Multi-Resolution Multicasting Over the Grassmann and Stiefel Manifolds

Karim G. Seddik, *Senior Member, IEEE*, Ramy H. Gohary, *Senior Member, IEEE*, Mohammad T. Hussien, *Student Member, IEEE*, Mohammad Shaqfeh, *Member, IEEE*, Hussein Alnuweiri, *Senior Member, IEEE*, and Halim Yanikomeroglu, *Fellow, IEEE*

***Abstract*— We consider the design of space-time codes for the multiple-input multiple-output multicast communication systems with two classes of receivers. The first class comprises high-resolution (HR) receivers which have access to reliable channel state information (CSI) and can perform coherent detection, and the second class comprises low-resolution (LR) receivers which do not have access to CSI and can only perform non-coherent detection. We propose a layered encoding structure in which LR information available to both classes of receivers is encoded using Grassmannian constellations, and an incremental component, which is available only to the HR receivers, is encoded in the particular bases of the transmitted Grassmannian constellation points, thereby giving rise to constellations on the Stiefel manifold. The proposed structure enables reliable coherent communication of the HR information without compromising the reliability with which the basic LR information is non-coherently communicated. To effect rate-efficient communication of the incremental, HR layer, we use optimization methods on the Stiefel manifold to develop a novel technique for designing the unitary constellations directly. This approach alleviates the restriction imposed by the traditional techniques in which unitary space-time codes are constructed from scalar constellations. As such, this approach enables better control of the distance spectrum of the developed constellations and more effective utilization of the degrees of freedom that underlie the Stiefel manifold. For the LR receivers, we use maximum likelihood detection, whereas for the HR receivers, we develop a computationally-efficient two-step sequential detector which detects the LR information prior to detecting the incremental component superimposed on it. The detectors and the layered structure with the aforementioned constellations enable full diversity and maximum degrees of freedom to be achieved on the Grassmann and Stiefel manifolds.**

***Index Terms*— Non-coherent Communications, Layered Coding, Grassmannian Codes, Multicasting, MIMO.**

I. Introduction

THE multiple-input multiple-output (MIMO) architecture and its massive variant are likely to prevail in future wireless networks that aim at achieving high spectral efficiencies within tightly restricted bandwidths [1]–[6]. Effective utilization of this architecture depends on the accuracy of the available channel state information (CSI). For instance, when reliable CSI is available at the receiver, the communication channel operates in a coherent mode, whereas when no CSI is available at the receiver, the channel operates in a non-coherent mode [7]–[10]. The coherent mode usually arises in low-mobility communication scenarios in which the channel exhibits slow variations and the resources expended by the transmitter to send pilot symbols to assist the receiver in acquiring accurate CSI are negligible. In contrast with its coherent counterpart, the non-coherent mode arises in communication scenarios with fast, and possibly abrupt, channel variations. Such situations arise in high-mobility communication scenarios, wherein a user might be served by various base stations during consecutive time blocks. In those scenarios, the resources expended in order for the receiver to acquire accurate CSI cannot be ignored and their effect must be taken into consideration.

In addition to the point-to-point case, another communication scenario that arises in practice is the multicast one. Such a scenario was considered in the context of beamforming and transmit optimization in e.g., [11]–[14]. In this scenario, one transmitter wishes to send a common message to multiple receivers. However, the receivers might experience disparate channel conditions; some receivers might be able to acquire accurate CSI and subsequently operate in the coherent mode, other receivers may not have this luxury and are therefore restricted to operate in the non-coherent mode [15]–[17]. Such scenarios arise in digital video broadcasting and in the prospective paradigms of Internet-of-Things and vehicle-to-vehicle (V2V) communications. For instance, in V2V communications, vehicles traveling at small relative velocities can communicate more detailed information than those traveling at higher ones. Multicasting schemes for such scenarios have been considered in [18] and [19] but without a non-coherent component.

Manuscript received December 10, 2016; revised March 24, 2017; accepted May 13, 2017. Date of publication May 29, 2017; date of current version August 10, 2017. The work of K. G. Seddik and M. T. Hussien was supported by NPRP under Grant # 05-401-2-161 from the Qatar National Research Fund (a member of Qatar Foundation). The work of the M. Shaqfeh and H. Alnuweiri was supported by NPRP under Grant # 8-1531-2-651 from the Qatar National Research Fund (a member of Qatar Foundation). The work of R. H. Gohary and H. Yanikomeroglu was supported in part by Huawei Canada Co., Ltd., and in part by the Ontario Ministry of Economic Development and Innovations ORF-RE program. This work was presented at the 2014 IEEE International Symposium on Information Theory and the 2015 IEEE Global Communications Conference. The associate editor coordinating the review of this paper and approving it for publication was W. H. Mow. *(Corresponding author: Karim G. Seddik.)*

K. G. Seddik is with the Electronics and Communications Engineering Department, The American University in Cairo, New Cairo 11835, Egypt (e-kseddik@aucegypt.edu).

R. H. Gohary and H. Yanikomeroglu are with the Department of Systems and Computer Engineering, Carleton University, Ottawa, ON K1S 5B6, Canada.

M. T. Hussien was with the Department of Electrical Engineering, Alexandria University, Alexandria 21544, Egypt. He is now with the School of Electrical Engineering and Computer Science, Pennsylvania State University, State College, PA 16802 USA.

M. Shaqfeh and H. Alnuweiri are with the Department of Electrical Engineering, Texas A&M University at Qatar, Doha, Qatar.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TWC.2017.2707549

In contrast with [18] and [19], in the framework considered herein, the transmitter adopts a layered architecture, whereby information to be transmitted is partitioned into basic low-resolution (LR) information that is communicated to all receivers and incremental high-resolution (HR) information that is communicated only to those privileged receivers that have access to accurate CSI.[1] Comparisons between the cases of coherent and non-coherent MIMO communications were considered in [15] and [16], but without considering the simultaneous multicasting framework considered herein. The focus in [15] and [16] was on investigating the communications degrees of freedom, rather than on providing explicit constructions that achieve them. In contrast, herein we provide explicit designs that enable space-time block codes to operate in a multilayer (multi-resolution) multicasting setup, first introduced in [17]. We analyze the degrees of freedom achieved by these designs and provide performance comparisons between them.

Communicating effectively and simultaneously to HR coherent and LR non-coherent receivers constitutes conflicting goals. To see that, we note that the capacity-achieving signals for the coherent receivers are in the form of Gaussian distributed signals [20], which for practical considerations, are usually replaced by unitarily-structured ones [2]. In contrast, capacity-achieving signals for non-coherent receivers possess an intricate structure, which constitutes the product of (tall) unitary matrices and diagonal matrices with non-negative entries [21]–[23]. This structure reduces to unitary matrices that span distinct subspaces when the signal-to-noise ratio (SNR) is asymptotically high and the coherence time of the channel, normalized to the symbol duration, equals at least the sum of the number of transmit and receive antennas. The set of such unitary matrices is known as the Grassmann manifold. This manifold possesses desirable properties which will enable us to mitigate the conflict arising from the fundamental difference between the signals that achieve capacity in coherent and non-coherent scenarios. The key feature of the Grassmann manifold is that each point on it can be represented by an equivalence class of unitary matrices, and each such class contains those unitary matrices that span the same subspace.

The design of high-SNR capacity-achieving non-coherent codes is equivalent to subspace packing on the Grassmannian manifold. See, e.g., [24] for packings on real Grassmannian manifolds and [7], [25]–[29] for several packing techniques on the complex Grassmann manifolds.

In this paper, we use capacity-achieving Grassmannian constellations to communicate the LR information to the non-coherent receivers, and square unitary constellations to communicate the HR information to the coherent receivers.[2] Simultaneous communication of both types of information is achieved by right-multiplying points from the Grassmannian constellation with points from the unitary one. Such a multiplication rotates the bases, but preserves the subspace spanned by the Grassmannian constellation point. With this construction, the resulting unitary matrices represent points on the Stiefel manifold, and a coherent receiver will be able to detect the HR information encoded by points on this manifold, whereas a non-coherent receiver will only be able to detect the LR information encoded by points on the underlying Grassmann manifold. Restricting the matrices used to communicate the HR information to be unitary does not achieve the coherent capacity but enables the non-coherent one to be achieved.

Effective realization of the proposed multi-resolution signalling scheme requires the design of Grassmannian and unitary constellations with favourable distance properties. Several approaches for designing arbitrary Grassmannian constellations are available, see e.g., [7], [8]. However, square unitary constellations are only available for specific cardinalities [30] and dimensions [31]. To overcome this difficulty, in this paper we use generic optimization methods [32], [33] to design unitarily-constrained constellations directly using a gradient descent algorithm over the group of square unitary matrices. To do so, we obtain expressions for the design objectives, the gradients and their projections on the unitary group. We will show that the resulting constellations together with the Grassmannian ones achieve the maximum degrees of freedom for the coherent HR channels with unitarily-constrained input signals. Finally, we develop a two-step detector, which, in addition to its computational efficiency, enables us to show that the proposed layered structure allows full diversity to be achieved by both the LR and HR receivers.

Our contributions are summarized in the following points.

- We proposed the only available method that uses the unitary invariance of the Grassmann manifold for simultaneous multicasting of multi-resolution information to two classes of receivers, coherent and non-coherent ones.
- We developed a generic approach of designing square unitary matrices with arbitrary cardinalities and dimensions. This approach yields constellations with distance properties that are more favorable than those of their existing counterparts.
- We developed a novel two-step detector that achieves the same diversity order as its optimal one-step counterpart but with a significantly less computational complexity.

The paper is organized as follows. Section II presents background material, the system model and the proposed signaling structure. In Section III, we develop a computationally-efficient two-step coherent detector that achieves full diversity. The degrees of freedom achieved by the proposed signaling structure are analyzed in this section. Design techniques for constellations on the group of unitary matrices are presented in Section IV. Simulation results are presented in Section V and conclusions are discussed in Section VI.

II. Preliminaries and System Model

In this section we will present preliminaries and the system model considered in the paper.

[1] Whereas all receivers are assumed to know the channel statistical model perfectly, those that operate non-coherently are assumed to have no information about the instantaneous channel gains.

[2] The HR layer refers to the layer containing incremental information; the HR receivers will decode the LR and HR layers.

A. Preliminaries

We begin by providing definitions needed for describing the proposed signaling scheme. (See e.g., [34] for more background on the unitary group, the Grassmann and the Stiefel manifolds.),

Definition 1: The unitary group, $\mathbb{U}_M$, is the set of square matrices in $\mathbb{C}$, where $\mathbb{C}$ is the field of complex numbers, which have orthonormal columns and rows. In particular,

$$\mathbb{U}_M = \{\mathbf{U} \in \mathbb{C}^{M\times M} | \mathbf{U}^\dagger\mathbf{U} = \mathbf{U}\mathbf{U}^\dagger = \mathbf{I}_M\},$$

where $\mathbf{I}_M$ is the $M \times M$ identity matrix. □

The product of two elements of $\mathbb{U}_M$ is an element of $\mathbb{U}_M$. The number of real dimensions that are spanned by the elements of $\mathbb{U}_M$, $\dim(\mathbb{U}_M) = M^2$.

A set that is closely related to the unitary group is the Stiefel manifold, which is defined next.

