Sphere Decoding Algorithms With Improved Radius Search

Wanlun Zhao and Georgios B. Giannakis, *Fellow, IEEE*

***Abstract*—We start by identifying a relatively efficient version of sphere decoding algorithm (SDA) that performs exact maximum-likelihood (ML) decoding. We develop novel algorithms based on an improved increasing radius search (IIRS), which offer error performance and decoding complexity between two extremes: the ML receiver and the nulling–canceling (NC) receiver with detection ordering. With appropriate choices of parameters, our IIRS offers the flexibility to trade error performance for complexity. We provide design intuitions and guidelines, analytical parameter specifications, and a semianalytical error-performance analysis. Simulations illustrate that IIRS achieves considerable complexity reduction, while maintaining performance close to ML.**

***Index Terms*—Closest-point algorithm, multiple-input multiple-output (MIMO) decoding, sphere decoding.**

I. Introduction

CONSIDER the following generic model:

$$\mathbf{y} = \mathbf{H}\,\mathbf{s} + \mathbf{v} \tag{1}$$

where $\mathbf{y}, \mathbf{v} \in \mathbb{R}^M$, $\mathbf{s} \in \mathbb{Z}^N$, $\mathbf{H} \in \mathbb{R}^{M\times N}$ has full column rank, $M \geq N$, and $\mathbb{Z}$ and $\mathbb{R}$ denote the sets of integers and real numbers, respectively. Operating on $\mathbf{s}$, the matrix $\mathbf{H}$ generates a lattice that we denote as $\Lambda(\mathbf{H}) := \{\mathbf{x} = \mathbf{H}\mathbf{s} | \mathbf{s} \in \mathbb{Z}^N\}$. The *closest-point problem* is: Given $\mathbf{y} \in \mathbb{R}^M$ and a lattice Λ with a known *generator* $\mathbf{H}$, find the lattice vector $\hat{\mathbf{x}} \in \Lambda$ that minimizes the Euclidean distance from $\mathbf{y}$ to $\hat{\mathbf{x}}$; that is, $\hat{\mathbf{x}} = \arg\min_{\mathbf{x}\in\Lambda} \|\mathbf{y} - \mathbf{x}\|^2$, where $\|\cdot\|$ represents the vector norm.

In a wireless communication context, $\mathbf{s}$, $\mathbf{y}$, and $\mathbf{v}$ are the transmitted, received, and the additive white Gaussian noise (AWGN) vectors, whereas $\mathbf{H}$ contains the channel coefficients. The distribution of $\mathbf{v}$ is $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, where $\mathcal{N}(\cdot,\cdot)$ represents the Gaussian distribution, and $\mathbf{H}$ is a random matrix, often with known statistical properties. Furthermore, instead of the whole integer lattice $\mathbb{Z}^N$, $\mathbf{s}$ is usually drawn from a finite subset $\mathcal{S}^N \subset \mathbb{Z}^N$. In block decoding, we are interested in determining the maximum-likelihood (ML) estimate of $\mathbf{s}$, subject to the finite alphabet (FA) constraints $\hat{\mathbf{s}}_{\text{ML}} = \arg\min_{\mathbf{s}\in\mathcal{S}^N} \|\mathbf{y} - \mathbf{H}\mathbf{s}\|^2$.

Under FA constraints, closest-point algorithms can be employed to find $\hat{\mathbf{s}}_{\text{ML}}$ in various applications, including space–time decoding, equalization of block transmissions, and multiuser detection. A well-known closest-point algorithm, the sphere decoding algorithm (SDA), was introduced to determine vectors with small norms in an arbitrary lattice [6], but has gained popularity in lattice-code decoding [14], symbol-synchronous code-division multiple-access (CDMA) detection [3], and space–time decoding [9], [16]. A variate of SDA, first used by Schnorr and Euchner and which appeared recently in both [1] and [4], includes an ordering mechanism to improve search efficiency. With AWGN and the random channel model, where the entries of $\mathbf{H}$ are independent and identically distributed (i.i.d.) $\mathcal{N}(0,1)$, the average complexity of the SDA was found in [9] and [10], along with a method to determine the initial search radius. To achieve ML error performance, SDA with increasing radius search (IRS) was also suggested in [10] and [14]. A novel closest-point algorithm that examines candidates for $\mathbf{s}$ in a descending probability order was developed in [17]. Soft versions of SDA are also available to enable near-capacity performance of multiple-antenna systems [11], [15]. Other related works on SDA include [5] and [8].

We first review SDA and its improved versions in Section II. In Section III, we derive several variants of SDA with improved IRS (IIRS). Exploiting the AWGN model only, our algorithm achieves near-ML error performance with considerably reduced complexity. We conclude in Section IV.

Notation: Upper (lower) bold face letters denote matrices (column vectors); $(\cdot)^T$ and $(\cdot)^\dagger$ denote transpose and pseudoinverse, respectively; $\chi^2_n(\sigma^2)$ denotes the Chi-square distribution with probability density function (pdf) $f(x) = 1/(2^{n/2}\sigma^n\Gamma(n/2))x^{n/2-1}\exp(-x/(2\sigma^2))$, where $\Gamma(n)$ represents the Gamma function. For brevity, the standard Chi-square distribution with $\sigma^2 = 1$ is denoted by χ^2_n.

II. Improved SDAs

The basic idea of SDA is to search in a hypersphere of radius r centered at the received vector $\mathbf{y}$. Even though points in this hypersphere are searched exhaustively, calculations are performed recursively, based on a search tree to enable reusing intermediate computations. For detailed discussions on SDA, interested readers can check, e.g., [1], [6], and [14].

