere decoder, we are interested in the performance gap to the *exact implementation* of the MMSE-preprocessed lattice decoder. As before, in the presence of $N_{\max} \doteq \rho^{c_L}$ flops for the lattice sphere decoder, we have a vanishing gap when

$$g(c_L) \triangleq \lim_{\rho\to\infty} \frac{P_L}{\mathrm{P}\left(\hat{\mathbf{s}}_{L} \neq \mathbf{s}\right)} = 1 \tag{12}$$

where P_L describes the error probability of the preprocessed lattice sphere decoder, and where $\mathrm{P}\left(\hat{\mathbf{s}}_{L} \neq \mathbf{s}\right)$ describes the error probability of the exact solution of MMSE-preprocessed lattice decoder.

II. Complexity of ML-based Sphere Decoding

The total number of visited nodes is commonly taken as a measure of the sphere decoder complexity[1] and is given by

$$N_{SD} = \sum_{k=1}^{\kappa} N_k, \tag{13}$$

where N_k denotes the number of visited nodes at layer k that corresponds to the kth component of the transmitted symbol vector $\mathbf{s}$ and is given by $N_k \triangleq |\mathcal{N}_k|$ where $\mathcal{N}_k \triangleq \{\hat{\mathbf{s}}_k \in \mathbb{S}_r^{\kappa} \mid \|\mathbf{r}_k - \mathbf{R}_k\hat{\mathbf{s}}_k\|^2 \leq \xi^2\}$.

We are interested in the ML-based SD complexity required to achieve a vanishing performance gap to brute force ML[2]. We recall that a ML-based SD with run-time constraints, in addition to making the ML errors ($\hat{\mathbf{s}}_{ML} \neq \mathbf{s}$), also makes errors when the run-time limit of ρ^x flops for $x > c(r)$ becomes active, as well as when the fixed search radius ξ causes $\mathcal{N}_\kappa = \emptyset$. Consequently the corresponding performance gap to the brute force ML decoder, takes the form (cf. (11))

$$g(x) = \lim_{\rho\to\infty} \frac{\mathrm{P}\left(\{\hat{\mathbf{s}}_{ML} \neq \mathbf{s}\} \cup \{N_{SD} \geq \rho^x\} \cup \{\mathcal{N}_\kappa = \emptyset\}\right)}{\mathrm{P}\left(\hat{\mathbf{s}}_{ML} \neq \mathbf{s}\right)}.$$

To bound the above gap, we apply the union bound in conjunction with (8) and the fact that $\mathrm{P}\left(\mathcal{N}_\kappa = \emptyset\right) \leq \mathrm{P}\left(\|\mathbf{Q}^H\mathbf{w}\| > \xi\right)$, to get

$$g(x) \leq \lim_{\rho\to\infty} \left(\frac{\mathrm{P}\left(\hat{\mathbf{s}}_{ML} \neq \mathbf{s}\right) + \mathrm{P}\left(N_{SD} \geq \rho^x\right)}{\mathrm{P}\left(\hat{\mathbf{s}}_{ML} \neq \mathbf{s}\right)}\right). \tag{14}$$

Thus a vanishing gap to brute force ML decoding requires that

$$\lim_{\rho\to\infty} \frac{\mathrm{P}\left(N_{SD} \geq \rho^x\right)}{\mathrm{P}\left(\hat{\mathbf{s}}_{ML} \neq \mathbf{s}\right)} = 0.$$

Now going back to (10), and having in mind appropriate timeout policies that guarantee a *vanishing* gap, the complexity exponent $c(r)$ can be bounded as $\underline{c}(r) \leq c(r) \leq \overline{c}(r)$, where

$$\overline{c}(r) \triangleq \inf\{x \mid -\lim_{\rho\to\infty} \frac{\log \mathrm{P}\left(N_{SD} \geq \rho^x\right)}{\log \rho} > d(r)\}, \tag{15a}$$

$$\underline{c}(r) \triangleq \sup\{x \mid -\lim_{\rho\to\infty} \frac{\log \mathrm{P}\left(N_{SD} \geq \rho^x\right)}{\log \rho} < d(r)\}. \tag{15b}$$

We note that $\overline{c}(r)$ and $\underline{c}(r)$ respectively denote sufficient and necessary conditions that guarantee a vanishing gap to ML performance.