Definition 2: The Stiefel manifold, $\mathbb{S}_{T,M} \subset \mathbb{C}^{T\times M}$, is the set of "tall" unitary matrices, i.e.,

$$\mathbb{S}_{T,M} = \{\mathbf{Q} \in \mathbb{C}^{T\times M} | \mathbf{Q}^\dagger\mathbf{Q} = \mathbf{I}_M\}, \qquad T \geq M.$$

□

Since $T \geq M$, the Stiefel manifold subsumes the unitary group wherein $T = M$. The number of real dimensions spanned by the elements of $\mathbb{S}_{T,M}$ is given by $\dim(\mathbb{S}_{T,M}) = 2TM - M^2$.

Left and right multiplication of an element of $\mathbb{S}_{T,M}$ with elements of $\mathbb{U}_T$ and $\mathbb{U}_M$, respectively, yields an element of $\mathbb{S}_{T,M}$. We highlight a key difference between left and right multiplications. When $T > M$, the columns of a matrix $\mathbf{Q} \in \mathbb{S}_{T,M}$ represent the basis of a particular M-dimensional subspace in $\mathbb{C}^T$, say $\mathcal{S}_Q$. It can be readily verified that right multiplication of $\mathbf{Q}$ by an element of $\mathbb{U}_M$ yields another basis for the same subspace, i.e., for any $\mathbf{U} \in \mathbb{U}_M$, $\mathcal{S}_{QU} = \mathcal{S}_Q$. In contrast, left multiplication of $\mathbf{Q}$ by an element of $\mathbb{U}_T$ yields an element of $\mathbb{S}_{T,M}$ that may span a subspace other than the one spanned by $\mathbf{Q}$. The set of all distinct M-dimensional subspaces in a T-dimensional subspaces is usually referred to as the Grassmann manifold, which we denote by $\mathbb{G}_{T,M}$. The construction of this manifold implies the following definition.

Definition 3: The Grassmann manifold is the quotient space $\mathbb{S}_{T,M}/\mathbb{U}_M$, $T \geq M$. In particular, let $[\mathbf{Q}] = \{\mathbf{QP} | \mathbf{P} \in \mathbb{U}_M\}$ be the equivalence class of $\mathbf{Q}$ containing the elements of $\mathbb{S}_{T,M}$ that span the same M-dimensional subspace. Then, the Grassmann manifold can be defined as

$$\mathbb{G}_{T,M} = \{[\mathbf{Q}] | \mathbf{Q} \in \mathbb{S}_{T,M}\}.$$

□

From this definition, it can be seen that the Grassmann manifold is the disjoint union of matrices that span distinct subspaces. Hence, the number of real dimensions that it spans is

$$\dim(\mathbb{G}_{T,M}) = \dim(\mathbb{S}_{T,M}) - \dim(\mathbb{U}_M) = 2M(T - M). \quad (1)$$

Since for each M-dimensional subspace in $\mathbb{C}^T$, there is a unique $(T - M)$-dimensional orthogonal subspace associated with it, it follows that it suffices to restrict attention to the case of $M \leq \lfloor \frac{T}{2} \rfloor$ when considering the Grassmann manifold. To see that, we note that the M-dimensional subspace can be described either by M T-dimensional linearly-independent vectors that span it, which gives rise to a $T \times M$ matrix, or by $(T-M)$ T-dimensional linearly-independent vectors that span its null space, which gives rise to a $(T-M)\times M$ matrix. Since both matrices represent the same geometric object, it suffices to consider the one with fewer columns, i.e., the one with $\min\{M, T-M\}$ columns. Now, the value of M that maximizes $\min\{M, T-M\}$ is $M = \lfloor T/2 \rfloor$. Hence, it suffices to consider values of $M \leq \lfloor \frac{T}{2} \rfloor$.

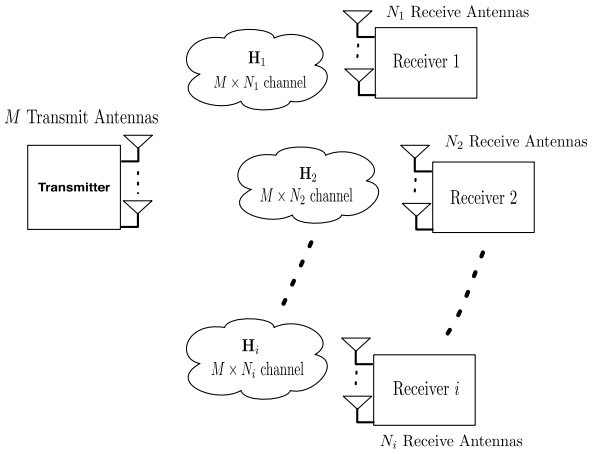


Fig. 1. The multi-resolution MIMO multicast system model ($N_i \geq M$, $\forall i$).

The aforementioned geometric objects will facilitate the description of the proposed signaling scheme. Towards that end, we will make the following definition.

Definition 4: Let $\mathbf{X}$ be a random matrix in $\mathbb{C}^{T\times M}$ with probability density function (pdf), $p(\mathbf{X})$. Then, $\mathbf{X}$ is isotropically distributed if, for any matrix $\mathbf{P} \in \mathbb{U}_T$, $p(\mathbf{X}) = p(\mathbf{PX})$. □

Having provided the mathematical definitions that will be necessary for subsequent developments, we will now describe the model of the system considered in this paper.

B. System Model

We consider a multicast MIMO communication system with one transmitter and two classes of receivers operating over block Rayleigh flat-fading channels as shown in Fig. 1. The number of antennas at the transmitter is denoted by M, and the number of antennas at receiver i is denoted by N_i. The channels are assumed to be constant over a coherence interval of T consecutive time slots. Using this model, the $T \times N_i$ received matrix of the i-th receiver can be expressed as

$$\mathbf{Y}_i = \mathbf{X}\mathbf{H}_i + \sqrt{\frac{M}{\rho T}}\mathbf{W}_i, \qquad i \in \mathcal{N}_C \cup \mathcal{N}_{NC}, \quad (2)$$

where $\mathcal{N}_C$ and $\mathcal{N}_{NC}$ denote the set of coherent receivers and the set of non-coherent receivers, respectively, ρ is the SNR, which is assumed identical for all users,[3] $\mathbf{X} \in \mathbb{C}^{T\times M}$ is the transmitted signal matrix, which is assumed to satisfy

[3]In practice, users within the $\mathcal{N}_C$ and $\mathcal{N}_{NC}$ sets will observe different SNRs, and the transmission rate of either the LR or the HR information will be limited by the lowest SNR within each set.

the power constraint $\mathrm{Tr}(\mathbf{X}\mathbf{X}^{\mathcal{H}}) = M$, and $\mathbf{H}_i \in \mathbb{C}^{M\times N_i}$ and $\mathbf{W}_i \in \mathbb{C}^{T\times N_i}$ are the channel and noise matrices of receiver i, $i \in \mathcal{N}_C \cup \mathcal{N}_{NC}$. The elements of $\{\mathbf{H}_i\}$ and $\{\mathbf{W}_i\}$ are assumed to be statistically independent identically distributed (i. i. d.) circularly-symmetric zero mean unit variance complex Gaussian random variables. Throughout, the number of receive antennas N_i of any receiver $i \in \mathcal{N}_C \cup \mathcal{N}_{NC}$ is assumed to satisfy $N_i \geq M$. For notational convenience, we will drop the index i.

Our objective is to develop a layered structure that enables $\mathbf{X}$ to communicate two messages, an LR one containing the basic information intended for all receivers including those without CSI, and an incremental HR message intended for receivers with reliable CSI.

When the SNR is sufficiently high, the capacity of the LR channel of any receiver can be achieved if $\mathbf{X}$ is isotropically distributed on $\mathbb{G}_{T,M}(\mathbb{C})$, $N \geq M$, $T \geq M+N$ and $M \leq \lfloor T/2 \rfloor$ [23]. Henceforth, these conditions will be assumed to be satisfied.

To comprise the LR and HR components, $\mathbf{X}$ is structured as follows:

$$\mathbf{X} = \mathbf{U}\mathbf{A}, \tag{3}$$

where $\mathbf{U}$ is isotropically distributed on $\mathbb{G}_{T,M}(\mathbb{C})$ and $\mathbf{A} \in \mathbb{C}^{M\times M}$ is isotropically distributed on $\mathbb{U}_M$. Since $\mathbb{G}_{T,M}(\mathbb{C})$ is invariant under right multiplication by square unitary matrices, cf. Section II-A, it can be readily seen that the proposed structure ensures that $\mathbf{X}$ is isotropically distributed on $\mathbb{G}_{T,M}(\mathbb{C})$, and is therefore guaranteed to achieve the high-SNR ergodic capacity of the LR channel. Since each point on $\mathbb{G}_{T,M}(\mathbb{C})$ represents an M-dimensional subspace, the LR information can be communicated by mapping it to the subspaces spanned by the matrix $\mathbf{U}$. This subspace can be recovered non-coherently by the LR receivers which do not have access to any CSI, or coherently by the HR receivers which have access to reliable CSI.

As for the HR information, we note that the structure proposed for $\mathbf{X}$ does not endow it with the Gaussian distribution that achieves the capacity of the HR channel. However, this structure has various practical advantages including admitting a computationally efficient detection technique that will be described in Section III-A.2. To communicate the HR information, this information is mapped to the matrix $\mathbf{A}$, which represents the particular basis of the subspace spanned by $\mathbf{U}$ and can be recovered by the HR receivers only, but not by the LR ones.

We will later show that, in addition to admitting efficient detection, restricting $\mathbf{A}$ in (3), which represents the incremental HR information, to possess a unitary structure preserves the distance characteristics of the Grassmannian non-coherent constellation. This implies that the performance of the LR (non-coherent) constellation will not be affected by the transmission of the HR information. Also, restricting $\mathbf{A}$ to be unitary does not increase the required transmitted power. This implies that the HR (coherent) layer is transparent to the LR (non-coherent) layer and no additional power provisioning is required to maintain the performance of the LR layer.

III. Diversity and Degrees-of-Freedom Analysis of the Layered Framework

A. Detectors

In this section, we will present coherent and non-coherent detectors for receivers with and without reliable CSI, respectively. For receivers without CSI, the incremental HR information will be lost and only the LR information can be detected. In contrast, for receivers with reliable CSI, both the LR and the HR information can be detected, and the optimal detector in this case is the conventional maximum likelihood (ML) one. Despite its optimality, the ML detector is computationally expensive to implement in practice. To circumvent this difficulty, we will develop a computationally-efficient suboptimum detector, and we will show that this detector achieves the same diversity order as the optimal ML one.

