Reduced-Basis Constrained Tree Search for Large-Scale MIMO Detection

Lanxin He[1], Zheng Wang[1], Jinming Wen[2]and Yongming Huang[1]
[1]School of Information Science and Engineering, Southeast University, Nanjing, China
[2]School of Mathematics, Jilin University, Jilin, China
E-mail: lanxin_he@seu.edu.cn; wznuaa@gmail.com; jinming.wen@mail.mcgill.ca; huangym@seu.edu.cn;

***Abstract*—Tree-search-based detection provides a promising performance-complexity trade-off for large-scale multiple-input multiple-output (MIMO) systems. In this work, we propose a reduced-basis constrained tree search (RB-CTS) framework for large-scale MIMO detection, where lattice reduction is incorporated to improve the conditioning of the effective channel. Specifically, Lenstra-Lenstra-Lovász (LLL) lattice reduction is employed to enable a modulation-insensitive constrained prefix search (PS), during which the probability of successful prefix retention is analytically characterized, followed by a Babai-based completion to construct full-length candidate solutions. Meanwhile, it is analytically shown that the proposed detector attains the full receive diversity order under i.i.d. Rayleigh fading. Numerical results demonstrate a favorable detection trade-off, especially for higher-order modulations.**

***Index Terms*—Large-scale MIMO detection, tree search, lattice reduction, receive diversity**

I. Introduction

Uplink detection in large-scale multiple-input multiple-output (MIMO) systems is a fundamental problem for realizing the potential performance gains [1], [2]. However, the maximum likelihood (ML) detector entails prohibitive computational complexity, rendering it impractical for hardware implementation and motivating the development of tree-search-based detection schemes that significantly reduce the search space [3], [4]. Despite their effectiveness, many existing tree-search-based methods — most notably sphere decoding — exhibit highly variable computational complexity and inherently sequential processing, which limits their suitability for real-time and hardware-constrained systems. To address these drawbacks, fixed-complexity variants and other structured tree search schemes have been proposed to provide predictable complexity while retaining near-ML performance.

In particular, the k-best algorithm and its variant [5], [6] perform tree expansion in a breadth-first manner, retaining only the best k partial candidates at each layer according to a predefined path metric, thereby achieving fixed and predictable computational complexity. Another representative approach is the fixed-complexity sphere decoder (FSD) [7], which adopts a depth-first search strategy with a predetermined expansion pattern: a subset of layers is fully expanded, while the remaining layers are resolved via successive interference cancellation (SIC). It has been shown that FSD can achieve full receive diversity provided that the number of full expansion (FE) layers p satisfies $p \geq \sqrt{n} - 1$ in an $n \times n$ complex MIMO system [8]. Various refinements of FSD have been proposed, including the introduction of channel-dependent switching points to reduce the average complexity [9], as well as efficient implementations tailored for large quadrature amplitude modulation (QAM) constellations in soft-output detection [10]. These developments make FSD particularly attractive for hardware implementation.

Despite these developments, existing FSD-based schemes still incur considerable computational overhead in the FE stage, as the number of visited nodes grows rapidly with the modulation order due to exhaustive enumeration. More broadly, this limitation reflects a common challenge shared by many structured tree-search-based detectors: controlling the growth of candidate expansions at each layer without sacrificing detection reliability. Motivated by this observation, we propose a reduced-basis constrained tree search (RB-CTS) framework for large-scale MIMO detection. By incorporating lattice reduction, the proposed framework improves the conditioning of the effective channel. As a result, the candidate distribution is more concentrated around the true transmitted symbols, thereby reducing the number of candidates required at each prefix search phase.

The main contributions of this paper are as follows. We propose a modulation-insensitive constrained prefix search that constructs a fixed-size candidate set independent of the modulation order, develop an analytical characterization of the corresponding successful prefix-retention probability, and show that the proposed RB-CTS framework achieves full receive diversity under i.i.d. Rayleigh fading.