We define $\mu_j \triangleq -\frac{\log \sigma_j(\mathbf{H}_C^H\mathbf{H}_C)}{\log \rho}$, $j = 1, \cdots, n_T$, where $\mu_1 \geq \cdots \geq \mu_{n_T}$ and where $\sigma_j(\mathbf{H}_C^H\mathbf{H}_C)$ denotes the j-th singular value of $\mathbf{H}_C^H\mathbf{H}_C$. The upper bound on the complexity exponent can be obtained following the footsteps of the proof for [1, Theorem 2] and is given by $\overline{c}(r) \leq \tilde{c}(r)$ where

$$\tilde{c}(r) \triangleq \max_{\boldsymbol{\mu}} T \sum_{j=1}^{n_T} \min\left(\frac{r}{n_T} - (1-\mu_j), \frac{r}{n_T}\right)^+ \tag{16a}$$

$$\text{s.t.} \quad I(\boldsymbol{\mu}) \leq d(r), \tag{16b}$$

$$\mu_1 \geq \cdots \geq \mu_{n_T} \geq 0, \tag{16c}$$

where $\boldsymbol{\mu}$ satisfies the large deviation principle with rate function $I(\boldsymbol{\mu})$. Equivalently for $\boldsymbol{\mu}^* = (\mu_1^*, \cdots, \mu_{n_T}^*)$ being one of the maximizing vectors such that $I(\boldsymbol{\mu}^*) = d(r)$, we have that $\tilde{c}(r) = T\sum_{j=1}^{n_T} \min\left(\frac{r}{n_T} - (1-\mu_j^*), \frac{r}{n_T}\right)^+$. Furthermore given the monotonicity of the rate function $I(\boldsymbol{\mu})$, and the fact that the objective function in (16) does not increase in μ_j beyond $\mu_j = 1$, we may also assume without loss of generality that $\mu_j^* \leq 1$ for $j = 1, \cdots, n_T$. It follows that

$$\overline{c}(r) \leq \tilde{c}(r) = T \sum_{j=1}^{n_T} \min\left(\frac{r}{n_T} - (1-\mu_j^*)\right)^+. \tag{17}$$

In order to establish a lower bound that matches the upper bound in (17) irrespective of the decoding ordering policy, we define the following lemma.

Lemma 1: Let $\mathbf{G}_p$ be the matrix consisting of any $2pT$ columns of the (fixed but) randomly chosen generator matrix $\mathbf{G}$ where entries of $\mathbf{G}$ are independently chosen from a continuous distribution over the real numbers. Let $\mathbf{V}_p$ be the matrix consisting of the $2p$ columns of the unitary matrix $\mathbf{V}$ corresponding to the $2p$ largest singular values of $\mathbf{H}_R$, where $\mathbf{V}$ is such that $\mathbf{H}_R = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^H$, where $\boldsymbol{\Sigma} \triangleq \mathrm{diag}\{\sigma_1(\mathbf{H}_R), \cdots, \sigma_{2n_T}(\mathbf{H}_R)\}$ and $\mathbf{V}\mathbf{V}^H = \mathbf{I}$. Then almost surely, in the choice of $\mathbf{G}$ we have that

$$rank((I_T \otimes \mathbf{V}_p^H)\mathbf{G}_p) = 2pT. \tag{18}$$

[1]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 total number of visited nodes, we considered the total number of flops spent by the decoder.

[2]In this case only, this same complexity can be shown to match the complexity required by ML-based sphere decoders to achieve an optimal DMT performance.