1) The Optimum Non-Coherent Detector: For the LR receivers, no CSI is available and the subspace spanned by $\mathbf{X}$ can be detected using the generalized likelihood ratio test (GLRT) detector [35]. In this detector, the supremum of the likelihood function over the unknown matrix $\mathbf{A}\mathbf{H}$ is computed and the symbol that maximizes this supremum is decided in favour of [36]. Performing the supremum operation yields that the GLRT detector will decide in favour of[4]

$$\hat{\mathbf{U}} = \arg\max_{\mathbf{U}\in\mathcal{C}_L} \mathrm{Trace}(\mathbf{Y}^{\mathcal{H}}\mathbf{U}\mathbf{U}^{\mathcal{H}}\mathbf{Y}). \tag{4}$$

From (4) it can be readily seen that the GLRT detector is identical to the ML one [35], and that encoding the HR information in the unitary matrix $\mathbf{A} \in \mathbb{U}_M$ does not compromise the performance of the non-coherent GLRT detector. In particular, we can write $\mathbf{H} \overset{d}{=} \mathbf{A}\mathbf{H}$, where $\overset{d}{=}$ denotes equality in distribution. This implies that the encoded HR information will 'see' an equivalent channel matrix $\mathbf{A}\mathbf{H}$ with the same statistics as the original channel matrix $\mathbf{H}$. As such, it can be readily seen that the GLRT detector will exhibit the same performance as if the HR information layer were not present. This fact can be also inferred from the upper bound on the pairwise error probability (PEP) [22] of the standard single layer non-coherent channel with unitary input matrices. This bound asserts that $\mathrm{PEP}(\mathbf{U}_1 \to \mathbf{U}_2)$, i.e., the probability of mistaking the transmitted signal $\mathbf{U}_1$ for $\mathbf{U}_2$ satisfies

$$\mathrm{PEP}(\mathbf{U}_1 \to \mathbf{U}_2) \leq \frac{1}{2}\prod_{m=1}^{M}\left[1+\frac{(\rho T/M)^2(1-s_m^2)}{4(1+\rho T/M)}\right]^{-N}, \tag{5}$$

where $1 \geq s_1 \geq \cdots \geq s_M \geq 0$ are the singular values of the $M\times M$ matrix $\mathbf{X}_2^{\mathcal{H}}\mathbf{X}_1$, which are equal to the singular values of the matrix $\mathbf{U}_2^{\mathcal{H}}\mathbf{U}_1$, irrespective of $\mathbf{A}$. By properly designing the Grassmannian signaling matrices of the LR non-coherent constellation, $\mathcal{C}_L = \{\mathbf{U}_i\}$, the greatest singular value of $\mathbf{U}_2^{\mathcal{H}}\mathbf{U}_1$, s_1, must be guaranteed to be strictly less than 1 for any two distinct matrices $\mathbf{U}_1, \mathbf{U}_2 \in \mathbb{G}_{T,M}(\mathbb{C})$. In such a case, it can be readily verified that the asymptotic SNR exponent

[4] A reduced-search detector that yields a potentially suboptimal solution was developed in [7]. This detector uses the QR-decomposition to limit the search for the transmitted Grassmannian symbol to the neighborhood of the received signal subspace.

equals $-MN$, which ensures that the ML detection of the LR layer achieves full diversity order regardless of whether the HR information is transmitted or not.

2) Coherent One-Step and Two-Step Detectors:

a) The Optimum One-Step Coherent Detector: The optimum coherent detector that "jointly" decodes the LR and HR information layers can be expressed as the following ML detector

$$\hat{\mathbf{X}} = \arg \max_{\mathbf{X} \in \mathcal{C}_L \times \mathcal{C}_H} p(\mathbf{Y}|\mathbf{X}, \mathbf{H}), \tag{6}$$

which, using the fact that the noise is additive and Gaussian-distributed, can be expressed as

$$\hat{\mathbf{X}} = \arg \min_{\mathbf{X} \in \mathcal{C}_L \times \mathcal{C}_H} \|\mathbf{Y} - \mathbf{X}\mathbf{H}\|^2. \tag{7}$$

The pairwise error probability (PEP) between two distinct data symbols, $\mathbf{X}_1$ and $\mathbf{X}_2$, of this one-step coherent detector can be upper bounded as follows [1]

$$\text{PEP}(\mathbf{X}_1 \to \mathbf{X}_2) \leq \frac{1}{2} \prod_{m=1}^{M} \left[1 + \frac{\rho T(1-\lambda_m)}{4M}\right]^{-N}, \tag{8}$$

where $\lambda_m \in [-1, 1]$ is the m-th eigenvalue of the matrix $\frac{\mathbf{X}_2^{\mathcal{H}}\mathbf{X}_1 + \mathbf{X}_1^{\mathcal{H}}\mathbf{X}_2}{2}$.

From (8), it can be readily verified that, similar to the non-coherent case, an appropriate design of the HR coherent constellation, $\mathcal{C}_H = \{\mathbf{A}_i\}$, must ensure that for any two distinct matrices, $\mathbf{X}_1$ and $\mathbf{X}_2$, none of the eigenvalues of $\frac{\mathbf{X}_2^{\mathcal{H}}\mathbf{X}_1 + \mathbf{X}_1^{\mathcal{H}}\mathbf{X}_2}{2}$ is 1. In that case, full diversity is achieved and the asymptotic SNR exponent equals $-MN$.

Unfortunately, the detector in (7) requires an exhaustive search over $|\mathcal{C}_L||\mathcal{C}_H|$ constellation points.[5] This detector is computationally expensive to implement in practice and to alleviate this difficulty, we will present a two-step detector, which is less computationally demanding than the one-step detector and achieves full diversity, i.e., yields an SNR exponent of $-MN$.

b) The Two-Step Coherent Detector: To mitigate the computational cost of the one-step detector in Section III-A.2.a, we now develop a sequential two-step detector that is significantly less complex and achieves full diversity. In the first step of this detector, the GLRT in (4) is used to detect the LR Grassmannian matrix. Let this matrix be denoted by $\hat{\mathbf{U}}$. Assuming that $\hat{\mathbf{U}}$ was actually transmitted, in the second step the detector uses $\hat{\mathbf{U}}$ to perform ML detection of the HR information in $\mathbf{A}$. In particular, the output of this ML detector is given by

$$\hat{\mathbf{A}} = \arg \min_{\mathbf{A} \in \mathcal{C}_H} \|\mathbf{Y} - \mathbf{U}\hat{A}H\|^2. \tag{9}$$

To see that this detector is significantly less complex than the one-step detector in (7), we note that it requires searching over $|\mathcal{C}_L| + |\mathcal{C}_H|$ matrices, as opposed to the $|\mathcal{C}_L||\mathcal{C}_H|$ matrices that are searched over in (7). From a performance perspective, the two-step detector does not take advantage of the available CSI when detecting the LR information in $\mathbf{U}$. This results in a performance degradation in comparison with one-step detector. However, as the following theorem shows, both detectors yield the same diversity order.

[5] $|\mathcal{T}|$ denotes the cardinality of the set $\mathcal{T}$.

Theorem 1: Let the LR constellation $\mathcal{C}_L = \{\mathbf{U}_i\}$ satisfy the full diversity singular values criterion for non-coherent codes, whereby $\max_m\{s_m(\mathbf{U}_i^{\mathcal{H}}\mathbf{U}_j)\} < 1$, for every $\mathbf{U}_i, \mathbf{U}_j \in \mathcal{C}_L$, $i \neq j$, where $s_m(\cdot)$ is the m-th singular value of the matrix argument. Let the HR constellation $\{\mathbf{A}_i\}$ satisfy the full diversity criterion $\prod_{m=1}^{M}(1-\sigma_m) > 0$, where σ_m is the m-th eigenvalue of $\frac{\mathbf{A}_\ell^{\mathcal{H}}\mathbf{A}_k + \mathbf{A}_k^{\mathcal{H}}\mathbf{A}_\ell}{2}$, for every $\mathbf{A}_\ell, \mathbf{A}_k$, $\ell \neq k$ (cf. (8)). Then, the sequential two-step coherent detector achieves the same diversity order of MN as the optimal one-step detector, i.e., full diversity.

Proof: See Appendix. ∎

From this theorem, it can be seen that superimposing the incremental information on the LR information as in (3) is not only beneficial from the rate perspective of the LR receivers, but also, using the two-step detector admits low-complexity detection for the HR receivers. This detector enables full diversity to be achieved, which is the key performance metric at asymptotically high SNRs.

Note that the two-step detector shares the philosophy of successive cancellation in the sense that the detector recovers one component of the received signal and subsequently uses it to facilitate recovering the other components. That being said, we note that there is a fundamental difference between the mechanisms that underlie successive cancellation and the proposed two-step detector. In particular, in successive cancellation, the contribution of the recovered component is subtracted (i.e., canceled) from the received signal and the other components are detected from the residual signal. In contrast, in the two-step detector, no subtraction or cancellation is performed. The detector first decides on the subspace containing the information that is communicated non-coherently, and then it decides on the particular bases that span this subspace. These bases contain the information that is communicated coherently.

We conclude this section by noting that although the LR and HR components of $\mathbf{X}$ in $\mathbf{U}$ and $\mathbf{A}$, respectively, are independently selected at the transmitter, the communication channel couples these two components. This implies that the two-step detector is generally suboptimal and explains the performance deterioration that will be observed in Section V below.

B. Degrees of Freedom

We now analyze the number of degrees of freedom achieved by the LR and HR information layers. This analysis complements diversity gain analysis by providing a high-SNR characterization of the rate of increase of capacity; the degrees of a freedom of a scheme depend on the dimensions that this scheme spans [37]. To determine the degrees of freedom achieved by the proposed scheme, we will use the fact that the HR information is sent over the Stiefel manifold, whereas the basic LR information is sent over the Grassmann manifold, cf. Section II.

Corollary 1: The achievable degrees of freedom for the conjoined LR and HR layers is $2M(T - M/2)$, whereas the achievable degrees of freedom for the LR layer is $2M(T-M)$.

Proof: By construction, the LR and HR information are encoded over matrices that are isotropically distributed on $\mathbb{G}_{T,M}(\mathbb{C})$, and $\mathbb{U}_M(\mathbb{C})$, respectively. The proof follows directly from the dimensions of these manifolds, cf. Section II. ■

Unfortunately, the construction in (3) does not achieve the maximum number of degrees of freedom for the HR receivers. In particular, for these receivers, when no constraints are imposed on $\mathbf{X}$, the maximum number of degrees of freedom is given by $2T\min\{M, N\}$, which, under the conditions in Section II, reduces to $2TM$. However, by restricting the transmitted signal, $\mathbf{X}$, to lie in $\mathbb{S}_{T,M}(\mathbb{C})$, the number of degrees of freedom is reduced by M^2. This reduction can be regarded as the price paid to ensure that the rate of the basic LR information, which can be detected by all receivers, is maximized. Restricting $\mathbf{X}$ to be in $\mathbb{S}_{T,M}(\mathbb{C})$ is equivalent to restricting $\mathbf{A}$ to be in $\mathbb{U}_M$, which offers the advantage of preserving the channel statistics of the LR channel; cf. Section III-A.1. Hence, the construction in (3) ensures that code designs that are favorable for point-to-point non-coherent MIMO systems can be readily utilized in the current multi-resolution layered scheme. In the following section, we will present methods for generating constellations for communicating the HR information component on $\mathbb{U}_M$.