II. System Model

Consider a real-valued $n \times n$ large-scale MIMO system obtained via an equivalent transformation of the complex model [4], where n denotes the number of receive and transmit antennas. Let $\mathbf{x} \in \mathcal{X}^n$ denote the transmitted signal, and the corresponding received signal $\mathbf{c} \in \mathbb{R}^n$ is given by

$$\mathbf{c} = \mathbf{H}\mathbf{x} + \mathbf{w}, \tag{1}$$

where $\mathbf{H} \in \mathbb{R}^{n \times n}$ is the channel matrix with i.i.d. real-valued Gaussian entries, $\mathbf{w} \in \mathbb{R}^n$ is the additive white Gaussian noise (AWGN) vector with zero mean and variance σ_{w}^2. The symbol alphabet $\mathcal{X} = \{\pm 1, \pm 3, \ldots, \pm(\sqrt{M} - 1)\}$ corresponds to an $\sqrt{M}$-ary pulse amplitude modulation (PAM) constellation

Corresponding author: Zheng Wang (email: wznuaa@gmail.com)

associated with square M-QAM. The optimal ML detection problem is formulated as

$$\hat{\mathbf{x}}_{\mathrm{ML}} = \arg\min_{\mathbf{x}\in\mathcal{X}^n} \|\mathbf{H}\mathbf{x} - \mathbf{c}\|^2, \tag{2}$$

which naturally corresponds to the closest vector problem (CVP) in lattice decoding [11]. The n-dimensional lattice Λ generated by the full-rank matrix $\mathbf{H}$ is defined as:

$$\Lambda = \{\mathbf{H}\mathbf{x} : \mathbf{x} \in \mathbb{Z}^n\}, \tag{3}$$

where $\mathbf{H}$ serves as the lattice basis.

III. Reduced-basis Constrained Tree Search

In the context of MIMO detection, lattice reduction is applied to the channel matrix by seeking a unimodular transformation $\mathbf{T} \in \mathbb{Z}^{n\times n}$ such that

$$\widetilde{\mathbf{H}} = \mathbf{H}\mathbf{T}. \tag{4}$$

The resulting basis $\widetilde{\mathbf{H}} = \{\widetilde{\mathbf{h}}_1, \ldots, \widetilde{\mathbf{h}}_n\}$ is said to be *LLL-reduced* if it satisfies the following two conditions [11]:

$$|\mu_{i,j}| \le \tfrac{1}{2}, \quad 1 \le j < i \le n, \tag{5}$$

$$\|\hat{\mathbf{h}}_i + \mu_{i,i-1}\hat{\mathbf{h}}_{i-1}\|^2 \ge \delta\|\hat{\mathbf{h}}_{i-1}\|^2, \quad 1 < i \le n, \tag{6}$$

where $\hat{\mathbf{h}}_i$ denotes the i-th Gram-Schmidt (GS) vector of the basis $\widetilde{\mathbf{H}}$, $\mu_{i,j}$ are the corresponding GS coefficients, and $\delta \in (1/4, 1)$ is the Lovász constant. These two conditions imply that the lengths of the GS vectors do not decrease too rapidly, since

$$\|\hat{\mathbf{h}}_i\|^2 \ge (\delta - \mu_{i,i-1}^2)\|\hat{\mathbf{h}}_{i-1}\|^2 \ge (\delta - \tfrac{1}{4})\|\hat{\mathbf{h}}_{i-1}\|^2. \tag{7}$$

As a result, the reduced basis $\widetilde{\mathbf{H}}$ typically exhibits improved orthogonality, which is beneficial for the subsequent tree search.

A. Algorithm Description

We first describe the proposed reduced-basis constrained tree search (RB-CTS) algorithm on an ordered upper-triangular system. Specifically, an appropriate column permutation $\mathbf{P}$ is applied to the reduced lattice basis $\widetilde{\mathbf{H}}$, followed by a QR decomposition,

$$\widetilde{\mathbf{H}}\mathbf{P} = \mathbf{Q}\mathbf{R}, \tag{8}$$

where $\mathbf{Q} \in \mathbb{R}^{n\times n}$ is an orthogonal matrix and $\mathbf{R} \in \mathbb{R}^{n\times n}$ is an upper-triangular matrix. The resulting equivalent upper-triangular system is given by

$$\mathbf{y} = \mathbf{R}\mathbf{z} + \mathbf{n}, \tag{9}$$

where $\mathbf{y} = \mathbf{Q}^{\mathrm{T}}\mathbf{c}$, $\mathbf{z} = \mathbf{P}^{\mathrm{T}}\mathbf{T}^{-1}\mathbf{x} \in \mathbb{Z}^n$, and $\mathbf{n} = \mathbf{Q}^{\mathrm{T}}\mathbf{w}$ with $\sigma_n^2 = \sigma_w^2$. Based on (9), the proposed RB-CTS algorithm performs reverse sequential detection from layer n to layer 1. Let $\hat{z}_i$ denote the estimate at the i-th layer. The associated *soft center* is calculated according to

$$\tilde{z}_i = \frac{1}{r_{i,i}}\Big(y_i - \sum_{k=i+1}^{n} r_{i,k}\hat{z}_k\Big), \tag{10}$$

where $r_{i,k}$ denotes the (i,k)-th entry of $\mathbf{R}$. Then, the estimate $\hat{z}_i$ in RB-CTS is successively determined through the following two phases:

*a) **Constrained Prefix Search (PS)**:* This phase operates on the last p layers of the ordered $\mathbf{R}$, with the index set $\mathcal{I}_{\mathrm{PS}} = \{n, n-1, \ldots, n-p+1\}$, where the *PS depth* p is a predefined parameter, and its selection guideline is provided in Section IV. A limited number of candidate symbols around the soft center $\tilde{z}_i$ are retained at each prefix layer. Specifically, the $2S$ nearest integer points centered at $\lfloor \tilde{z}_i \rfloor$ are selected, i.e.,

$$\mathcal{C}_{\mathrm{PS}}^i \triangleq \{\lfloor \tilde{z}_i \rfloor - S+1, \ldots, \lfloor \tilde{z}_i \rfloor, \ldots, \lfloor \tilde{z}_i \rfloor + S\}, \quad i \in \mathcal{I}_{\mathrm{PS}}, \tag{11}$$

where $\lfloor \cdot \rfloor$ denotes the floor operation and the integer S is the per-layer *search radius*. In practice, a relatively small value of S is sufficient (e.g., $S = 2$ for 64-QAM). The resulting prefix candidate set is defined as the Cartesian product

$$\mathcal{Z}_{\mathrm{PS}} \triangleq \mathcal{C}_{\mathrm{PS}}^{n-p+1} \times \cdots \times \mathcal{C}_{\mathrm{PS}}^{n}, \tag{12}$$

where each element $\hat{\mathbf{z}}_{n-p+1:n} \in \mathcal{Z}_{\mathrm{PS}}$ represents a surviving prefix vector.

*b) **Babai Completion**:* This phase completes the symbol estimates for the remaining $n - p$ layers, with the index set $\mathcal{I}_{\mathrm{B}} = \{n-p, n-p-1, \ldots, 1\}$. For each retained prefix $\hat{\mathbf{z}}_{n-p+1:n}$, the remaining layers are deterministically determined via Babai's nearest-plane algorithm [12] according to

$$\hat{z}_i = \lfloor \tilde{z}_i \rceil, \qquad i \in \mathcal{I}_{\mathrm{B}}, \tag{13}$$

where $\lfloor \cdot \rceil$ denotes rounding to the nearest integer. This procedure yields a unique Babai completion

$$\mathbf{z}_{\mathrm{B}}(\hat{\mathbf{z}}_{n-p+1:n}) \triangleq \hat{\mathbf{z}}_{1:n-p}. \tag{14}$$

In the context of MIMO detection, this Babai completion is equivalent to SIC detection. The completed lattice point corresponding to a given prefix $\hat{\mathbf{z}}_{n-p+1:n}$ is constructed via the mapping

$$\Phi(\hat{\mathbf{z}}_{n-p+1:n}) \triangleq \begin{bmatrix} \mathbf{z}_{\mathrm{B}}(\hat{\mathbf{z}}_{n-p+1:n}) \\ \hat{\mathbf{z}}_{n-p+1:n} \end{bmatrix} \in \mathbb{Z}^n. \tag{15}$$

Among all completed points generated from the prefix set, the final hard decision is obtained as

$$\hat{\mathbf{z}}_{\text{RB-CTS}} = \arg\min_{\mathbf{z}\in\Phi(\mathcal{Z}_{\mathrm{PS}})} \big\|\mathbf{y} - \mathbf{R}\mathbf{z}\big\|^2. \tag{16}$$

Compared with FSD [8], which enumerates all constellation points during its full-expansion stage and generates $\sqrt{M}^p$ candidates, the proposed RB-CTS reduces the candidate set cardinality to $(2S)^p$. This reduction becomes increasingly pronounced for higher-order modulations.