It then follows that for any fixed or dynamically changing column permutation matrix $\mathbf{\Pi}$, and for $\mathbf{G}_{|p}$ denoting the first $2pT$ columns of the matrix $\mathbf{G\Pi}$, it holds that

$$\mathrm{P}\left(\{\sigma_1\left((\mathbf{I}_T \otimes \mathbf{V}_p^H)\mathbf{G}_{|p}\right) \geq u\}\right) \doteq \rho^0, \quad u > 0. \tag{19}$$

Proof: For $\mathbf{v}_i^H,\ i = 1, \cdots, 2pT$ denoting the $2pT$ linearly independent rows of the matrix $I_T \otimes \mathbf{V}_p^H$ with rank $2pT$ and for $\mathbf{g}_i,\ i = 1, \cdots, \kappa$ denoting the κ linearly independent columns of the full rank matrix $\mathbf{G}$, then

$$(I_T \otimes \mathbf{V}_p^H)\mathbf{G} = \begin{bmatrix} \mathbf{v}_1^H\mathbf{g}_1 & \dots & \mathbf{v}_1^H\mathbf{g}_\kappa \\ \vdots & \ddots & \vdots \\ \mathbf{v}_{2pT}^H\mathbf{g}_1 & \dots & \mathbf{v}_{2pT}^H\mathbf{g}_\kappa \end{bmatrix}.$$

Since $\mathbf{v}_i^H,\ i = 1, \cdots, 2pT$ are fixed and linearly independent, any $2pT$ columns of $(I_T \otimes \mathbf{V}_p^H)\mathbf{G}$ are linearly independent ($rank((I_T \otimes \mathbf{V}_p^H)\mathbf{G}_p) = 2pT$), with probability one. This in turn implies that, given such linear lattice codes that are drawn with probability one, there exists a unitary matrix $\mathbf{V}_p$ such that irrespective of the fixed or dynamically changing column permutation matrix $\mathbf{\Pi}$, it is the case that $\mathrm{rank}\big((\mathbf{I}_T \otimes (\mathbf{V}_p)^H)\mathbf{G}_{|p}\big) = 2pT$. Consequently, it follows that $\sigma_1\left((\mathbf{I}_T \otimes (\mathbf{V}_p)^H)\mathbf{G}_{|p}\right) > 0$.

Now, by the continuity of singular values [8], it follows for sufficiently small $u > 0$ that $\mathrm{P}\left(\{\sigma_1\left((\mathbf{I}_T \otimes \mathbf{V}_p^H)\mathbf{G}_{|p}\right) \geq u\}\right) > 0$, which implies[3] that $\mathrm{P}\left(\{\sigma_1\left((\mathbf{I}_T \otimes \mathbf{V}_p^H)\mathbf{G}_{|p}\right) \geq u\}\right) \doteq \rho^0$ as $(\mathbf{I}_T \otimes (\mathbf{V}_p)^H)\mathbf{G}_{|p}$ is independent of ρ. This proves Lemma 1. ■

A. Lower Bound on Complexity

To tighten the lower bound and show that $\underline{c}(r) = \tilde{c}(r)$, we will use Lemma 1 to show that, irrespective of the ordering policy, the sphere decoder visits a total number of nodes that approaches $\rho^{\tilde{c}(r)}$, and does so with probability that is large compared to the ML error probability. We consider the codes that satisfy (18) and which appear almost surely as shown in Lemma 1.

Towards this we let $q \in \{1, \kappa\}$ be the largest integer for which $\frac{r}{n_T} - (1 - \mu_q^*) > 0$, in which case (17) takes the form

$$\tilde{c}(r) = T\sum_{j=1}^{q}\left(\frac{r}{n_T} - (1 - \mu_j^*)\right). \tag{20}$$

We quickly note that without loss of generality we can assume that $q \geq 1$ as otherwise $\bar{c}(r) = c(r) = 0$. Consequently it is the case that $\mu_j^* > 0$ for $j = 1, \cdots, q$.