IV. Constructions Methods for Square Unitary Constellations

The ralization of the proposed multi-resolution scheme relies on non-coherent constellations designed on $\mathbb{G}_{T,M}(\mathbb{C})$ and coherent constellations designed on $\mathbb{U}_M$. Several methods for designing Grassmannian constellations are available, see e.g., [7] and [8]. However, methods for designing square unitary constellations are relatively scarce and are only available for particular dimensions and with specific cardinalities, see e.g., [30], [38], [39], [40] for parametric approaches for designing unitary constellations used in low SNR low-rate differential signaling for $M = 2$ transmit antennas. To facilitate communicating the HR information in the current layered framework, in this section we will present generic approaches for designing square unitary constellations for coherent signaling with arbitrary dimensions and cardinalities. Unfortunately, the design methods developed in [7] and [8] for designing constellations on $\mathbb{G}_{T,M}(\mathbb{C})$ are not readily applicable for designing constellations on $\mathbb{U}_M$. Indeed, these two spaces possess fundamentally different topological properties. For instance, the singluar values of $\mathbf{U}_1^\dagger\mathbf{U}_2$, $\mathbf{U}_1, \mathbf{U}_2 \in \mathbb{G}_{T,M}(\mathbb{C})$ represent the cosine of the principal angles between the subspaces spanned by $\mathbf{U}_1$ and $\mathbf{U}_2$ and hence less than 1, whereas the singular values of $\mathbf{A}_1^\dagger\mathbf{A}_2$, $\mathbf{A}_1, \mathbf{A}_2 \in \mathbb{U}_M$ are equal to 1. Being square, $\mathbf{A}_1$ and $\mathbf{A}_2$ do not represent subspaces. Rather, each spans the entire space, $\mathbb{C}^M$. The fundamental differences between $\mathbb{G}_{T,M}(\mathbb{C})$ and $\mathbb{U}_M$ render the distance metrics and the gradient computations and projections that underlie the design problem significantly different. Hence, applying the methods in [7] and [8] directly on $\mathbb{U}_M$ may compromise the efficacy of the resulting constellations.

Before presenting our design approaches, we note that a candidate constellation on $\mathbb{U}_M$ can be obtained from the standard 2×2 Alamouti structure [2], whereby a matrix $\mathbf{A}$ has the form:

$$\mathbf{A} = \frac{1}{\sqrt{2}}\begin{bmatrix} q_1 & q_2 \\ -q_2^* & q_1^* \end{bmatrix}, \tag{10}$$

where q_1 and q_2 are two complex symbols drawn from a constant modulus constellation, e.g., phase-shift keying (PSK). Despite its simplicity, this approach suffers from a fundamental drawback. In particular, since the modulus of q_1 and q_2 is fixed, each of these symbols spans one dimension, which implies that the matrix $\mathbf{A}$ spans two real dimensions only. Since the number of real dimensions spanned by a matrix in $\mathbb{U}_M$ is $M^2 = 4$, it can be seen that a matrix constructed as in (10) wastes half the available degrees of freedom. Therefore, although Alamouti code enjoys simple symbol-by-symbol detection, its restrictive structure results in degraded performance in comparison with generic unitary codes which are not necessarily constructed from phase-shift keying symbols. To avoid this drawback, in the next sections, we will present two approaches for designing constellations on $\mathbb{U}_M$, a greedy approach for designing the constellation points sequentially and a direct approach for designing them jointly. The philosophy of these approaches parallels that of the approaches developed in [7] for designing Grassmannian constellations, cf. Section V. Despite this analogy, the optimization details on the Grassmann manifold are fundamentally different from their counterparts on the unitary group. For instance, while the distance between points on $\mathbb{G}_{T,M}(\mathbb{C})$ are described by the angles between the subspaces that they span, the corresponding distance on $\mathbb{U}_M$ are described by the Frobenius norm of the difference between the respective unitary matrices. Another major difference lies in the way in which the gradients are computed on $\mathbb{G}_{T,M}(\mathbb{C})$ and $\mathbb{U}_M$ and the operators that project the gradients computed in the Euclidean space onto these spaces.

A. The Greedy Design

In this section we develop a sequential technique for designing the square unitary constellation, $\mathcal{C}_H$, which will be used to communicate the incremental information component to the HR receivers. The objective of this technique is to generate a constellation in which the minimum distance between points is large. As such, this technique aims at obtaining good sphere-packings that enable the capacity of the HR channel to be approached. In contrast with this technique, are the techniques that aim at maximizing the diversity gain described in Theorem 1. Although diversity gain does not directly relate to sphere-packing, the constellations generated by sphere-packing techniques can always be perturbed to ensure full diversity while having marginal impact on the packing tightness. Such a perturbation is possible because the diversity gain is a discrete criterion, whereas the distance between constellations points is a continuous one.

In the technique proposed in this section an arbitrary point on $\mathbb{U}_M$ is selected to be the first point of $\mathcal{C}_H$. Without loss of generality, this point can be the identity matrix $\mathbf{I}_M$. The following point in $\mathcal{C}_H$ is then taken to be the one that maximizes the Frobenius distance to the first point. The third

point of $\mathcal{C}_H$ is now taken to be the one that maximizes the minimum distance to the two point already designed, and so on. In particular, in the i-th step of this algorithm, the i-th point of $\mathcal{C}_H$ is chosen to be the point on $\mathbb{U}_M$ that maximizes the minimum distance to the $i-1$ points already designed. Hence, the points generated by this algorithm can be expressed as

$$\mathbf{A}_i = \arg\max_{\mathbf{A}\in\mathbb{U}_M} \min_{1\le j\le i-1} \|\mathbf{A}-\mathbf{A_j}\|^2{}_F \tag{11}$$

$$= \arg\min_{\mathbf{A}\in\mathbb{U}_M} \max_{1\le j\le i-1} \mathrm{Tr}\big(\Re(\mathbf{A}_j^{\mathcal{H}}\mathbf{A})\big). \tag{12}$$

The objective in (12) is not differentiable and hence not amenable to derivative-based optimization. To circumvent this difficulty, we use the Jacobi logarithm approximation, whereby

$$\max_{1\le j\le i-1} x_j \approx \left(\log\left(\sum_{j=1}^{i-1} \mathrm{e}^{x_j^r}\right)\right)^{\frac{1}{r}}. \tag{13}$$

In (13), r is a design parameter; a larger r yields a more accurate approximation, but may result in numerical difficulties. Using (13), the objective in (12) can be approximated by

$$f(\mathbf{A}) = \left(\log\Big(\sum_{j=1}^{i-1} \mathrm{e}^{\left(\mathrm{Tr}\left(\Re(\mathbf{A}_j^{\mathcal{H}}\mathbf{A})\right)\right)^r}\Big)\right)^{\frac{1}{r}}. \tag{14}$$

Our approach for finding the i-th point of the desired constellation is based on a version of the gradient descent algorithm on the unitary group [33]. Let $\mathbf{A}^{(0)} \in \mathbb{U}_M$ be the initial point, and let $\nabla_{\mathbf{A}^{(0)}} f$ be the projection of the gradient of $f(\mathbf{A})$ onto the tangent space of $\mathbb{U}_M$ at $\mathbf{A}^{(0)}$. A closed-form expression for this projection was given in [32] and [33]. To obtain this expression for $f(\mathbf{A})$ in (14), we denote the standard gradient[6] of $f(\mathbf{A})$ at $\mathbf{A}^{(0)}$ by $\mathbf{G}_{\mathbf{A}^{(0)}}$, i.e., $\mathbf{G}_{\mathbf{A}^{(0)}}$ is the complex $M \times M$ matrix whose $k\ell$-th entry contains the Wirtinger derivative $\frac{df(\mathbf{A})}{d\mathbf{A}_{k\ell}}\Big|_{\mathbf{A}=\mathbf{A}^{(0)}}$ [41]. Hence, this derivative can be expressed as [42] $\mathbf{G}_{\mathbf{A}}\big|_{\mathbf{A}=\mathbf{A}^{(0)}}$, where

$$\mathbf{G}_{\mathbf{A}} = \frac{\left(\log\left(\sum_{j=1}^{i-1} e^{\mathrm{Tr}^r\left(\Re\left(\mathbf{A}_j^{\mathcal{H}}\mathbf{A}\right)\right)}\right)\right)^{\frac{1}{r}-1}}{\sum_{j=1}^{i-1} e^{\mathrm{Tr}^r\left(\Re\left(\mathbf{A}_j^{\mathcal{H}}\mathbf{A}\right)\right)}} \times \sum_{j=1}^{i-1} e^{\mathrm{Tr}^r\left(\Re\left(\mathbf{A}_j^{\mathcal{H}}\mathbf{A}\right)\right)} \mathrm{Tr}^{r-1}\left(\Re\left(\mathbf{A}_j^{\mathcal{H}}\mathbf{A}\right)\right)\mathbf{A}_j, \tag{15}$$

where in writing (15), we have used the fact that the Wirtinger derivative

$$\frac{d\,\mathrm{Tr}\left(\Re\left(\mathbf{A}_j^{\mathcal{H}}\mathbf{A}\right)\right)}{d\mathbf{A}} = \mathbf{A}_j. \tag{16}$$

Using (15) and (16), the projection $\nabla_{\mathbf{A}^{(0)}} f$ can be readily expressed as [32], [33]

$$\nabla_{\mathbf{A}^{(0)}} f = \mathbf{S}_{\mathbf{A}}\mathbf{A}\Big|_{\mathbf{A}=\mathbf{A}^{(0)}}, \tag{17}$$

[6]The standard gradient refers to the gradient computed in the Euclidean space prior to projection, i.e., without considering the unitary group space in which the function $f(\mathbf{A})$ lies.

where $\mathbf{S}_{\mathbf{A}}$ is a skew Hermitian matrix given by

$$\mathbf{S}_{\mathbf{A}} = \mathbf{G}_{\mathbf{A}}\mathbf{A}^{\mathcal{H}} - \mathbf{A}\mathbf{G}_{\mathbf{A}}^{\mathcal{H}}. \tag{18}$$

In the standard gradient descent approach, subsequent iterations of the algorithm would generate the iterates $\mathbf{A}^{(n)} = \mathbf{A}^{(n-1)} - \tau^{(n)}\nabla f_{\mathbf{A}^{(n-1)}}$, where $\tau^{(n)} > 0$ is the step size at the n-th iteration. However, the iterates generated by this expression do not necessarily lie on $\mathbb{U}_M$. To overcome this difficulty, it was proposed in [33] to modify the projection expression (17) in order to ensure that the iterates of the gradient descent approach remain on $\mathbb{U}_M$. At the n-th iteration, the modified projection proposed in [33] can be expressed as

$$\tilde{\nabla}_{\mathbf{A}^{(n)}} f = \frac{1}{2}\mathbf{S}_{\mathbf{A}^{(n-1)}}\big(\mathbf{A}^{(n-1)} + \mathbf{A}^{(n)}\big). \tag{19}$$

Using $\tilde{\nabla}_{\mathbf{A}^{(n)}} f$, the iterates of the gradient descent approach, can be expressed as

$$\mathbf{A}^{(n)} = \mathbf{A}^{(n-1)} - \tau^{(n)}\tilde{\nabla} f_{\mathbf{A}^{(n-1)}} = \mathbf{A}^{(n-1)} - \frac{\tau^{(n)}}{2}\mathbf{S}_{\mathbf{A}^{(n-1)}}\big(\mathbf{A}^{(n-1)} + \mathbf{A}^{(n)}\big), \tag{20}$$

which, after rearrangement, yields

$$\mathbf{A}^{(n)} = \big(\mathbf{I} + \frac{\tau^{(n)}}{2}\mathbf{S}_{\mathbf{A}^{(n-1)}}\big)^{-1}\big(\mathbf{I} - \frac{\tau^{(n)}}{2}\mathbf{S}_{\mathbf{A}^{(n-1)}}\big)\mathbf{A}^{(n-1)}. \tag{21}$$

We note that because, by construction, the matrix $\mathbf{S}_{\mathbf{A}}$ is skew-Hermitian, cf. (18), the matrix $\big(\mathbf{I} + \frac{\tau^{(n)}}{2}\mathbf{S}_{\mathbf{A}^{(n-1)}}\big)^{-1}\big(\mathbf{I} - \frac{\tau^{(n)}}{2}\mathbf{S}_{\mathbf{A}^{(n-1)}}\big) \in \mathbb{U}_M$. In fact, this matrix represents the so-called Cayley transform [10] which maps the linear space of skew-Hermitian matrices to the non-linear space of unitary matrices. Hence, it can be seen that using the modified projection in (19) ensures that for every $n = 1, 2, \ldots,$ the iterates are confined to $\mathbb{U}_M$.