In this work, we adopt a column ordering strategy similar to that used in FSD [8] to obtain the permutation matrix $\mathbf{P}$, where an iterative ordering is performed on the columns of $\widetilde{\mathbf{H}}$. At iteration i, a QR decomposition

$$\widetilde{\mathbf{H}}_i = \mathbf{Q}_i\mathbf{R}_i \tag{17}$$

is computed, and the metric

$$d_k = \big\|\mathbf{R}_i^{-1}\mathbf{e}_k\big\|_2^2, \qquad k = 1, \ldots, i \tag{18}$$

is evaluated, where $\mathbf{e}_k$ is the kth standard basis vector, and $\mathbf{Q}_i$ and $\mathbf{R}_i$ denote the orthogonal and upper-triangular matrices at

the i-th iteration, respectively. During iterations corresponding to the Babai completion phase, i.e., $i \in \mathcal{I}_{\mathrm{B}}$, the column associated with the minimum d_k is selected, whereas for the constrained PS phase, i.e., $i \in \mathcal{I}_{\mathrm{PS}}$, the column with the maximum d_k is chosen. This criterion amounts to selecting layers with smaller or larger zero-forcing noise amplification, respectively [13]. The resulting column ordering is represented by the permutation matrix $\mathbf{P}$.

B. Reliability of the PS Phase

To quantify the reliability of the PS phase, we examine how likely the correct symbol vector is retained during this phase. Substituting $y_i = r_{i,i}z_i + \sum_{k>i} r_{i,k}z_k + n_i$ into (10) yields

$$\tilde{z}_i - z_i = \frac{n_i}{r_{i,i}} + \sum_{k=i+1}^{n} \frac{r_{i,k}}{r_{i,i}}(z_k - \hat{z}_k), \tag{19}$$

where $n_i \sim \mathcal{N}(0, \sigma_{\mathrm{n}}^2)$. Equation (19) shows that the deviation of the soft center $\tilde{z}_i$ from the true symbol z_i consists of two components: a noise term scaled by $1/|r_{i,i}|$ and an interference term caused by error propagation from previously detected layers. As ensured by (7), LLL reduction improves the orthogonality of the effective lattice basis, resulting in more balanced diagonal entries $|r_{i,i}|$ and reduced relative off-diagonal coupling $|r_{i,k}/r_{i,i}|$. Consequently, both noise amplification and error propagation are mitigated, leading to a tighter concentration of $\tilde{z}_i$ around z_i.

Motivated by this concentration behavior, we next bound the probability that the correct symbol vector is retained during the PS phase, denoted as $\Pr(\mathbf{x} \in \mathcal{S}_{\mathrm{PS}})$. Here, with a slight abuse of notation, the event $\mathbf{x} \in \mathcal{S}_{\mathrm{PS}}$ is defined to indicate that the PS prefix of the associated true lattice vector $\mathbf{z}$ is retained, i.e., $\mathbf{z}_{n-p+1:n} \in \mathcal{Z}_{\mathrm{PS}}$.

Lemma 1. *The probability of successful prefix retention during the PS phase satisfies*

$$\Pr(\mathbf{x} \in \mathcal{S}_{\mathrm{PS}}) \geq 1 - p \cdot \exp\left(-\frac{S^2 \min_i |r_{i,i}|^2}{2\kappa\sigma_{\mathrm{n}}^2}\right), \tag{20}$$

where p is the PS depth, S is the per-layer prefix search radius, and $\kappa > 0$ is a constant determined by the residual interference level.

Proof. The result follows by first upper bounding the probability that the correct prefix is excluded during the PS phase. Conditioned on previous decisions, the interference term in (19) can be regarded as a data-dependent perturbation. Accordingly, we introduce an analytical dispersion parameter

$$\varepsilon_i^2 \triangleq \kappa \frac{\sigma_{\mathrm{n}}^2}{|r_{i,i}|^2}, \tag{21}$$

where the constant κ absorbs the effect of the interference term and depends on the quality of the reduced basis. In general, a larger δ is associated with a smaller κ.

At layer i, the event that the correct symbol z_i is not retained in the PS candidate set $\mathcal{C}_{\mathrm{PS}}^i$ implies $|\tilde{z}_i - z_i| > S$. Applying a Gaussian tail bound yields

$$\Pr\big(z_i \notin \mathcal{C}_{\mathrm{PS}}^i\big) \leq 2Q\left(\frac{S}{\varepsilon_i}\right) \leq \exp\left(-\frac{S^2}{2\varepsilon_i^2}\right), \tag{22}$$

where $Q(x) \triangleq \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-t^2/2}\,\mathrm{d}t$ denotes the Gaussian Q-function. Applying the union bound over the p PS layers yields