A. Schnorr–Euchner Variate of SDA

Recently, a variant of the SDA appeared in both [1] and [4]. Since this version of SDA was first used by Schnorr and Euchner, it is abbreviated as the SE-SDA [1]. The key difference of the SE-SDA from the conventional SDA lies in a simple ordering of the candidates at each dimension. Specifically, the candidates are examined in a descending probabilistic order. A careful examination of SE-SDA reveals that the first candidate of $\mathbf{s}$ checked is always $\mathbf{s}_{\text{NC}}$, which is the nulling–canceling (NC)

Paper approved by B. Hochwald, the Editor for Communication and Coding Theory of the IEEE Communications Society. Manuscript received October 22, 2003; revised June 2, 2004 and October 10, 2004. This work was supported by the Army Research Office/Collaborative Technology Alliance (ARO/CTA) under Grant DAAD19-01-2-0011. This paper was presented in part at the IEEE Wireless Communication and Networking Conference, Atlanta, GA, March 2004.

The authors are with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455 USA (e-mail: wlzhao@ece.umn.edu; georgios@ece.umn.edu).

Digital Object Identifier 10.1109/TCOMM.2005.851590

estimate from [7]. Under the AWGN model, SE-SDA enables considerable complexity reduction.

B. SDA With Detection Ordering

SDA with detection ordering was introduced by Fincke and Pohst as a useful heuristic [6]. After rearranging columns of $\mathbf{H}$, the first lattice point examined by SE-SDA is the NC solution with received symbol-energy-based detection ordering, which appeared recently in [12]. Under a well-accepted random channel model, we provided a statistical justification in [17]. SDA with detection ordering proceeds as follows. Rearrange the columns of $\mathbf{H} := [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_N]$ to obtain $\mathbf{H}_o := [\mathbf{h}_{\sigma(1)}, \mathbf{h}_{\sigma(2)}, \ldots, \mathbf{h}_{\sigma(N)}]$, such that $\|\mathbf{h}_{\sigma(1)}\| \leq \|\mathbf{h}_{\sigma(2)}\| \leq \ldots \leq \|\mathbf{h}_{\sigma(N)}\|$. Permute the entries of $\mathbf{s}$ and $\hat{\mathbf{s}}$ accordingly; and then apply SDA. This ordering allows for considerable decoding speedup without bringing further complications to SDA.

C. SDA With IRS

A simple method to determine the search radius based on the AWGN model was introduced in [9], which is effective for the medium-to-high signal-to-noise ratio (SNR) regime. For a fixed search radius, there is always a probability that no candidate is found. Hence, increasing the radius is needed to achieve ML or near-ML performance while maintaining the SDA's efficiency. The SDA with IRS is as follows. Let $r_{P_1} < r_{P_2} < \ldots < r_{P_n}$ be a set of sphere radii. Execute the SDA with search radius r_{P_1}. If a candidate is found, terminate the program; otherwise, run SDA again with the next radius until r_{P_n}. This algorithm was initially mentioned in [14] without explicitly giving the set of radii, which was suggested in [10] as $\{r_{1-\epsilon}, r_{1-\epsilon^2}, \ldots\}$, where ϵ is a small number and r_x is determined by $\Pr(\|\mathbf{v}\|^2 \leq r_x) = x$. A closed-form expression for the average complexity of the SDA with IRS was derived in [10]. Here, we fix $\epsilon = 0.1$ in our simulations. An efficient SDA achieving *exact* ML error performance is the SE-SDA with detection ordering and IRS.

III. SDA With Improved IRS

By exploiting the AWGN model, IRS can improve the computational efficiency of the conventional SDA. However, there is an apparent waste of computations in the SDA with IRS. Specifically, for any sphere radius r_{P_i}, there is always a probability that this sphere does not contain any valid lattice point. When this happens, the SDA increases the search radius from r_{P_i} to $r_{P_{i+1}}$, and searches again. Computations in the search with radius r_{P_i} are discarded, but they are recalculated in the search with radius $r_{P_{i+1}}$. To reduce this loss and provide a mechanism to further lower search complexity, we will develop improved IRS (IIRS) algorithms.

The intuition behind the new IIRS is as follows. Whenever the SDA search with radius r_{P_i} fails, an incomplete search tree is constructed, from which promising paths can often be identified. Our IIRS exploits the valuable information on likely candidates conveyed by this partial tree. An incomplete tree for a four-dimensional (4-D) binary search is depicted in Fig. 1, where the initial radius is r_{P_1}. Each branch in the kth level of the tree is associated with a candidate of s_k. Starting from the root, each complete path corresponds to a candidate of $\mathbf{s}$, where the path metric is the sum of its branch metrics. From Fig. 1, it can be observed that paths 1 and 2 are more promising than path 3.

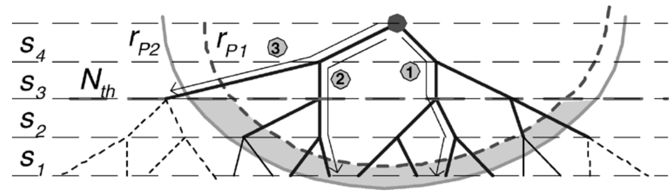

Fig. 1. Search tree of the SDA.