To implement the gradient descent approach, a step size $\tau^{(n)}$ that ensures convergence must be determined [33], [43]. Candidate values for such a $\tau^{(n)}$ are those that meet the strong Wolfe conditions [44], [45], which stipulate that, for some $\rho_1 \in [0, 1]$ and $\rho_2 \in [\rho_1, 1]$, $\tau^{(n)}$ must satisfy

$$f\big(\mathbf{A}^{(n)}\big) \le f\big(\mathbf{A}^{(n-1)}\big) + \rho_1\tau^{(n)}\frac{df\big(\mathbf{A}^{(n)}\big)}{d\tau^{(n)}}\Big|_{\tau^{(n)}=0}, \tag{22}$$

$$\Big|\frac{df\big(\mathbf{A}^{(n)}\big)}{d\tau^{(n)}}\Big| \le \rho_2\Big|\frac{df\big(\mathbf{A}^{(n)}\big)}{d\tau^{(n)}}\Big|_{\tau^{(n)}=0}\Big|. \tag{23}$$

The derivative $\frac{df\left(\mathbf{A}^{(n)}\right)}{d\tau^{(n)}}$ can be computed using the chain rule, whereby we can write

$$\frac{df\big(\mathbf{A}^{(n)}\big)}{d\tau^{(n)}} = \mathrm{Tr}\left(\Big(\frac{df(\mathbf{A})}{d\mathbf{A}}\Big|_{\mathbf{A}=\mathbf{A}^{(n)}}\Big)^{\mathcal{H}}\frac{d\mathbf{A}^{(n)}}{d\tau^{(n)}}\right) \tag{24}$$

$$= \mathrm{Tr}\Big(\mathbf{G}_{\mathbf{A}^{(n)}}^{\mathcal{H}}\frac{d\mathbf{A}^{(n)}}{d\tau^{(n)}}\Big). \tag{25}$$

where (25) follows from invoking (15). Using (20), the derivative $\frac{d\mathbf{A}^{(n)}}{d\tau^{(n)}}$ can be shown to be

$$\frac{d\mathbf{A}^{(n)}}{d\tau^{(n)}} = -\frac{1}{2}\Big(\mathbf{I} + \frac{\tau^{(n)}}{2}\mathbf{S}_{\mathbf{A}^{(n-1)}}\Big)^{-1}\mathbf{S}_{\mathbf{A}^{(n-1)}}\Big(\mathbf{A}^{(n-1)} + \mathbf{A}^{(n)}\Big). \tag{26}$$

The condition in (22) resembles the standard backtracking line search. In particular, it ensures that the value of the objective at the n-th iteration is smaller than its value at the $n-1$-th iteration by a constant times the derivative. This constant is controlled by ρ_1, which offers a trade off between accuracy and convergence speed; a smaller value of ρ_1 improves accuracy but slows convergence. In contrast with (22), the condition in (23) ensures that the value of the gradient at $\tau^{(n)}$ is closer to zero than its value at $\tau^{(n-1)}$. Similar to ρ_1, the role of ρ_2 is to control convergence and accuracy; a smaller value of ρ_2 implies faster convergence but lower accuracy.

In other words, ρ_1 and ρ_2 control the convergence of the gradient-descent algorithm. Choosing ρ_1 to be close to zero and ρ_2 to be close to one ensures accurate but slow convergence of the algorithm to a local optimum. In contrast, choosing ρ_1 to be close to one and ρ_2 to be close to ρ_1 implies faster convergence but potentially to a point relatively far from the nearest local optimum. For reasonable convergence, ρ_1 and ρ_2 are typically chosen to be 0.1 and 0.9, respectively. However, since the constellations herein are designed off-line, smaller values were chosen, e.g., $\rho_1 = 0.01$ and $\rho_2 = 0.99$. The proposed gradient descent greedy algorithm is summarized in Algorithm 1.

Algorithm 1 Gradient Descent Greedy Algorithm for Designing Constellations on $\mathbb{U}_M$

1: Given a set of existing constellation points $\{\mathbf{A}_j\}_{j=1}^{i-1}$ and an initial point for $\mathbf{A}_i$, $\mathbf{A}^{(0)} \in \mathbb{U}_M$.
2: Initialization: Set $n \leftarrow 1$, $\epsilon \geq 0$ and $0 \leq \rho_1 \leq \rho_2 \leq 1$
3: **while** true **do**
4: Prepare: Generate $\mathbf{S}_{\mathbf{A}^{(n-1)}}$, cf. (18).
5: Find a suitable step size $\tau^{(n)}$ that satisfies the strong Wolfe conditions, cf. (22) and (23).
6: Update: $\mathbf{A}^{(n-1)} \rightarrow \mathbf{A}^{(n)}$, cf. (21).
7: Stopping Check:
8: **if** $\|\tilde{\nabla}_{\mathbf{A}^{(n)}} f\| \leq \epsilon$ **then** stop, cf. (19),
9: **else** $n \leftarrow n+1$ and **continue**

We note that, because $\mathbb{U}_M$ is not convex, the problem in (12) and its approximated version are not convex. Hence, the constellations generated by Algorithm 1 are not necessarily optimal.

The greedy algorithm is easy to implement and can be used to generate constellations with tens of thousands of points. For larger constellations, the greedy algorithm might be inefficient and more systematic design approaches need to be considered. To obtain constellations that perform better that those generated by the greedy approach, in the next section we will develop a joint design technique.

B. The Direct Design

The greedy algorithm yields constellations with good distance spectra. However, constellations with more desirable distance spectra can be generated by using techniques in which the constellation points are designed jointly, rather than sequentially. In particular, in the joint technique, the constellation design problem can be cast in the following form:

$$\{\mathbf{A}_s\}_{s=1}^{|\mathcal{C}_H|} = \arg\min_{\{\mathbf{A}_s\}\in\mathbb{U}_M} \left(\log \left(\sum_{i=1}^{|\mathcal{C}_H|-1} \sum_{j=i+1}^{|\mathcal{C}_H|} \mathrm{e}^{\left(\mathrm{Tr}\left(\Re\left(\mathbf{A}_j^{\mathcal{H}}\mathbf{A}_i\right)\right)\right)^r} \right) \right)^{\frac{1}{r}}, \tag{27}$$

where (13) is used to approximate $\max\{\cdot\}$ and r is the approximation parameter, cf. SectionIV-A.

By placing the matrices $\{\mathbf{A}_r\}_{r=1}^{|\mathcal{C}_H|}$ along the main diagonal of a $|\mathcal{C}_H|M \times |\mathcal{C}_H|M$ matrix, the problem of jointly designing $|\mathcal{C}_H|$ points on $\mathbb{U}_M$ can be seen to be equivalent to a problem of designing one point $\tilde{\mathbf{A}}$ on $\mathbb{U}_{|\mathcal{C}_H|M}$. In particular, the design argument can be expressed as

$$\tilde{\mathbf{A}} = \arg\min_{\tilde{\mathbf{A}}\tilde{\mathbf{A}}^{\mathcal{H}}=\mathbf{I}} \left(\log \left(\sum_{i=1}^{|\mathcal{C}_H|-1} \sum_{j=i+1}^{|\mathcal{C}_H|} \mathrm{e}^{u_{ij}^r} \right) \right)^{\frac{1}{r}}, \tag{28}$$

where $u_{ij} = \mathrm{Tr}\left(\Re\left(\mathbf{I}_M^j \tilde{\mathbf{A}}^{\mathcal{H}} \mathbf{I}_M^{j\mathcal{H}} \mathbf{I}_M^i \tilde{\mathbf{A}} \mathbf{I}_M^{i\mathcal{H}}\right)\right)$, and $\mathbf{I}_M^i$ is a $M \times |\mathcal{C}_H|M$ zero matrix with $\mathbf{I}_M$ in the i^{th} block. Using the chain rule, the gradient of the objective in (28), $\tilde{\mathbf{G}}_{\tilde{\mathbf{A}}}$, can be expressed as

$$\tilde{\mathbf{G}}_{\tilde{\mathbf{A}}} = \frac{\left(\log\left(\sum_{i=1}^{|\mathcal{C}_H|-1}\sum_{j=i+1}^{|\mathcal{C}_H|} e^{u_{ij}^r}\right)\right)^{\frac{1}{r}-1}}{\sum_{i=1}^{|\mathcal{C}_H|-1}\sum_{j=i+1}^{|\mathcal{C}_H|} e^{u_{ij}^r}} \times \sum_{i=1}^{|\mathcal{C}_H|-1}\sum_{j=i+1}^{|\mathcal{C}_H|} e^{u_{ij}^r u_{ij}^{r-1}} \frac{du_{ij}}{d\tilde{\mathbf{A}}}, \tag{29}$$

and using the techniques in [46] the derivative $\frac{d\tilde{u}_{ij}}{d\tilde{\mathbf{A}}}$ can be expressed as

$$\frac{du_{ij}}{d\tilde{\mathbf{A}}} = \mathbf{I}_M^{j\mathcal{H}} \mathbf{I}_M^i \tilde{\mathbf{A}} \mathbf{I}_M^{i\mathcal{H}} \mathbf{I}_M^j + \mathbf{I}_M^{i\mathcal{H}} \mathbf{I}_M^j \tilde{\mathbf{A}} \mathbf{I}_M^{j\mathcal{H}} \mathbf{I}_M^i. \tag{30}$$

Using $\tilde{\mathbf{G}}_{\tilde{\mathbf{A}}}$, the counterpart of the modified projection in (19) at a given iterate $\tilde{\mathbf{A}}^{(n)}$ is given by $\tilde{\nabla}_{\tilde{\mathbf{A}}^{(n)}}$ and a gradient descent algorithm that resembles the one used in the greedy design technique described in Section IV-A can be readily implemented.

We conclude our exposition of the direct design approach by noting that although the matrix $\tilde{\mathbf{A}}$ lies in $\mathbb{U}_{|\mathcal{C}_H|M}$, which spans $|\mathcal{C}_H|^2M^2$ real dimensions, the space spanned by such matrices has only $|\mathcal{C}_H|M^2$ real dimensions. This is because $\tilde{\mathbf{A}}$ belongs to the submanifold of $\mathbb{U}_{|\mathcal{C}_H|M}$ that is spanned by block diagonal matrices, which can be shown to imply that the gradient $\tilde{\mathbf{G}}_{\tilde{\mathbf{A}}}$ in (29) also possesses a block diagonal structure. This further implies that choosing the initial matrix of the gradient descent approach, $\tilde{\mathbf{A}}^{(0)}$, to be block diagonal ensures subsequent iterates to be also block diagonal. This observation significantly reduces complexity and renders the direct approach more amenable to implementation for designing practical constellations on $\mathbb{U}_M$.

C. Complexity Analysis

We herein provide a comparison between the per-iteration computational complexity of the greedy design and the direct one.