$$\begin{aligned}\Pr(\mathbf{x} \notin \mathcal{S}_{\mathrm{PS}}) &\leq \sum_{i \in \mathcal{I}_{\mathrm{PS}}} \Pr\big(z_i \notin \mathcal{C}_{\mathrm{PS}}^i\big) \leq p \cdot \max_{i \in \mathcal{I}_{\mathrm{PS}}} \Pr\big(z_i \notin \mathcal{C}_{\mathrm{PS}}^i\big) \\ &= p \cdot \exp\left(-\frac{S^2 \min_i |r_{i,i}|^2}{2\kappa\sigma_{\mathrm{n}}^2}\right), \end{aligned} \tag{23}$$

which yields the bound in (20). □

Lemma 1 reveals two important implications for the proposed RB-CTS algorithm. First, LLL reduction improves the orthogonality of the effective lattice basis, which has two beneficial effects on the exponential bound in (20). On one hand, by balancing the diagonal entries $|r_{i,i}|$, LLL reduction alleviates extremely small values and thus tends to increase the minimum diagonal entry $\min_i |r_{i,i}|$. On the other hand, the reduced basis exhibits weaker inter-layer coupling, which mitigates error propagation and effectively reduces the interference-related constant κ. As a result, the exponential term in (20) decays faster, significantly lowering $\Pr(\mathbf{x} \notin \mathcal{S}_{\mathrm{PS}})$. Second, owing to this exponential concentration behavior, a relatively small integer radius S is sufficient to ensure a high probability of successful prefix retention during the PS phase. This observation justifies retaining only a limited number of nearest candidates at each PS layer, thereby substantially reducing the number of visited search nodes.

C. Complexity Analysis

We now evaluate the computational complexity of the proposed RB-CTS algorithm in terms of (i) the number of visited tree nodes N_{node} and (ii) the floating-point operations (FLOPs) per visited node. Let N_{PS} and N_{B} denote the numbers of visited nodes in the PS phase and the Babai completion phase, respectively, such that $N_{\mathrm{node}} = N_{\mathrm{PS}} + N_{\mathrm{B}}$.

At the PS phase, $2S$ candidate symbols are retained at each layer. As a result, the total number of visited nodes during the PS phase is given by

$$N_{\mathrm{PS}} = \sum_{\ell=1}^{p} (2S)^\ell = \frac{(2S)^{p+1} - 2S}{2S - 1}. \tag{24}$$

Since $(2S)^p$ PS prefixes survive and each prefix is deterministically completed over the remaining $(n-p)$ layers via Babai completion, the number of visited nodes in the completion phase is

$$N_{\mathrm{B}} = (n-p)(2S)^p. \tag{25}$$

Therefore, the total number of visited nodes of the proposed RB-CTS algorithm is

$$N_{\mathrm{node}} = \frac{(2S)^{p+1} - 2S}{2S - 1} + (n-p)(2S)^p = O(n(2S)^p), \tag{26}$$

which is essentially insensitive to the modulation order for a fixed small integer S.

Each visited node accounts for one partial Euclidean distance update on the upper-triangular system, whose computational cost scales linearly with the dimension n. As a result,

the overall search complexity of RB-CTS can be approximated as

$$C_{\text{RB-CTS}} = O(n)\, N_{\text{node}} = O\big(n^2(2S)^p\big)\,. \tag{27}$$

In addition to the tree search, the proposed detector requires a one-time LLL lattice reduction as preprocessing, whose computational complexity is $O(n^3 \log n)$ [14]. The column ordering applied after lattice reduction involves a sequence of QR decompositions and metric evaluations, incurring at most $O(n^3)$ computational cost. Since the subsequent tree search complexity grows exponentially with the PS depth p, both the LLL reduction and the ordering overhead are asymptotically dominated by the search phase.

IV. Receive Diversity Analysis

In this section, we analyze the receive diversity of the proposed RB-CTS detector. Define the minimum distance of a lattice generated by a full-column-rank matrix $\mathbf{A}$ as

$$d(\mathbf{A}) \triangleq \min_{\mathbf{z}\in\mathbb{Z}^n\setminus\{\mathbf{0}\}} \|\mathbf{A}\mathbf{z}\|_2. \tag{28}$$

Due to the invariance of the lattice minimum distance under unimodular, permutation, and orthogonal transformations [12], we have $d(\widetilde{\mathbf{H}}) = d(\mathbf{H})$, and the following Lemma holds [15].

Lemma 2. *There exists a constant $c > 0$, depending only on the dimension n and the LLL parameter δ, such that if*

$$\|\mathbf{w}\|_2 \le c\, d(\widetilde{\mathbf{H}}), \tag{29}$$

then all decisions in the Babai completion phase are correct.