A. Ordering Promising Paths

To check paths efficiently, we will examine promising paths according to an ascending order of their predicted average Euclidean distance. When the SDA search with radius r_{P_i} fails, a path in an N-dimensional problem often consists of two segments. The first segment comprises n_1 branches, where branch metrics have been calculated. Let the sum of these branch metrics be $d_{n_1}^2 > r_{P_i}^2$. The parameters n_1 and $d_{n_1}^2$ are generated by the SDA search with r_{P_i}. For the second segment with n_2 branches, branch metrics remain unknown. It is clear that $N = n_1 + n_2$. We predict the average Euclidean distance of a path based on the parameters n_1 and $d_{n_1}^2$ next. Let us assume that promising paths correspond to either $\mathbf{s}$ or its immediate neighbors. Hence, path metrics are either $\|\mathbf{v}\|^2$ or $\|\mathbf{v} + \mathbf{h}\|^2$, where $\mathbf{h}$ is a column vector of $\mathbf{H}$, $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, and $\mathbf{h} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Given parameters n_1 and d_{n_1} of the first segment, we determine the probability of the null hypothesis that this partial path corresponds to $\mathbf{s}$, which is denoted by $P_o := \Pr(\|\mathbf{v}\|^2 | n_1, d_{n_1}^2)$. Similarly, the probability of the alternative hypothesis is denoted by $P_a := \Pr(\|\mathbf{v} + \mathbf{h}\|^2 | n_1, d_{n_1}^2)$. Due to the simplifying assumption, we have $P_o + P_a = 1$. Denote the segment of a vector $\mathbf{x}$ starting from index a to b as $\mathbf{x}_{a:b}$. Since $\|\mathbf{v}_{N-n_1+1:N}\|^2 \sim \chi_{n_1}^2(\sigma^2)$ and $\|(\mathbf{v} + \mathbf{h})_{N-n_1+1:N}\|^2 \sim \chi_{n_1}^2(1 + \sigma^2)$, it follows that:

$$\frac{P_o}{P_a} = \left(1 + \frac{1}{\sigma^2}\right)^{\frac{n_1}{2}} \exp\left(-\frac{d_{n_1}^2}{2\sigma^2(1+\sigma^2)}\right) = \exp\left(c_1 n_1 - c_2 d_{n_1}^2\right)$$

where $c_1 := \ln(1 + 1/\sigma^2)/2$ and $c_2 := 1/(2\sigma^2)/(1 + \sigma^2)$ need to be calculated only once for a certain SNR, and are introduced to reduce computational cost in implementation. It follows that $P_a = 1/(1 + \exp(c_1 n_1 - c_2 d_{n_1}^2))$. The average squared Euclidean distance d_{av}^2 of a path can be calculated as

$$d_{\text{av}}^2 = d_{n_1}^2 + n_2(P_a + \sigma^2). \tag{2}$$

Checking paths in an ascending order of their average distances maximizes the average probability to find the vector with minimum distance early.

B. Additional SDA Constraints

Ordering promising paths induces a probabilistic structure. However, keeping track of all paths results in an undesirable exponentially growing memory. Our ultimate goal is to design algorithms with linear memory and near-ML error performance. To reduce memory requirements, we rely on several additional constraints. First, we employ two sphere radii $r_{P_i} < r_{P_{i+1}}$ in a single SDA search. The radius r_{P_i} plays the role of search radius as in the conventional SDA, whereas $r_{P_{i+1}}$ upper bounds distance of promising paths when the SDA search with r_{P_i} fails. Second, we employ a threshold $N_{\rm th}$ to confine the search to promising paths. This divides paths in two categories: promising or unlikely. If a path satisfies $n_2 < N_{\rm th}$, it is a promising path, and is ordered according to its average distance calculated by (2); otherwise, it is unlikely, and is ignored. These constraints are illustrated in Fig. 1, where only paths falling in the shaded area are checked, whenever the SDA search with r_{P_1} returns no candidate.

To further clarify our design intention and specify design parameters, let $C_{1-\epsilon}$ be the complexity of the version of SDA discussed in Section II with search radius $r_{1-\epsilon}$. No exact analytical expression is available for $C_{1-\epsilon}$. The complexity of IRS can be written as $C_{1-\epsilon} + \epsilon C_{1-\epsilon^2} + \epsilon^2 C_{1-\epsilon^3} + \cdots$, where the first two terms $C_{1-\epsilon} + \epsilon C_{1-\epsilon^2}$ comprise most of the search complexity for a suitable choice of ϵ, say $\epsilon = 0.1$. Furthermore, the probability of finding the ML estimate with complexity $C_{1-\epsilon} + \epsilon C_{1-\epsilon^2}$ is $1-\epsilon^2$. Our goal here is to determine suitable $N_{\rm th}$, r_{P_1}, and r_{P_2} values such that the first improved SDA search has considerably less complexity than $C_{1-\epsilon} + \epsilon C_{1-\epsilon^2}$, yet its probability of success is approximately $1-\epsilon^2$. Due to the lack of analytical expressions for both complexity and probability of error, the optimal parameters are rather difficult, if not impossible, to specify. Here, we will pursue a suboptimal approach. As a first step, we set r_{P_2} to be $r_{1-\epsilon^2}$.