1) Greedy Design: The objective $f(\mathbf{A})$ in (14) requires $|\mathcal{C}_H| - 1$ exponential computations, the derivative $\mathbf{G}_\mathbf{A}$ in (15) requires $O(|\mathcal{C}_H|M^2)$ complex multiplications, the projector, $\mathbf{S}_\mathbf{A}$ in (18), the updating formula for $\mathbf{A}^{(n)}$ in (21), and the derivative for computing the step-size in (26), all require $O(M^3)$ complex multiplications.

2) Direct Design: The objective in (27) requires $|\mathcal{C}_H|(|\mathcal{C}_H| - 1)/2$ exponential computations, the derivative $\mathbf{G}_\mathbf{A}$ in (29) requires $O(|\mathcal{C}_H|(|\mathcal{C}_H| - 1)M^2/2)$ complex multiplications, the projector, corresponding to $\mathbf{S}_\mathbf{A}$ in (18), the updating formula corresponding to in (21), and the derivative for computing the step-size corresponding to (26), all require $O(|\mathcal{C}_H|M^3)$ complex multiplications.

This comparison, implies that the direct design multiplies the order of the number of computations in the greedy design by $|\mathcal{C}_H|$, which is significant because generally $|\mathcal{C}_H| \gg M$.

V. Simulation Results

In this section, we provide numerical evaluations of the performance of the layered signalling scheme described in Section II using the detectors described in Section III-A and the coherent unitary constellations described in Section IV. The reported simulations are based on at least 100 error events, which imply that, under general assumptions, the 95% confidence interval lies within $[8\times10^{-(\nu+1)}, 2\times10^{-\nu}]$, where ν is the error probability [47]. For the non-coherent codes, we will use the Grassmannian constellations generated by the exponential parameterization method developed in [8] and the constellations generated using the direct method developed in [7].

For the method in [8], the Grassmannian constellation $\{\mathbf{U}_i\}$ is obtained by mapping a standard PSK and QAM scalar constellation. In particular,

$$\mathbf{U}_i = \exp\left(\begin{bmatrix} \mathbf{0}_M & \alpha\mathbf{V}_i \\ -\alpha\mathbf{V}_i^{\mathcal{H}} & \mathbf{0}_M \end{bmatrix}\right)\mathbf{I}_{T,M}, \qquad i = 1,\ldots,|\mathcal{C}_L|, \tag{31}$$

where $\mathbf{V}_i \in \mathbb{C}^{M\times(T-M)}$ represents the i-th point of the standard coherent constellation, $\mathbf{I}_{T,M} \in \mathbb{C}^{T\times M}$ is the matrix containing the first M columns of the $T\times T$ identity matrix, and α is a homothetic factor which ensures that the singular values of $\mathbf{V}_i$ are less than $\pi/2$ [8].[7]

For the direct method in [7], the Grassmannian constellation is obtained by finding approximate solutions to the following optimization problem:

$$\begin{aligned} &\min_{\{\mathbf{U}_r\}_{r=1}^{|\mathcal{C}_L|}} \max_{1\le i,j\le\mathcal{C}_L} \quad \mathrm{Tr}(\mathbf{\Sigma}_{ij}) \\ &\text{subject to } \mathbf{U}_k \in \mathbb{G}_{T,M}(\mathbb{C}), \quad k = 1,\ldots,|\mathcal{C}_L|. \end{aligned} \tag{32}$$

[7]The condition that the singular values of $\mathbf{V}_i$ must be less than $\pi/2$ ensures that (31) will yield two distinct subspaces for any two distinct matrices, $\mathbf{V}_i$, and $\mathbf{V}_j$, $i \neq j$ [8].

Although solving this problem directly is difficult, using (13) along with the conjugate gradient technique developed in [32] for optimization on $\mathbb{G}_{T,M}(\mathbb{C})$ were shown to yield constellations that exhibit superior performance to constellations generated using other approaches [7].

Similar to the design methods presented in Section IV for constellations on $\mathbb{U}_M$, neither the exponential mapping method nor the direct design method for designing constellations on the Grassmann manifold directly relate to the full diversity criterion that the maximum singular value $\max_{m=1,\ldots,m} s_m(\mathbf{U}_i^{\mathcal{H}}\mathbf{U}_j) < 1$, $i \neq j$; cf. Theorem 1. This is because the full diversity criterion is derived from minimizing the PEP, whereas the design criteria in the exponential mapping and the direct design method are derived from capacity-approaching sphere-packing considerations. That being said, in the numerical results presented hereinafter we verified that the constellations generated by either method satisfy the full diversity criterion. If, however, these constellations happen to have points that do not satisfy the diversity criterion, these points can be perturbed so that the singular values are bound away from 1. Such a perturbation will have marginal impact on the tightness of sphere-packing but will ensure that the diversity criterion is satisfied.

Example 1: In this example we consider a scenario in which the coherence time $T = 4$ and the number of transmit and receive antennas is $M = N = 2$. The LR information is communicated using 256-point Grassmannian constellations, whereas the HR information is communicated using the Alamouti scheme with two 4-QAM symbols. This setup yields an LR rate of 2 bits per channel use (bpcu) and an HR rate of 1 bpcu. We compare the symbol error rate (SER) when two Grassmannian constellations are used for the LR layer, one designed using the direct method and the other is designed using the exponential parameterization method.

The performance of each LR constellation is evaluated when optimal the optimal one-step detector in Section III-A.2.a and the two-step detector in Section III-A.2.b are used for detection. However, for the HR constellation we use the Alamouti detection scheme. Because of the Alamouti structure, the simplicity ML detection carries over to the case when the two-step detector is used instead of the one-step one. This can be observed by multiplying both sides of (2) by $\hat{\mathbf{U}}^{\mathcal{H}}$, where $\hat{\mathbf{U}}$ is the Grassmannian symbol detected in the first step of the two-step detector. Now, an incorrect decision on $\hat{\mathbf{U}}$ is independent of the Alamouti scheme and $\mathbf{X}$ is erroneously detected. Hence, it suffices to consider the case in which $\hat{\mathbf{U}}$ is the correct symbol. Since $\hat{\mathbf{U}}$ is unitary, the multiplication does not change the noise statistics and the Alamouti detection scheme can be readily utilized, but with the equivalent received signal matrix $\hat{\mathbf{U}}^{\mathcal{H}}\mathbf{Y}$.

For the LR constellation designed using the exponential parameterization method in (31), the space-time code matrix, $\mathbf{V}$, was chosen to be the given by [5]

$$\mathbf{V} = \begin{bmatrix} s_1 + \theta s_2 & \phi(s_3 + \theta s_4) \\ \phi(s_3 - \theta s_4) & s_1 - \theta s_2 \end{bmatrix}, \tag{33}$$

where $\phi^2 = \theta = e^{i\frac{\pi}{4}}$ and s_i, $i = 1,\cdots,4$, are the four 4-QAM symbols to be transmitted on the LR layer. In this case,

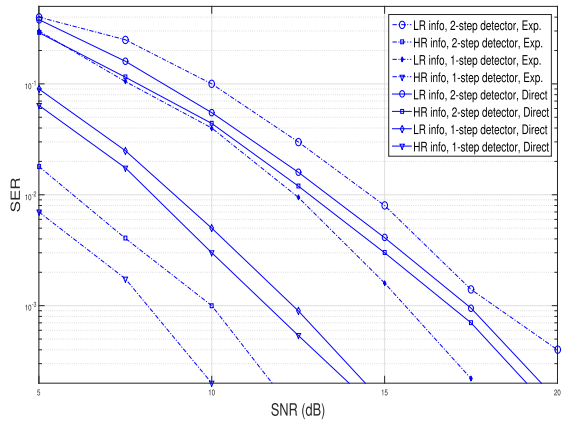


Fig. 2. The LR layer is constructed on $\mathbb{G}_{4,2}(\mathbb{C})$ using the direct design and the exponential parameterization technique; the HR layer is constructed using the 2×2 Alamouti scheme.

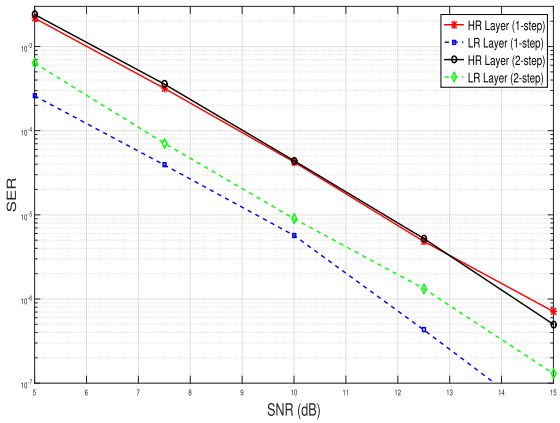


Fig. 3. SER Comparison of the two-step and the one-step detectors with a 0.25 bpcu LR layer.

the homothetic factor that maximizes the product distance is $\alpha = 0.3$ [8].

From Fig. 2, it can be seen that all the SER curves have identical high-SNR slopes. This confirms that the underlying constellations satisfy the full-diversity criteria in Theorem 1 and confirms that the performance of the non-coherent layer receiver is unaffected by the transmission of the HR layer. We also note from this figure that while directly generated Grassmannian constellations outperform those generated with the exponential map in the non-layered framework, the situation can be reversed in the layered one. In fact, the constellations with close-to-optimal performance in non-layered communication might drift away from optimality when used in a layered framework. □

Example 2: The purpose of this example is to further investigate the performance of the two-step detector. We consider a scenario similar to the one considered in Example 1, i.e., with $T = 4$, $M = N = 2$ and the HR information is communicated using the Alamouti scheme with two 4-QAM symbols. However, in this example, we consider the extreme case in which the cardinality of the LR constellation is $|\mathcal{C}_L| = 2$, corresponding to an LR rate of 0.25 bpcu. For this scenario, the performance of the two-step detector is compared with the one-step one in Fig. 3. The latter jointly decodes the composite signal containing the LR and HR components and subsequently separates these components. From this figure, it can be seen that, for the LR layer, the one-step detector has an SNR advantage of 1 dB at an SER of 10^{-5}. However, for the HR layer, the one-step detector and the two-step one yield essentially the same performance. Comparing this scenario with the one considered in Example 1, it can be seen that the performance gap between the two-step detector and the optimal one-step one depends on the LR rate; the higher the rate the bigger the performance gap. □

Example 3: In this example, we compare the performance of the unitary constellations designed using the techniques outlined in Section IV with those designed in [30]. The objective of the latter techniques is to generate differential space-time constellations with an underlying group-structure

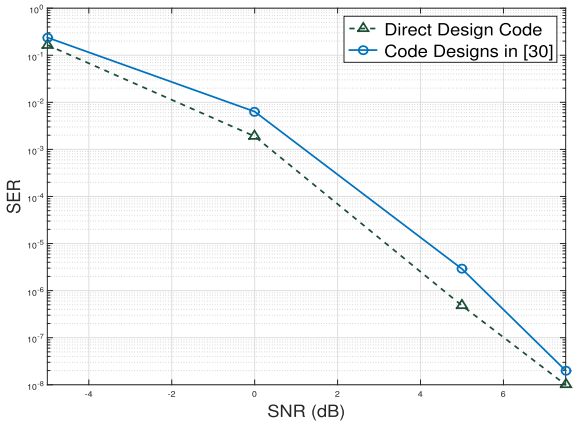


Fig. 4. Comparison with the 4×4 constellations designs in [30].

to facilitate detection in non-coherent communication frameworks. In contrast, the objective of the design techniques proposed herein is to generate unitary constellations that perform well in the multi-resolution framework, but without necessarily possessing a group structure. Unfortunately, the group-structure restriction renders the design methods in [30] suitable only for generating constellations with particular cardinalities, rather than arbitrary ones that may arise in practice.