We proceed to define four events Ω_1, Ω_2, Ω_3 and Ω_4 which we will prove to be jointly sufficient so that at layer $k = 2qT$, for some $q \in \{1, n_T\}$ the total number of nodes visited by sphere decoder is close to $\rho^{\tilde{c}(r)}$. These events are given by

$$\begin{aligned}\Omega_1 \triangleq \{&\mu_j^* - 2\delta < \mu_j < \mu_j^* - \delta, j = 1, \cdots, q \\ &0 < \mu_j < \delta, j = q+1, \cdots, n_T\},\end{aligned} \tag{21}$$

for a given small $\delta > 0$,

$$\Omega_2 \triangleq \{\sigma_1\left((\mathbf{I}_T \otimes \mathbf{V}_p^H)\mathbf{G}_{|p}\right) \geq u\}, \tag{22}$$

for some given $u > 0$ independent of ρ, where $p \triangleq n_T - q$,

$$\Omega_3 \triangleq \left\{\|\mathbf{w}\|^2 < \xi^2\right\}, \tag{23}$$

$$\Omega_4 \triangleq \left\{\|\mathbf{s}\| < \frac{1}{2}\rho^{\frac{rT}{\kappa}}\right\}. \tag{24}$$

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

It is shown in [1] that in the presence of events Ω_1, Ω_2 and Ω_3 we can remove the boundary constraints of ML-based SD in (6), which allows us to lower bound the number of nodes visited at layer k as (cf. [1, Lemma 1])

$$N_k \geq \prod_{i=1}^{k}\left[\frac{2\xi}{\sqrt{k}\sigma_i(\mathbf{R}_k)} - \sqrt{k}\right]^+. \tag{25}$$

Following the footsteps of [1, Lemma 2] it can be shown that in the presence of events Ω_1, Ω_2, Ω_3 and Ω_4

$$\sigma_i(\mathbf{R}_k) \dot{\leq} \rho^{-\frac{r}{2n_T} + \frac{3}{2}\delta + \frac{1}{2}(1-\mu^*_{\iota_{2T}(i)})}, \quad i = 1, \cdots, 2qT, \tag{26}$$

where $\iota_{2T}(i) \triangleq \left\lceil \frac{i}{2T}\right\rceil$. Consequently, going back to (25), we have that for $k = 2qT$ with $q \in \{1, n_T - 1\}$ we have that

$$N_{2qT} \dot{\geq} \rho^{\left(\sum_{i=1}^{2qT}\left(\frac{rT}{\kappa} - \frac{1}{2}(1-\mu^*_{\iota_{2T}(i)})\right) - 3qT\delta\right)} = \rho^{(\tilde{c}(r) - 3qT\delta)},$$

where the last equality follows from (20). For the case of $q = n_T$, it can be shown that

$$N_{2qT} \dot{\geq} \rho^{\sum_{i=1}^{\kappa}\left(\frac{rT}{\kappa} - \delta - \frac{1}{2}(1-\mu^*_{\iota_{2T}(i)})\right)} = \rho^{(\tilde{c}(r) - \kappa\delta)}.$$

Consequently for $q \in \{1, n_T\}$ we have that $N_{SD} \dot{\geq} \rho^{\tilde{c}(r) - K\delta}$ for small $\delta > 0$, where $K \in \{3qT, \kappa\}$. We note that (21)-(24) jointly imply that $N_{SD} \dot{\geq} \rho^{\tilde{c}(r) - K\delta}$. For some $\delta' \triangleq K\delta + \delta_1$, where $\delta > \delta_1 > 0$, it follows that

$$\mathrm{P}\left(N_{SD} \geq \rho^{\tilde{c}(r) - \delta'}\right) \geq \mathrm{P}\left(\Omega_1 \cap \Omega_2 \cap \Omega_3 \cap \Omega_4\right) \doteq \mathrm{P}\left(\Omega_1\right) \tag{27}$$

where exponential equality follows from the independence of the events Ω_1, Ω_2, Ω_3 and Ω_4 for i.i.d. Rayleigh fading statistics (see [9] on the independence of eigenvalues and eigenvectors) and from the fact that $\mathrm{P}(\Omega_2) \doteq \rho^0$ (cf. (19)), $\mathrm{P}(\Omega_3) \doteq \rho^0$ (cf.(8)) and $\mathrm{P}(\Omega_4) \doteq \rho^0$.With Ω_1 being an open set, we have that

$$-\lim_{\rho\to\infty}\frac{\mathrm{P}(\Omega_1)}{\log\rho} \leq \inf_{\boldsymbol{\mu}\in\Omega_1} I(\boldsymbol{\mu}) = I(\tilde{\boldsymbol{\mu}}) < I(\boldsymbol{\mu}^*) = d(r) \tag{28}$$