In communications, the dimension N is often an even integer $N = 2D$, where $D \in \mathbb{N}$. To simplify our analysis, we further assume that $N_{\rm th} = 2D_{\rm th}$ and $D_{\rm th} \in \mathbb{N}$. Letting $X := \|\mathbf{v}_{N_{\rm th}+1:N}\|^2$ and $Y := \|\mathbf{v}_{1:N_{\rm th}}\|^2$, we have $X + Y = \|\mathbf{v}\|^2$, and $X \sim \chi^2_{N-N_{\rm th}}(\sigma^2)$ is independent of $Y \sim \chi^2_{N_{\rm th}}(\sigma^2)$. For an improved SDA with parameters r_{P_1}, r_{P_2}, and $N_{\rm th}$, the conditional probability that the path corresponding to $\mathbf{s}$ is promising can be calculated as

$$\begin{aligned} P_c &:= \Pr\left(X \le r_{P_1}^2 | r_{P_1}^2 \le X + Y \le r_{P_2}^2\right) \\ &= \frac{1}{P_2 - P_1} \\ &\quad \times \sum_{k=0}^{D_{\rm th}-1} \sum_{l=0}^{k} \frac{(-1)^l k_1^{D-D_{\rm th}+l} \left[e^{-k_1} k_1^{k-l} - e^{-k_2} k_2^{k-l}\right]}{(k-l)!\, l!\, (D-D_{\rm th}-1)!\, (D-D_{\rm th}+l)} \end{aligned} \tag{3}$$

where $k_1 := r_{P_1}^2/(2\sigma^2)$ and $k_2 := r_{P_2}^2/(2\sigma^2)$ are actually independent of SNR, and we have used the fact that $f_{XY}(x,y) = f_X(x) f_Y(y)$. The probability that the first improved SDA search is successful can be approximated by $P_{\rm suc} \approx P_1 + P_c(P_2 - P_1)$, where we have assumed that the search is successful when there is a promising path with distance less than r_{P_2}. To guarantee that $P_{\rm suc}$ is close to P_2, we choose parameters $N_{\rm th}$ and r_{P_1} such that $P_c > 0.995$.

C. IIRS-A

Relying on path ordering and the additional constraints, we consider a modification to the original SDA with radius r_{P_i}. Based on the parameters r_{P_i}, $r_{P_{i+1}}$, and $N_{\rm th}$, SDA-A is as follows. Before any candidate is found within r_{P_i} during an SDA search, we calculate the average distance for each path satisfying $n_2 < N_{\rm th}$ and $r_{P_i} < d_{n_1} < r_{P_{i+1}}$, retaining only information about the most promising path. Whenever the SDA with r_{P_i} fails, SDA-A either provides information about the most promising path, or indicates that there is no such path. If the latter is true, or the actual distance of the most promising path is greater than $r_{P_{i+1}}$, SDA-A fails; otherwise, it returns the candidate corresponding to the most promising path. Except for the small overhead in finding the most promising path, SDA-A enjoys the same complexity as SDA with radius r_{P_i}. Replacing SDA in IRS with SDA-A results in IIRS-A, which occupies linear memory.

We compare the average decoding complexity of IIRS-A and IRS via Monte Carlo simulations. The average number of floating point operations per decoding $N_{\rm flop}$ is used to indicate decoding complexity. We will provide complexity-exponent plots, where the complexity exponent is defined by $\log_N(N_{\rm flop})$. Furthermore, we will consider a block fading channel, where many symbol vectors $\mathbf{s}$ are transmitted through the same channel realization $\mathbf{H}$. These symbol blocks share the same preprocessing steps that include detection ordering and QR decomposition of $\mathbf{H}$. Henceforth, we ignore preprocessing computations for both IIRS and IRS.

We illustrate the decoding complexity reduction and symbol-error performance loss of IIRS-A relative to IRS with an example. For IRS, we employ $\{r_{1-\epsilon}, r_{1-\epsilon^2}, r_{1-\epsilon^3}, \ldots\}$ as the set of search radii, where $\epsilon = 0.1$. Both detection ordering and SE ordering are employed. For IIRS-A, we use the set of radius pairs $\{(r_{P_1}, r_{1-\epsilon^2}), (r_{1-\epsilon^2}, r_{1-\epsilon^4}), (r_{1-\epsilon^4}, r_{1-\epsilon^6}), \ldots\}$, where r_{P_1} along with $N_{\rm th}$ are determined for the first SDA-A search, as described in Section III-B. If increasing the radius is necessary, the same $N_{\rm th}$ will be employed in all subsequent searches.

Example 1—Part 1: The IIRS-A for $M = N = 32$ is considered with a 4-pulse-amplitude modulation (PAM) constellation. For the first SDA-A search, we fix $N_{\rm th} = 20$, $r_{P_2} = r_{0.99}$, and minimize P_1 under the constraint $P_c > 0.995$. With *Mathematica*, we were able to determine a suitable P_1 value as $P_1 = 0.4$, where the corresponding $P_c = 0.9964$, and the parameters k_1 and k_2 in (3) are $k_1 = 14.67$ and $k_2 = 26.74$, respectively. Based on these parameters, we find that $P_{\rm suc} \approx 0.9879$. When we examine the first SDA-A of IIRS-A and the SDA of conventional IRS, the complexity reduction brought by IIRS-A becomes apparent. SDA-A searches with the radius $r_{0.4} = 29.34\sigma^2$, and succeeds with a probability near 0.99; whereas SDA uses a considerably larger radius $r_{0.9} = 42.58\sigma^2$, but succeeds with smaller probability 0.90. The ratio of average flops per decoding for IRS over IIRS-A is reported in the second row of Table I. It can be observed that considerable complexity reduction is achieved for all SNR values at the expense of 0.5 dB symbol-error rate (SER) degradation, as depicted in Fig. 2. Furthermore, we observe from the complexity exponent plot in Fig. 3 that both IIRS-A and IRS exhibit polynomial complexity. In fact, for the target SER $= 10^{-5}$, the decoding complexity is less than cubic, which is within the reach of current technology.