In Fig. 4 we compare the performance of the 4×4 unitary constellation obtained in [30] using the $K_{1,1,-1}$ group with 240 elements with a 240-point constellation generated using our direct design methodology. From this figure, it can be seen that, the design method proposed here has performance advantage over the one proposed in [30]. For instance, at an SER of 10^{-4}, the constellation generated by the method proposed herein yields a performance advantage of about 1 dB over the constellations proposed in [30]. This can be attributed to the fact that the objective in [30] is not generate constellation with desirable distance properties, but rather to generate constellations that are suitable for differential non-coherent MIMO communications. □

Example 4: In this example, we investigate the minimum distance of the unitary constellations generated by the direct design methodology of Section IV-B and the corresponding

TABLE I

A $d_{\min}$ COMPARISON FOR THE UNITARY CONSTELLATIONS DEVELOPED HEREIN AND IN [30] AND [40] FOR $M = 2$

	$SL_2(\mathbb{F}_5)$ [30]	Orthogonal [40]	Numerical [40]	Geometric [40]	Direct (herein)
$\lvert\mathcal{C}_H\rvert$	120	122	121	120	120
$d_{\min}$	0.8740	0.5634	1.0991	1.1756	1.2275

TABLE II

A $d_{\min}$ COMPARISON FOR THE UNITARY CONSTELLATIONS DEVELOPED HEREIN AND IN [30] FOR $M = 4$

	$K_{1,1,-1}$ [30]	Direct (herein)
$\lvert\mathcal{C}_H\rvert$	240	240
$d_{\min}$	1.8784	2.2448

distance for the constellations developed in [30] and [40]. Another design of unitary constellations is provided in [39], but this design is subsumed by the ones [40].

Maximizing the minimum distance between points on the unitary group is the objective of the design methodology of Section IV-B. This objective is equivalent to maximizing the sphere-packing density on the unitary group[8] and is given by

$$d_{\min} = \min_{\mathbf{A}_i, \mathbf{A}_j \in \mathcal{C}_H,\, \mathbf{A}_i \neq \mathbf{A}_j} \|\mathbf{A}_i - \mathbf{A}_j\|_F. \tag{34}$$

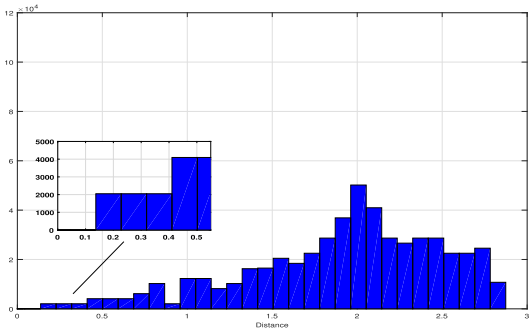
The values of $d_{\min}$ corresponding to the designs provided herein and in [30] and [40] are summarized in Table I for the case of $M = 2$, and the values of $d_{\min}$ corresponding to the designs provided herein and in [30] are summarized in Table II for the case of $M = 4$; no explicit designs are provided in [40] for the case of $M = 4$.

From Tables I and II, it can be seen that the direct design proposed in Section IV-B achieves larger minimum distances, $d_{\min}$, than all exisiting designs for the cases of $M = 2$ and $M = 4$. This advantage is due in part to the versatility of the proposed approach. Indeed, the unitary constellations in [30], [39], and [40] are constrainted to have a group (or a weak group) structure to facilitate their usage in differential signalling. This restriction is not present in the multi-layer signalling framework considered herein, which enables our design methodology to yield larger values of $d_{\min}$. Unlike the designs in [30], [39], and [40], our design approach enables the generation of unitary constellations with arbitrary cardinalities and dimensions, which is an important feature to enable multi-resolution multicasting in practical wireless networks.

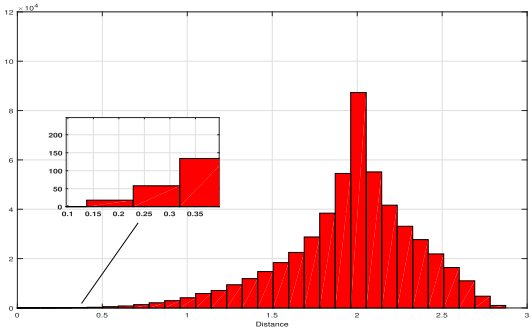
Finally, we note that, while the design objective in the framework of Section IV-B is to maximize $d_{\min}$, other design objectives can be readily incorporated in this framework, e.g., maximizng the diversity product, which governs SER performance at high SNRs [40].

Example 5: In this example, we compare the distance spectra and performance of two 1024-point constellations designed on the unitary group $\mathbb{U}_2$. Both constellations are designed using the gradient descent algorithm. However, one constellation is designed using the greedy approach in Section IV-A whereas the other constellation is designed using the direct approach in Section IV-B. The distance spectra and the performance of these constellations are compared with their Alamouti scheme counterparts at the same rate. To maintain unitarity the entries of the Alamouti matrices are chosen from PSK constellations.

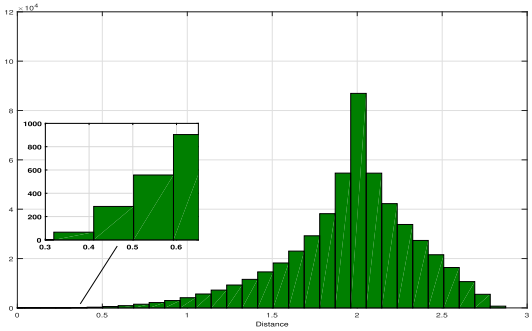
For the distance spectra, in Figs. 5(a), 5(b) and 5(c) we show the histogram of the Frobenius distances for the 1024-point 2×2 Alamouti, greedy and direct designs, respectively. It can



(a) Alamouti code ($d_{\min} = 0.196$)



(b) Greedy approach ($d_{min} = 0.1597$)



(c) Direct approach ($d_{min} = 0.3209$)

Fig. 5. Histogram of the Frobenius distances for the 1024-point HR constellation designed on $\mathbb{U}_2$ for different design approaches.

[8]This objective is also equivalent to maximizing the diversity sum, which is given by $\frac{1}{2\sqrt{M}} d_{\min}$.

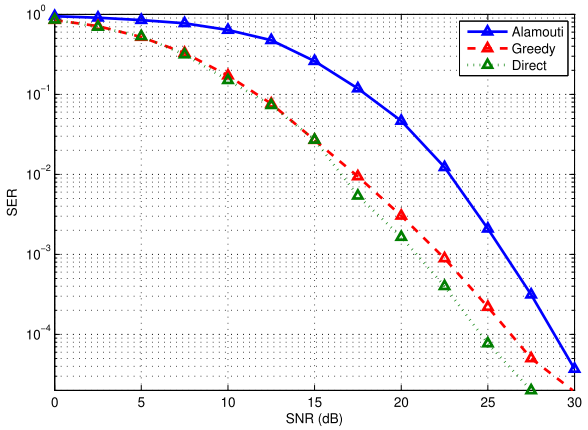
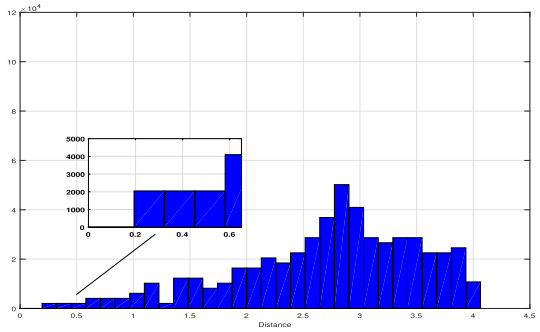


Fig. 6. Comparison between SER of the 1024-point 2×2 constellations designed onr $\mathbb{U}_2$ using the greedy and direct approaches and the corresponding, same rate, Alamouti-based constellation.

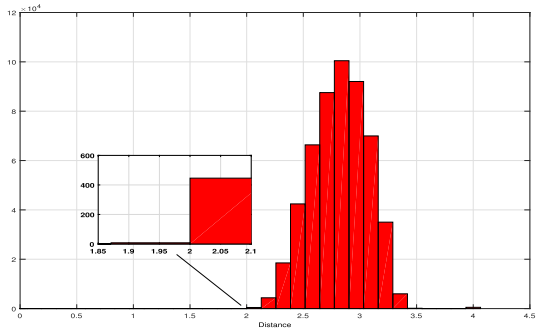
be seen from the figures that the minimum Frobenius distance in the case of the directly-designed constellation is larger than the corresponding distance in both the greedy and the Alamouti constellations. The effect of this distance advantage is investigated in Fig. 6. In particular, this figure shows this constellation yields significantly lower SERs than the other two constellations. For instance, at an SER of 10^{-4} the directly-designed constellation outperforms the greedy one by about 1 dB and the Alamouti-based one by more than 4 dB. Note that although the histograms for the greedy and direct designs look similar, at high SNRs, the direct design offers a performance advantage of about 1 dB over the greedy one. This advantage is due to the larger minimum distance between points of directly-designed constellations, cf. Figs. 5(b) and 5(c).

From Figs. 5(a), 5(b) and 6, it can be seen that, even though the minimum distance, $d_{\min}$, of the Alamouti-based constellation is greater than $d_{\min}$ of its greedy counterpart, the SER performance of the latter is better than that of the former. This is because maximizing $d_{\min}$ aims at developing dense packings on the unitary group, which is a capacity-based criterion rather than a performance-based one. Indeed, high-SNR performance is dominated by the product diversity [40]. In essence, this criterion could be, but is not explicitly, optimized in the designs proposed in Sections IV-A and IV-B. The effect of a larger $d_{\min}$ will manifest itself in coded systems operating close to capacity. □

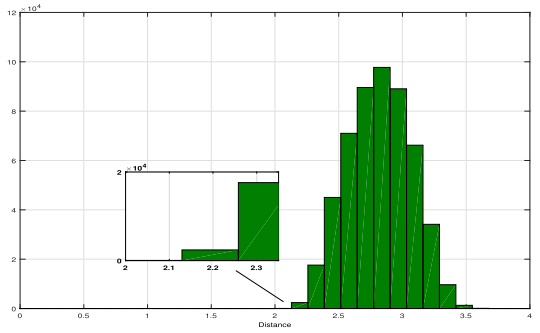
Example 6: In this example we compare the performance of the counterparts of the constellation considered in Example 5 but when the underlying design space is the unitary group $\mathbb{U}_4$. The first constellation is designed using the greedy approach, whereas the other constellation is design using the direct approach. The gradient-descent algorithm described in Section IV is used in both cases. The distance spectra and performance of these constellations are compared with the distance spectrum and performance of the corresponding 4×4 Alamouti-based constellation. The matrices of the latter constellations are structured as



(a) Alamouti-based code ($d_{\min} = 0.2772$)



(b) the greedy approach ($d_{\min} = 1.9907$)



(c) the direct approach ($d_{\min} = 2\ 1439$)

Fig. 7. Histogram of the Frobenius distances for the 1024-point HR code designed on $\mathbb{U}_4$ for different design approaches.

follows:

$$\mathbf{A} = \frac{1}{\sqrt{2}} \begin{bmatrix} \hat{\mathbf{A}} & \hat{\mathbf{A}} \\ -\hat{\mathbf{A}} & \hat{\mathbf{A}} \end{bmatrix}, \tag{35}$$

where $\hat{\mathbf{A}} = \frac{1}{\sqrt{2}} \begin{bmatrix} s_1 & s_2 \\ -s_2^* & s_1^* \end{bmatrix}$, where s_1 and s_2 are two PSK symbols. For instance, to design a 1024-point Alamouti-based constellation we use two 32-PSK symbols.