Then, we establish the Theorem below.

Theorem 1. *For an i.i.d. Rayleigh fading MIMO channel, the proposed RB-CTS achieves the full receive diversity order, i.e.,*

$$\lim_{\rho\to\infty} -\frac{\log P_e^{\text{RB-CTS}}}{\log \rho} = n, \tag{30}$$

where $\rho = 1/\sigma_{\text{n}}^2$ denotes the signal-to-noise ratio (SNR).

Proof. The error event can be decomposed by conditioning on whether the correct vector is retained after the PS phase and whether the noise realization lies within the reliable decoding region of the Babai completion phase, yielding

$$\begin{aligned} P_e^{\text{RB-CTS}} &= \Pr(\hat{\mathbf{x}} \neq \mathbf{x}) \\ &\le \Pr(\mathbf{x} \notin \mathcal{S}_{\text{PS}}) + \Pr(\hat{\mathbf{x}} \neq \mathbf{x} \mid \mathbf{x} \in \mathcal{S}_{\text{PS}},\, \|\mathbf{w}\|_2 \le c\, d(\widetilde{\mathbf{H}})) \\ &\quad + \Pr(\|\mathbf{w}\|_2 > c\, d(\widetilde{\mathbf{H}})), \end{aligned} \tag{31}$$

where $c > 0$ is the constant specified in Lemma 2. From Lemma 1, the probability that the correct vector is excluded in the PS phase satisfies

$$\Pr(\mathbf{x} \notin \mathcal{S}_{\text{PS}}) \le p \cdot \exp\left(-\frac{S^2 \min_i |r_{i,i}|^2}{2\kappa\sigma_{\text{n}}^2}\right) = o(\rho^{-n}), \tag{32}$$

i.e., it decays exponentially fast in ρ and is asymptotically negligible compared with ρ^{-n}.

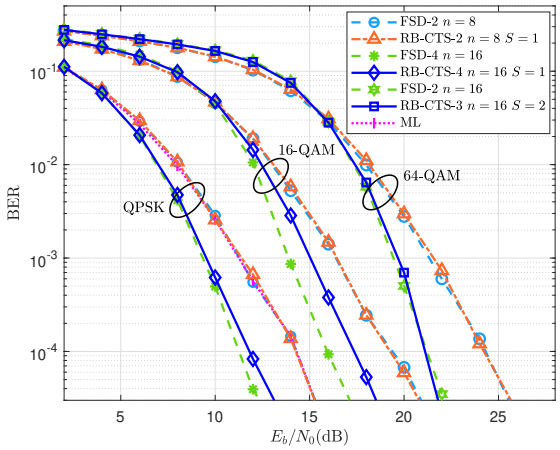

Fig. 1. Performance comparison for $n \times n$ real-valued MIMO systems.

Conditioned on $\mathbf{x} \in \mathcal{S}_{\text{PS}}$ and $\|\mathbf{w}\|_2 \le c\, d(\widetilde{\mathbf{H}})$, Lemma 2 implies that the true vector $\mathbf{x}$ is contained in the final candidate list. Consequently, the proposed detector performs ML selection over this list, yielding

$$\Pr(\hat{\mathbf{x}} \neq \mathbf{x} \mid \mathbf{x} \in \mathcal{S}_{\text{PS}},\, \|\mathbf{w}\|_2 \le c\, d(\widetilde{\mathbf{H}})) \le P_e^{\text{ML}}, \tag{33}$$

where P_e^{ML} denotes the error probability of the ML detector. It remains to bound $\Pr(\|\mathbf{w}\|_2 > c\, d(\widetilde{\mathbf{H}}))$. Choose any $\alpha \in (0, \frac{1}{2})$. Since $d(\widetilde{\mathbf{H}}) = d(\mathbf{H})$, by the union bound,

$$\Pr(\|\mathbf{w}\|_2 > c d(\mathbf{H})) \le \Pr(d(\mathbf{H}) \le \rho^{-\alpha}) + \Pr(\|\mathbf{w}\|_2 > c\rho^{-\alpha}). \tag{34}$$

Let $\dot{\le}$ denote exponential inequality in ρ, i.e., $f(\rho) \dot{\le} g(\rho)$ if $\limsup_{\rho\to\infty} \frac{\log f(\rho)}{\log \rho} \le \limsup_{\rho\to\infty} \frac{\log g(\rho)}{\log \rho}$. It then follows that, by choosing $\alpha \in (0, \frac{1}{2})$ arbitrarily close to $\frac{1}{2}$,

$$\Pr(\|\mathbf{w}\|_2 > c d(\mathbf{H})) \dot{\le} \rho^{-n}, \tag{35}$$

similar to the proof in [15]. Since ML detection achieves the full receive diversity order n in i.i.d. Rayleigh fading, combining (32), (33) and (35) implies that the proposed RB-CTS attains the full receive diversity order. □