D. IIRS-B

By considering the most promising path only, our IIRS-A achieves a SER performance within 0.5 dB of an ML detector,

TABLE I
COMPLEXITY REDUCTION IN AVERAGE NUMBER OF FLOPS PER DECODING BROUGHT BY IIRS ALGORITHMS WHEN $M = N = 32$ AND 4-PAM CONSTELLATION

SNR (dB)	18	19	20	21	22
IRS / IIRS-A	1.7369	1.9550	2.1517	2.1730	2.0064
IRS / IIRS-B	1.7240	1.8287	1.8291	1.7667	1.6924
IRS / IIRS-C	1.7449	1.8361	1.8222	1.7319	1.6576

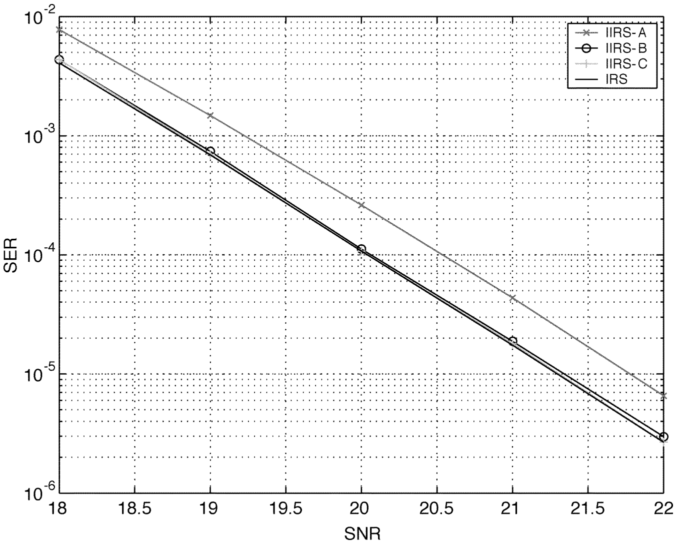

Fig. 2. SER comparison between IIRS algorithms and the conventional IRS for $M = N = 32$ and 4-PAM constellation.

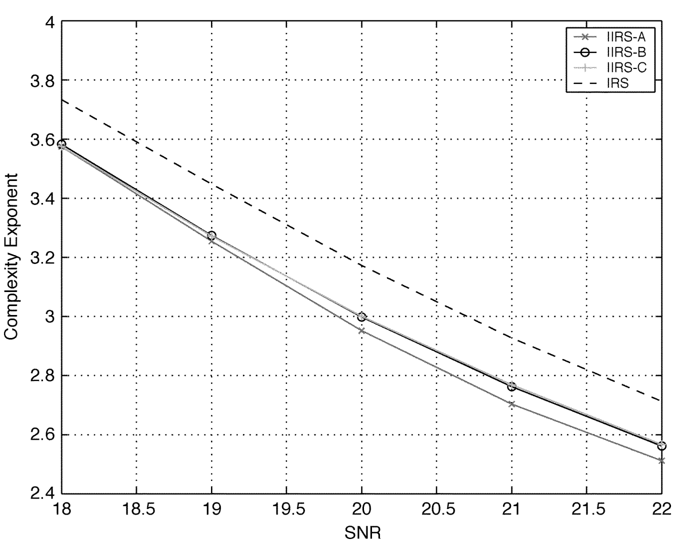

Fig. 3. Comparison of the average number of flops per decoding, in terms of complexity exponents between IIRS algorithms and the conventional IRS, for $M = N = 32$ and 4-PAM constellation.

which suggests that ML performance can be approached by exploiting a few most promising paths. To obtain near-ML performance, we develop SDA-B with parameters r_{P_i}, $r_{P_{i+1}}$, and $N_{\rm th}$ as follows. Before any candidate is found within r_{P_i} during an SDA search, we calculate the average distance for each path satisfying $n_2 < N_{\rm th}$ and $r_{P_i} < d_{n_1} < r_{P_{i+1}}$, and keep complete information about the three most promising paths. Whenever SDA with r_{P_i} fails, SDA-B provides more information on promising paths than SDA-A. If no such path exists, or no candidate within $r_{P_{i+1}}$ can be determined from these paths, SDA-B fails; otherwise, we test on the first candidate found with distance $d < r_{P_{i+1}}$. Suppose that n_2 and d_{n_1} are the parameters of the next promising path. If they are not available, we use $N_{\rm th}$ and r_{P_i} instead. The probability of a promising path with parameters n_2 and d_{n_1} to beat the current best distance d is $\Pr(Z < d^2 - d_{n_1}^2)$, where $Z := \|\mathbf{v}_{1:n_2}\|^2$ and $Z \sim \chi_{n_2}^2(\sigma^2)$. A threshold $P_{\rm th}$ can be chosen based on the performance gap between IIRS-A and conventional IRS. For a given $P_{\rm th}$ and each n_2, a threshold $Z_{\rm th}(n_2)$ can be determined by *Mathematica*. If $d^2 - d_{n_1}^2 < Z_{\rm th}(n_2)$, we decide that the current candidate is $\mathbf{s}$; otherwise, if we have information about the promising path under consideration, we calculate its distance with the upper bound d. If a better candidate is found, we update d. Perform the test again with d for the next path. If no candidate is found reliable, then SDA-B fails. Combining SDA-B with IRS results in our IIRS-B, which enjoys near-ML performance and also linear memory.

SDA-B offers two improvements over SDA-A. First, the three most promising paths are tracked by SDA-B, which enables near-ML performance with linear memory. Second, the testing is an effective mechanism to guarantee the reliability of a candidate. It is clear that this testing relies only on the AWGN model.