where $\tilde{\boldsymbol{\mu}} = \{\mu_1^* - 2\delta \cdots, \mu_q^* - 2\delta, 0, \cdots, 0\}$, where the last inequality follows from the monotonicity of the rate function $I(\boldsymbol{\mu})$ and where last equality follows from the fact that, by definition, $I(\boldsymbol{\mu}^*) = d(r)$. Consequently (27),(28) along with (15b) imply that $\underline{c}(r) = \tilde{c}(r)$, for arbitrarily small $\delta > 0$. The following directly holds.

Theorem 1: For $n_T \times n_R$ ($n_R \geq n_T$) quasi-static MIMO channels with i.i.d. Rayleigh fading statistics and any DMT optimal full-rate linear code (fixed but) randomly chosen from a continuous distribution over the real numbers, irrespective of the fixed or dynamically changing decoding order, the complexity exponent of the ML-based sphere decoder *almost surely*, in the choice of the DMT optimal lattice code, matches

[3]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 $\mathbf{V}_{2n_T}$.

the universal upper bound on the complexity exponent given by (16).

We caution reader that Theorem 1 does not imply that dynamic decoding orders can not provide exponential reduction in complexity for any code. Instead what it does imply is that, if a code is picked at random then almost surely its complexity exponent matches the upper bound given by (16). For any given specific code finding dynamically changing decoding orders that can guarantee reduction in the complexity exponent as compared to (16), remains a challenging open problem.

III. Complexity of MMSE-preprocessed lattice sphere decoding

It follows from [2, Theorem 1] that the universal upper bound on the complexity exponent of MMSE-preprocessed lattice sphere decoding, which holds irrespective of any fading statistics, full-rate code and decoding ordering policies, matches the upper bound on the complexity exponent for ML-based sphere decoding and is given by (16). In order to derive a lower bound on the complexity exponent of MMSE-preprocessed lattice sphere decoding in the presence of Lemma 1 and the codes specified therein, following the footsteps of the [2, Proof of Theorem 1] and the considering arguments presented in the previous section for the existence of the lower bound for ML-based sphere decoder that also hold for MMSE-preprocessed lattice sphere decoder, it can be shown that for all decoding ordering policies, the lower bound on the complexity exponent of MMSE-preprocessed lattice sphere decoder almost surely matches the universal upper bound. As a result, it implies that irrespective of the decoding order policies, ML and MMSE-preprocessed lattice sphere decoder share the same complexity for almost all DMT optimal full-rate linear codes. The following holds.

Theorem 2: Irrespective of the fixed or dynamically changing decoding order, the complexity exponent for MMSE-preprocessed lattice sphere decoding any (fixed but) randomly and uniformly chosen code (from an ensemble of DMT optimal full-rate linear codes) over the quasi-static MIMO channel with i.i.d. Rayleigh fading statistics almost surely, in the choice of DMT optimal lattice code, matches the complexity exponent of ML-based SD with or without MMSE preprocessing.

Under a basic *richness of codes* assumption, results of Theorem 1 and Theorem 2 are in fact further extended to hold, with probability one, over all full-rate codes. The specific assumption asks that the ensemble of full-rate lattice codes that achieve any suboptimal DMT performance, is sufficiently large so that the entries of the generator matrix corresponding to a certain DMT performance $d(r)$ accept a continuous distribution across the real numbers. This in turn implies that, almost surely in the random choice of a full-rate lattice design with a certain $d(r)$, a randomly picked generator matrix has independent columns and that the rank criterion in (18) holds.

We note that the above does not exclude the existence of full-rate lattice codes that reduce the complexity away from the universal bound. Such designs may exist, but they will belong to a set of measure zero.

The basic assumption of *richness of codes* allows for a meaningful rate-reliability-complexity tradeoff that holds, almost surely in the random choice of the full-rate code.