Remark 1. *The Lovász constant δ affects only the constant c in Lemma 2 and thus influences the SNR gap but not the diversity order [15]. Besides, inspired by the ordering analysis in [8] associated with* (18)*, for a real-valued system, selecting the PS depth as $p \ge 2\big(\sqrt{n/2} - 1\big)$ is sufficient in practice to retain full receive diversity of the proposed RB-CTS.*

V. Numerical Results

Simulation results are provided to evaluate the bit error rate (BER) performance, diversity order, and complexity characteristics of the proposed RB-CTS detector. Within RB-CTS, the classical FSD [8] can be viewed as a representative special case that fully enumerates all constellation points over the first p layers; hence, it is included as a natural benchmark. The Lovász constant in the applied LLL reduction is fixed to $\delta = 0.75$. Fig. 1 compares the BER performance of RB-CTS and FSD over $n \times n$ real-valued MIMO systems under QPSK, 16-QAM, and 64-QAM. Here, "RB-CTS-p" denotes RB-CTS with p PS layers, while "FSD-p" denotes the corresponding

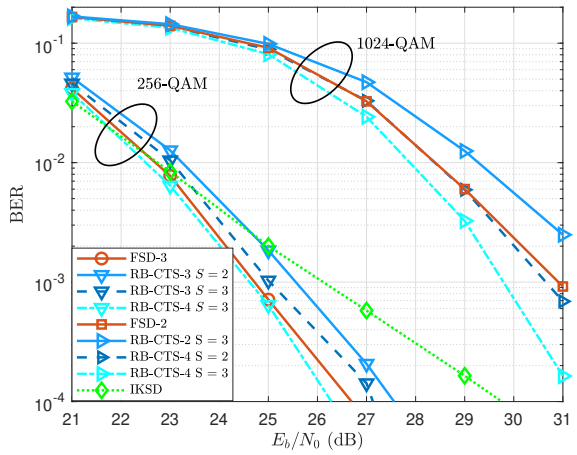

Fig. 2. Performance comparison under 32×32 real-valued MIMO system.

FSD with p fully-expanded layers. For most configurations, the parameter p satisfies $p \geq 2(\sqrt{n/2}-1)$ (i.e., $p=2$ for 8×8 and $p=4$ for 16×16 systems), under which the FSD curves exhibit the same asymptotic slope as ML detection and can be regarded as full-diversity references [8]. Although, for the 16×16 system, a clear performance gap is observed under 16-QAM, RB-CTS employs only 2^4 candidate vectors, compared to $4^4 = 2^8$ candidates used by FSD. Under 64-QAM, FSD with $p = 4$ is computationally infeasible and thus omitted; instead, a pair of curves with equal candidate-node complexity is highlighted, corresponding to FSD with $p = 2$ and RB-CTS with $S = 2$, $p = 3$, for reference. Overall, the results confirm that RB-CTS achieves the expected full diversity order.

Fig. 2 compares the BER performance under a 32×32 real-valued MIMO system with 256-QAM and 1024-QAM. The corresponding numbers of visited search nodes are summarized in Table I. The improved k-best sphere decoding (IKSD) [6] with $k = 16$ and a fixed threshold $\Delta = 3\sigma_n^2$ is also included as a reference. Under higher-order modulation, the proposed RB-CTS achieves comparable BER performance with substantially fewer visited nodes than FSD. Specifically, under 256-QAM, RB-CTS-4 with $S = 4$ slightly outperforms FSD-3 while visiting only 37,842 nodes, which is significantly lower than the 123,152 nodes required by FSD-3 (see Table I). A similar trend is observed under 1024-QAM, where RB-CTS-4 with $S = 2$ achieves comparable performance to FSD-2 with a lower search complexity. Moreover, RB-CTS-4 with $S = 3$ under 1024-QAM significantly outperforms FSD-2 while incurring only a modest increase in the number of visited nodes. These results demonstrate the favorable performance-complexity trade-off enabled by RB-CTS in high-order modulation scenarios. For completeness, we note that applying FSD-3 under 1024-QAM would require approximately 9.8×10^5 visited nodes, which is computationally prohibitive and therefore omitted from Fig. 2.