Example 1—Part 2: For the same system setting as in *Example 1—Part 1*, we compare SER and complexity of IIRS-B against the conventional IRS. To mitigate the 0.5 dB error-performance loss of IIRS-A, we use $P_{\rm th} = 0.1$. The SER performance is shown in Fig. 2, from which we observe that IIRS-B closely approaches the ML performance. The computational speedup is reported in the third row of Table I, and the complexity exponent is depicted in Fig. 3. Considerable complexity reduction is achieved by IIRS-B; while both IRS and IIRS-B entail affordable complexity.

E. Eliminating Channel-Model Dependency

The IIRS variants of SDA that we developed so far depend on the random model of $\mathbf{H}$ only through the ordering of promising paths. We will derive a new ordering here to eliminate this dependency. Under the assumption that a promising path with parameters n_1, n_2, and d_{n_1} corresponds to the transmitted vector $\mathbf{s}$, the average path distance can be calculated as $d_{\rm av}^2 = d_{n_1}^2 + n_2\sigma^2$. Ordering promising paths according to the new $d_{\rm av}$, we obtain a different ordering, which is independent of the statistical model of $\mathbf{H}$. Since there is only one path corresponding to $\mathbf{s}$, the predicted $d_{\rm av}$ is often much smaller than the actual path distance. Hence, the probability structure becomes less apparent than ordering according to (2). Nonetheless, the probability P_a in (2) is often near zero for a few very promising paths. For these leading paths, the probability order remains approximately invariant under both definitions of $d_{\rm av}$. Since our IIRS algorithms track only a few most promising paths, we expect IIRS to achieve similar SER under both path orderings. Furthermore, this new $d_{\rm av}$ is easier to compute than that in (2). Replacing the path ordering in IIRS-B with our new ordering mechanism results in IIRS-C.

Example 1—Part 3: We here compare SER and decoding complexity of the IIRS-C against the conventional IRS. Based on previous calculations, we set $N_{\text{th}} = 20$ and $P_{\text{th}} = 0.1$. The SER performance is shown in Fig. 2, from which it is evident that IIRS-C also closely approaches the ML performance. The ratio of the average flops per decoding is also reported in Table I. We can observe that considerable complexity reduction is achieved by IIRS-C without exploiting the random model of the channel matrix $\mathbf{H}$.

F. SER Degradation From ML Performance

Here, we analyze the error-probability loss of IIRS algorithms relative to ML. We employ $\{(r_{P_1}, r_{P_2}), (r_{P_2}, r_{P_3}), \ldots\}$ as the set of search radius pairs for IIRS. To simplify analysis, we assume the most promising path corresponds to either $\mathbf{s}$ or its immediate neighbor.

We examine the performance degradation of IIRS-A first, considering one SDA-A search with radius pair $(r_{P_i}, r_{P_{i+1}})$ and no candidate within distance r_{P_i}. For a noise realization $\mathbf{v}$, we determine the parameter n_1 such that $\|\mathbf{v}_{N-n_1+1:N}\|^2 > r_{P_i}^2$ and $\|\mathbf{v}_{N-n_1+2:N}\|^2 < r_{P_i}^2$. For this n_1, we define $X_1 := \|\mathbf{v}_{N-n_1+1:N}\|^2$ and $X_2 := \|\mathbf{v}_{1:n_2}\|^2$, where $n_1 + n_2 = N$. Given $N_1 = n_1$, X_1, and X_2 with $X_1 > r_{P_i}^2$ and $X_1 + X_2 < r_{P_{i+1}}^2$, we calculate the conditional error probability first, which corresponds to the event that an immediate neighbor of $\mathbf{s}$ looks more promising based on path-length prediction, yet its actual distance is larger than $\|\mathbf{v}\|$ but less than $r_{P_{i+1}}$. Given $\mathbf{v}$ and n_1, we define $Y_1 := \min\{Z_1, \ldots, Z_N\}$ and $Y_2 := \|(\mathbf{v}+\mathbf{h})_{1:n_2}\|$, where $Z_i := \|(\mathbf{v}+\mathbf{h}_i)_{N-n_1+1:N}\|^2$ and $\mathbf{h}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ are independent columns of $\mathbf{H}$. We determine the conditional pdf of the random variable Y_1 next. Conditioned on X_1, random variables $\{Z_i\}$ are independent and identically noncentral Chi-square distributed [13], with the cumulative distribution function (cdf) given by $F_{Z_i|X_1}(z_i) = 1 - \mathcal{Q}_{n_1/2}(\sqrt{x_1}, \sqrt{z_i})$, where $\mathcal{Q}_n(\alpha, \beta)$ denotes the nth-order generalized Marcum Q function. It follows by definition that Y_1 is the smallest order statistic of N noncentral Chi-square random variables, and the cdf of Y_1 is given by $F_{Y_1|X_1}(y_1) = 1 - [\mathcal{Q}_{n_1/2}(\sqrt{x_1}, \sqrt{y_1})]^N$ [2]. Similarly, Y_2 follows a noncentral Chi square distribution with cdf $F_{Y_2|X_2}(y_2) = 1 - \mathcal{Q}_{n_2/2}(\sqrt{x_2}, \sqrt{y_2})$, and Y_2 is independent from Y_1. For $n_1 > N - N_{\text{th}}$, the desired conditional symbol-error probability can be found as