IV. Rate-Reliability-Complexity Tradeoff

Looking at (16) it can be seen how one can potentially tradeoff diversity gain with complexity by, for example, reducing the diversity gain to get a reduction in the required complexity exponent. This complexity and diversity gain relationship can be succinctly described by a rate-reliability-complexity tradeoff, which identifies with a concise description of the optimal diversity gain achievable in the presence of any computational constraint. The following holds for the quasi-static Rayleigh fading MIMO channel with $n_R \geq n_T$.

Theorem 3: With probability one in the random choice of a full-rate lattice design, the achievable diversity performance $d_{\mathcal{D}}(r)$ for ML-based SD with a run-time constraint $\rho^{c_{\mathcal{D}}(r)}$ flops, is uniquely described by

$$d_{\mathcal{D}}(r) = \min\{d(r), d_{\mathcal{D}}(r,x)\} \ \ \forall \ \ c_{\mathcal{D}}(r) \geq 0, \qquad (29)$$

where $d(r)$ is the optimal diversity gain (of uninterrupted brute force ML, for the given code), where $d_{\mathcal{D}}(r,x) \triangleq \lim_{\epsilon \to 0^+} d_{\mathcal{D}}(r, c_{\mathcal{D}}(r)+\epsilon)$, where $d_{\mathcal{D}}(r, c_{\mathcal{D}}(r)+\epsilon) \triangleq \inf I(\boldsymbol{\mu})$, such that $T\sum_{j=1}^{n_T}\left(\frac{r}{n_T} - (1-\mu_j)\right)^+ \geq c_{\mathcal{D}}(r)+\epsilon$, where $1 \geq \mu_1 \geq \cdots \geq \mu_{n_T} \geq 0$, and where the above holds irrespective of the fixed or dynamically changing decoding order.

Proof: The proof for Theorem 3 is given in [10]. ■

Following corollary holds directly from the Theorem 3.

Corollary 3a: For the $n \times n$ MIMO Rayleigh fading channel, and with probability one in the random choice of a full-rate lattice design, the achievable diversity gain of the ML and of the (MMSE-preprocessed) lattice SD with a run-time constraint of $\rho^{c_{\mathcal{D}}(r)}$ flops, is uniquely described for any fixed or dynamic decoding ordering policy by $d_{\mathcal{D}}(r) = \min\{d_{\mathrm{ML}}(r), d_{\mathcal{D}}(r, c_{\mathcal{D}})\}$, where $d_{\mathcal{D}}(r, c_{\mathcal{D}}) = K^2 + (2K+1)(\frac{c_{\mathcal{D}}(r)}{T} + 1 - (K+1)\frac{r}{n})$, and where $K = \left\lfloor \frac{n c_{\mathcal{D}}(r)}{rT} \right\rfloor$.

Acknowledgments

The research has received funding FP7 CONECT project, from the FP7 CELTIC SPECTRA project, from the Swedish Foundation for Strategic Research (SSF) / grant ICA08-0046, and from the project ANR-IMAGENET.

References

[1] J. Jaldén and P. Elia, "Sphere decoding complexity exponent for decoding full rate codes over the quasi-static MIMO channel," *IEEE Trans. Inf. Theory*, vol. 58, no. 9, pp. 5785 – 5803, Sep. 2012.

[2] A. Singh, P. Elia, and J. Jaldén, "Achieving a vanishing performance gap to exact lattice decoding at a subexponential complexity," *IEEE Trans. Inf. Theory*, vol. 58, no. 6, pp. 3692–3707, Jun. 2012.

[3] ——, " Complexity analysis for ML-based sphere decoder achieving a vanishing performance-gap to brute force ML decoding," in *Proc. Int. Zurich Seminar on Communications (IZS)*, Mar. 2012, pp. 127–130.

[4] E. Viterbo and J. Boutros, "A universal lattice code decoder for fading channels," *IEEE Trans. Inf. Theory*, vol. 45, no. 7, pp. 1639–1642, Jul. 1999.

[5] 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.

[6] 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.

[7] 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.

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

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

[10] A. Singh, P. Elia, and J. Jaldén. Supplement to ISIT 2013 submission - Rate-Reliability-Complexity tradeoff for ML and Lattice decoders. [Online]. Available: http://arunthenua.webs.com/suppISIT2013.pdf