VI. Conclusion

In this paper, we proposed a reduced-basis constrained tree search (RB-CTS) framework for large-scale MIMO detection. While leveraging standard lattice reduction and Babai completion, the proposed method is distinguished by a modulation-insensitive constrained prefix search with fixed candidate size independent of the modulation order. Its reliability is characterized via a prefix-retention probability analysis, and full receive diversity is shown to be preserved despite the reduced search space.

TABLE I
Average visited nodes for Fig. 2.

Method	256-QAM	1024-QAM
RB-CTS	1,940 ($p=3$, $S=2$) 6,522 ($p=3$, $S=3$) 37,842 ($p=4$, $S=3$)	1,122 ($p=2$, $S=3$) 7,508 ($p=4$, $S=2$) 37,842 ($p=4$, $S=3$)
FSD	123,152	31,776

Acknowledgment

This work was supported in part by the National Key R&D Program of China under Grants No.2023YFC2205501, and in part by National Natural Science Foundation of China under Grants No.62371124.

References

[1] M. Ke, Z. Gao, Y. Wu, X. Gao, and R. Schober, "Compressive sensing-based adaptive active user detection and channel estimation: Massive access meets massive MIMO," *IEEE Transactions on Signal Processing*, vol. 68, pp. 764–779, 2020.

[2] Y. Gao, Q. Chen, L. He, Z. Wang, Y. Huang, J. Zhang, Y. Gao, and Z. Xu, "Artificial intelligence enabled joint channel estimation and signal detection for massive MIMO systems," *Chinese Journal of Electronics*, vol. 35, no. 1, pp. 178–195, 2026.

[3] E. Agrell, T. Eriksson, A. Vardy, and K. Zeger, "Closest point search in lattices," *IEEE Transactions on Information Theory*, vol. 48, no. 8, pp. 2201–2214, 2002.

[4] Z. Wang, C. Ling, S. Jin, Y. Huang, and F. Gao, "Probabilistic searching for MIMO detection based on lattice Gaussian distribution," *IEEE Transactions on Communications*, vol. 72, no. 1, pp. 85–100, 2024.

[5] Z. Guo and P. Nilsson, "Algorithm and implementation of the K-best sphere decoding for MIMO detection," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 3, pp. 491–503, 2006.

[6] S. Han, T. Cui, and C. Tellambura, "Improved K-best sphere detection for uncoded and coded MIMO systems," *IEEE Wireless Communications Letters*, vol. 1, no. 5, pp. 472–475, 2012.

[7] L. G. Barbero and J. S. Thompson, "Fixing the complexity of the sphere decoder for MIMO detection," *IEEE Transactions on Wireless Communications*, vol. 7, no. 6, pp. 2131–2142, 2008.

[8] J. Jalden, L. G. Barbero, B. Ottersten, and J. S. Thompson, "The error probability of the fixed-complexity sphere decoder," *IEEE Transactions on Signal Processing*, vol. 57, no. 7, pp. 2711–2720, 2009.

[9] K.-C. Lai, C.-C. Huang, and J.-J. Jia, "Variation of the fixed-complexity sphere decoder," *IEEE Communications Letters*, vol. 15, no. 9, pp. 1001–1003, 2011.

[10] Y.-M. Chen, Y.-X. Dai, S.-J. Jhang, and Y.-L. Ueng, "An efficient soft-output fixed-complexity sphere decoder for large QAM constellations," *IEEE Transactions on Vehicular Technology*, vol. 74, no. 12, pp. 19308–19322, 2025.

[11] A. K. Lenstra, H. W. Lenstra, and L. Lovász, "Factoring polynomials with rational coefficients," *Mathematische Annalen*, vol. 261, no. 4, pp. 515–534, 1982.

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

[13] B. Hassibi and H. Vikalo, "On the sphere-decoding algorithm I. expected complexity," *IEEE Transactions on Signal Processing*, vol. 53, pp. 2806–2818, Aug. 2005.

[14] C. Ling, W. H. Mow, and N. Howgrave-Graham, "Reduced and fixed-complexity variants of the LLL algorithm for communications," *IEEE Transactions on Communications*, vol. 61, no. 3, pp. 1040–1050, 2013.

[15] M. Taherzadeh, A. Mobasher, and A. K. Khandani, "LLL reduction achieves the receive diversity in MIMO decoding," *IEEE Transactions on Information Theory*, vol. 53, pp. 4801–4805, Dec. 2007.