$$\begin{aligned} P_{i,c} &= \Pr\Big(Y_1 < x_1; x_1 + x_2 < Y_1 \\ &\qquad + Y_2 < r_{P_{i+1}}^2 | X_1 = x_1, X_2 = x_2, N_1 = n_1\Big) \\ &= \int_0^{x_1} \Big[\mathcal{Q}_{\frac{n_2}{2}}(\sqrt{x_2}, \sqrt{x_1 + x_2 - y_1}) \\ &\qquad - \mathcal{Q}_{\frac{n_2}{2}}\left(\sqrt{x_2}, \sqrt{r_{P_{i+1}}^2 - y_1}\right)\Big] dF_{Y_1|X_1}(y_1) \end{aligned} \quad (4)$$

where the condition $Y_1 < x_1$ indicates that there is a candidate more promising than $\mathbf{s}$, and the condition $x_1 + x_2 < Y_1 + Y_2 < r_{P_{i+1}}^2$ insures that it does incur a symbol error. The reason that there is no ordering for Y_2 is that SDA-A keeps the most promising path only. It is known that the derivative of $\mathcal{Q}_n(\alpha, \beta)$ with respect to β is $\partial\mathcal{Q}_n(\alpha, \beta)/\partial\beta = -(\beta^n/\alpha^{n-1})\exp(-(\alpha^2+\beta^2)/2)I_{n-1}(\alpha\beta)$, where $I_n(\alpha)$ is the nth-order modified Bessel function of the first kind. Based on this equality, we can easily reduce (4) to a single integration with finite limits, which can be evaluated with *Matlab* or *Mathematica*. For $n_1 \le N - N_{\text{th}}$, the path corresponding to $\mathbf{v}$ is not promising, and $P_{i,c}$ can be similarly calculated as

$$\begin{aligned} P_{i,c} &= \Pr\Big(Y_1 < r_{P_i}^2; x_1 + x_2 < Y_1 \\ &\qquad + Y_2 < r_{P_{i+1}}^2 | X_1 = x_1, X_2 = x_2, N_1 = N - N_{\text{th}}\Big). \end{aligned} \quad (5)$$

The unconditional probability P_i is the error probability of SDA-A with the search radius pair $(r_{P_i}, r_{P_{i+1}})$, and is given by $P_i = \sum_{n_1=1}^{N} \int\int P_{i,c}(x_1, x_2, n_1) f_{X_1,X_2,N_1}(x_1, x_2, n_1) dx_1 dx_2$, where the analytical expression for $P_{i,c}(x_1, x_2, n_1)$ is given in (4) and (5), and the joint distribution $f_{X_1,X_2,N_1}(x_1, x_2, n_1) = g(f_{\mathbf{v}}(\mathbf{v}))$ is a mixed distribution that is rather difficult to derive analytically. Nonetheless, Monte Carlo integration can be employed to evaluate P_i by randomly generating a large amount of $\mathbf{v}$ vectors, determining x_1, x_2, and n_1 from these realizations, and averaging. Finally, the additional error probability of IIRS-A relative to ML can be approximated by $P_A \approx \sum_i P_i$, where the approximation accuracy increases with SNR.

Error-performance degradation of IIRS-B and IIRS-C are similar, and can be approximated by $P_B \approx P_C \approx P_{\text{th}} P_A$, where P_{th} is the probability threshold used in testing.

G. Practical Considerations

Example 2: We touch upon practical aspects of IRS and IIRS-C and further test their efficiency. We consider a system with $M = N = 64$ and 2-PAM constellation. IIRS-C is compared with SE-SDA with IRS and detection ordering. Based on the procedure described in Section III-B, we set $N_{\text{th}} = 34$ and $r_{P_1} = r_{0.2}$. The corresponding probability of success for the first search is $P_{\text{suc}} = 0.9876$. Hence, the set of IIRS-C search radius pair is $\{(r_{0.2}, r_{1-\epsilon^2}), (r_{1-\epsilon^2}, r_{1-\epsilon^4}), \ldots\}$. For the conventional IRS, we set the search radius as $\{r_{1-\epsilon}, r_{1-\epsilon^2}, r_{1-\epsilon^3}, \ldots\}$. In both cases, $\epsilon = 0.1$. It is clear that IIRS-C uses a much smaller radius $r_{0.2}$ in the first search. The SER performance of IIRS-C, IRS, and NC with detection order is plotted in Fig. 4.

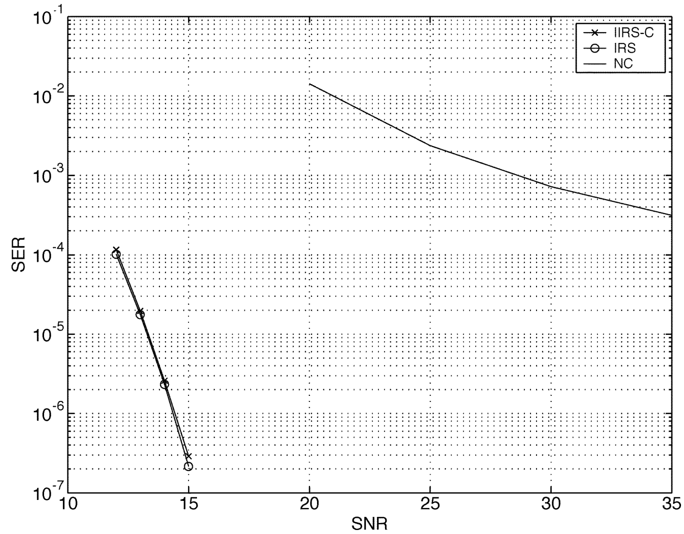

Fig. 4. SER comparison between IIRS-C, IRS, and NC for $M = N = 64$ and 2-PAM constellation.

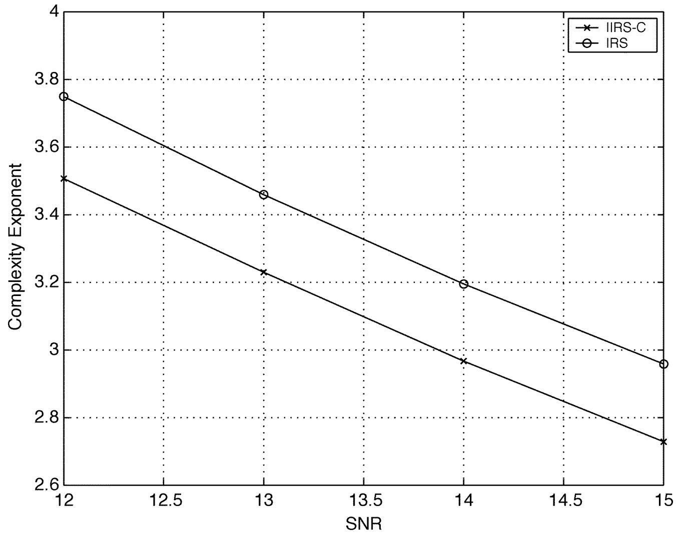

Fig. 5. Complexity exponent comparison between IIRS-C and IRS for $M = N = 64$ and 2-PAM constellation.

It can be observed that the SER of IIRS-C closely approaches the ML performance; whereas there is a huge performance gap between NC and ML performance. The complexity exponents of IRS and IIRS-C are depicted in Fig. 5. The decoding complexity for both algorithms decreases with increasing SNR. Nonetheless, the asymptotic complexity is $O(N^2)$, which is the complexity of NC. For SNR from 12 to 15 dB, the speedup for IIRS-C relative to IRS are 2.7352, 2.5954, 2.5856, and 2.5988, respectively. One key observation is that for those SNR values where NC with detection ordering provides satisfactory error performance, the complexity of the IRS or IIRS-C receiver is also rather low, yet these receivers offer considerable performance gain.

IV. Conclusions

Based on existing works, we have identified a relatively efficient ML-optimal SDA as the combination of SE-SDA with IRS and detection ordering. Using this version of SDA as a benchmark, we have developed IIRS algorithms to further reduce search complexity. Relying on the AWGN model, our novel IIRS-C algorithm can closely approach the ML error performance with considerably reduced complexity, and without dependence on the underlying channel model. We provided design guidelines, analytical parameter specifications, and error-degradation analysis. Simulations confirmed our design steps, and indicated that IIRS is effective for a wide range of dimensions N and SNR values.

References

[1] 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.

[2] N. Balakrishnan and A. C. Cohen, *Order Statistics and Inference Estimation Methods*. New York: Academic, 1991.

[3] L. Brunel and J. J. Boutros, "Lattice decoding for joint detection in direct-sequence CDMA systems," *IEEE Trans. Inf. Theory*, vol. 49, no. 4, pp. 1030–1037, Apr. 2003.

[4] A. Chan and I. Lee, "A new reduced-complexity sphere decoder for multiple antenna systems," in *Proc. Int. Conf. Commun.*, vol. 1, New York, NY, Apr. 28–May 2 2002, pp. 460–464.

[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. 2389–2402, Oct. 2003.

[6] U. Fincke and M. Pohst, "Improved methods for calculating vectors of short length in a lattice, including a complexity analysis," *Math. Comput.*, vol. 44, pp. 463–471, Apr. 1985.

[7] G. J. Foschini, "Layered space-time architecture for wireless communication in a fading environment when using multielement antennas," *Bell Labs Tech. J.*, vol. 1, no. 2, pp. 41–49, Autumn 1996.

[8] R. Gowaikar and B. Hassibi, "Efficient near-ML decoding via statistical pruning," in *Proc. Int. Symp. Inf. Theory*, Jul. 2003, p. 274.

[9] B. Hassibi and H. Vikalo, "On the expected complexity of sphere decoding," in *Proc. 35th Asilomar Conf. Signals, Syst., Computers*, Pacific Grove, CA, Nov. 2001, pp. 1051–1055.

[10] On the Expected Complexity of Sphere Decoding I. Theory, B. Hassibi and H. Vikalo. [Online]. Available: http://www.its.cal-tech.edu/hvikalo/publications.html

[11] B. M. Hochwald and S. ten Brink, "Achieving near-capacity on a multiple-antenna channel," *IEEE Trans. Commun*, vol. 51, no. 3, pp. 389–399, Mar. 2003.

[12] J. Luo, K. R. Pattipati, P. Willett, and G. M. Levchuk, "Fast optimal and suboptimal anytime algorithms for CDMA multiuser detection based on branch-and-bound," *IEEE Trans. Commun.*, vol. 52, no. 4, pp. 632–642, Apr. 2004.

[13] J. G. Proakis, *Digital Communications*, 4th ed. Englewood Cliffs, NJ: McGraw-Hill, 2001.

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

[15] R. Wang and G. B. Giannakis, "Approaching MIMO capacity with reduced-complexity soft sphere-decoding," in *Proc. Wireless Commun. Netw. Conf.*, vol. 3, Atlanta, GA, Mar. 2004, pp. 1620–1625.

[16] Y. Xin, Z. Wang, and G. B. Giannakis, "Space–time diversity systems based on linear constellation precoding," *IEEE Trans. Wireless Commun.*, vol. 2, no. 2, pp. 294–309, Mar. 2003.

[17] W. Zhao and G. B. Giannakis, "Reduced-complexity closest-point algorithms for random lattices," *IEEE Trans. Wireless Commun.*, to be published.