For the distance spectra, in Figs. 5(a), 5(b) and 5(c) we show the histogram of the Frobenius distances for the 1024-point 4×4 Alamouti-based, greedy and direct constellations, respectively. From these figures, it can be seen that $d_{\min}$ of the constellations generated by the greedy and the direct approaches are greater than the corresponding $d_{\min}$ of the Alamouti-based constellation. In particular, $d_{\min}$ for the constellations generated by the greedy and the direct approaches

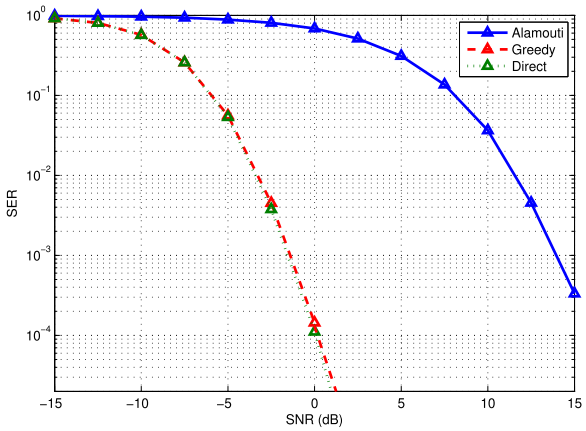


Fig. 8. Comparison between SER of the 1024-point 4×4 codes designed over $\mathbb{U}_4$ using the greedy approach, the direct approach and the corresponding Alamouti-based constellation.

are 1.9907 and 2.0576, respectively, whereas $d_{\min}$ for the Alamouti-based constellation is 0.2772. This relatively large distance advantage of the greedy and the direct constellations over the Alamouti-based one enables these constellations to achieve a significantly better SER performance.

For example, from Fig. 8 it can be seen that both the greedy and the direct 1024-point 4×4 constellations outperform the Alamouti-based constellation by about 15 dB at an SER of 10^{-3}; the performances of the greedy and the direct constellations are almost identical in this case, which is expected given the comparable values of their respective $d_{\min}$. □

VI. Conclusion

In this paper, we proposed a new layered multi-resolution multicast space-time coding scheme which allows the simultaneous transmission of LR non-coherent information to all receivers, including those with no CSI, and HR coherent information to those receivers that have reliable CSI. The proposed scheme ensures that the communication of the HR layer is transparent to the underlying LR layer. We showed that both the non-coherent and coherent receivers achieve full diversity, and we showed that the proposed scheme achieves the maximum number of communication degrees of freedom for non-coherent LR channels and coherent HR channels with unitarily-constrained input signals. To reduce the complexity of detecting high-rate HR constellations, we developed a two-step detector that was shown to achieve the same diversity order as the optimal one-step ML detector, but with much less computational cost.

Finally, we developed two generic techniques that rely on the gradient descent approach to generate unitary constellations with desirable distance spectra. The first method uses a low-complexity sequential design approach, whereas the second method uses a joint design approach that yields better constellations, but with a higher computational cost than its sequential counterpart. Simulation results confirm our theoretical findings and show that the constellations generated with these methods exhibit a significant performance advantage over existing unitary constellations when used to communicate the incremental HR information.

Appendix
Proof of Theorem 1

A decoding error event occurs in the two-step coherent detector if any of the two stages in the receiver fails to detect its layer correctly. Consider the PEP between any two pairs of distinct signals $\mathbf{X}_1 = \mathbf{U}_1\mathbf{A}_1$ and $\mathbf{X}_2 = \mathbf{U}_2\mathbf{A}_2$. The probability of mistaking $\mathbf{X}_2$ for $\mathbf{X}_1$ is

$$\begin{aligned}&\mathrm{PEP}(\mathbf{X}_1 \to \mathbf{X}_2)\\ &= \Pr(\mathbf{U}_1 \to \mathbf{U}_2) + \Pr(\mathbf{A}_1 \to \mathbf{A}_2|\mathbf{U}_1 \to \mathbf{U}_1|)\Pr(\mathbf{U}_1 \to \mathbf{U}_1)\\ &\le \Pr(\mathbf{U}_1 \to \mathbf{U}_2) + \Pr(\mathbf{A}_1 \to \mathbf{A}_2|\mathbf{U}_1 \to \mathbf{U}_1). \qquad (36)\end{aligned}$$

To compute $\Pr(\mathbf{U}_1 \to \mathbf{U}_2)$, we note that because $\mathbf{A}_1 \in \mathbb{U}_M$, $\mathbf{A}_1\mathbf{H}_i \overset{d}{=} \mathbf{H}_i$, $\forall i \in \mathcal{N}_C \cup \mathcal{N}_{NC}$. Hence, the ML detector in the first step will decide in favour of $\mathbf{U}_2$ if $p(\mathbf{Y}|\mathbf{U}_2) > p(\mathbf{Y}|\mathbf{U}_1)$, which, using (2) and the analysis in [7], amounts to the probability that $\|\mathbf{Y}_i^{\mathcal{H}}\mathbf{U}_2\| > \|\mathbf{Y}^{\mathcal{H}}\mathbf{U}_1\|$. This probability is independent of $\mathbf{A}_1$ and decays as ρ^{-MN_i} if and only if the non-coherent constellation, $\mathcal{C}_L$, achieves full diversity.

To compute $\Pr(\mathbf{A}_1 \to \mathbf{A}_2|\mathbf{U}_1 \to \mathbf{U}_1)$, we note that, because $\mathbf{H}_i$ is known, this probability is equivalent to the probability that $\|\mathbf{U}_1^{\mathcal{H}}\mathbf{Y}_i - \mathbf{A}_1\mathbf{H}_i\| > \|\mathbf{U}_1^{\mathcal{H}}\mathbf{Y}_i - \mathbf{A}_2\mathbf{H}_i\|$, $i \in \mathcal{N}_C \cup \mathcal{N}_{NC}$. Since $\mathbf{U}_1$ is already detected correctly in the first step, this probability decays as ρ^{-MN_i} if and only if the coherent constellation, $\mathcal{C}_H$, achieves full diversity, cf. [1].

Acknowledgment

The authors would like to acknowledge the valuable feedback provided by the anonymous reviewers.

References

[1] V. Tarokh, N. Seshadri, and A. R. Calderbank, "Space-time codes for high data rate wireless communication: Performance criterion and code construction," *IEEE Trans. Inf. Theory*, vol. 44, no. 2, pp. 744–765, Mar. 1998.

[2] S. Alamouti, "A simple transmit diversity technique for wireless communications," *IEEE J. Sel. Areas Commun.*, vol. 16, no. 8, pp. 1451–1458, Oct. 1998.

[3] B. Hassibi and B. Hochwald, "Linear dispersion codes," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2001, p. 325.

[4] S. Sandhu and A. Paulraj, "Space-time block codes: A capacity perspective," *IEEE Commun. Lett.*, vol. 4, no. 12, pp. 384–386, Dec. 2000.

[5] M. O. Damen, K. Abed-Meraim, and J. C. Belfiore, "Diagonal algebraic space-time block codes," *IEEE Trans. Inf. Theory*, vol. 48, no. 3, pp. 628–636, Mar. 2002.

[6] M. O. Damen and N. C. Beaulieu, "On diagonal algebraic space-time block codes," *IEEE Trans. Commun.*, vol. 51, no. 6, pp. 911–919, Jun. 2003.

[7] R. H. Gohary and T. N. Davidson, "Noncoherent MIMO communication: Grassmannian constellations and efficient detection," *IEEE Trans. Inf. Theory*, vol. 55, no. 3, pp. 1176–1205, Mar. 2009.

[8] I. Kammoun, A. M. Cipriano, and J. C. Belfiore, "Non-coherent codes over the Grassmannian," *IEEE Trans. Wireless Commun.*, vol. 6, no. 10, pp. 3657–3667, Oct. 2007.

[9] B. M. Hochwald and W. Sweldens, "Differential unitary space-time modulation," *IEEE Trans. Commun.*, vol. 48, no. 12, pp. 2041–2052, Dec. 2000.

[10] B. Hassibi and B. M. Hochwald, "Cayley differential unitary space-time codes," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1485–1503, Jun. 2002.

[11] S. X. Wu, W.-K. Ma, and A. M.-C. So, "Physical-layer multicasting by stochastic transmit beamforming and Alamouti space-time coding," *IEEE Trans. Signal Process.*, vol. 61, no. 17, pp. 4230–4245, Sep. 2013.

[12] J. Joung, H. D. Nguyen, P. H. Tan, and S. Sun, "Multicast linear precoding for MIMO-OFDM systems," *IEEE Commun. Lett.*, vol. 19, no. 6, pp. 993–996, Jun. 2015.
[13] S. Schwarz and M. Rupp, "Transmit optimization for the MISO multicast interference channel," *IEEE Trans. Commun.*, vol. 63, no. 12, pp. 4936–4949, Dec. 2015.
[14] E. Koyuncu, C. Remling, X. Liu, and H. Jafarkhani, "Outage-optimized multicast beamforming with distributed limited feedback," *IEEE Trans. Wireless Commun.*, vol. 16, no. 4, pp. 2069–2082, Apr. 2017.
[15] Y. Li and A. Nosratinia, "Grassmannian–Euclidean superposition for MIMO broadcast channels," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2012, pp. 2491–2495.
[16] Y. Li and A. Nosratinia, "Product superposition for MIMO broadcast channels," *IEEE Trans. Inf. Theory*, vol. 58, no. 11, pp. 6839–6852, Nov. 2012.
[17] M. T. Hussien, K. G. Seddik, R. H. Gohary, M. Shaqfeh, H. Alnuweiri, and H. Yanikomeroglu, "Multi-resolution broadcasting over the Grassmann and Stiefel manifolds," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2014, pp. 1907–1911.
[18] M. Morimoto, M. Okada, and S. Komaki, "A hierarchical image transmission system in a fading channel," in *Proc. IEEE Int. Conf. Univ. Pers. Commun.*, Nov. 1995, pp. 769–772.
[19] J. Liu and A. Annamalai, "Multi-resolution signaling for multimedia multicasting," in *Proc. IEEE Veh. Technol. Conf.*, Sep. 2004, pp. 1088–1092.
[20] İ. E. Telatar, "Capacity of multi-antenna Gaussian channels," *Eur. Trans. Telecommun.*, vol. 10, no. 6, pp. 585–595, Dec. 1999.
[21] T. L. Marzetta and B. M. Hochwald, "Capacity of a mobile multiple-antenna communication link in Rayleigh flat fading," *IEEE Trans. Inf. Theory*, vol. 45, no. 1, pp. 139–157, Jan. 1999.
[22] B. M. Hochwald and T. L. Marzetta, "Unitary space-time modulation for multiple-antenna communications in Rayleigh flat fading," *IEEE Trans. Inf. Theory*, vol. 46, no. 2, pp. 543–564, Mar. 2000.
[23] L. Zheng and D. N. C. Tse, "Communication on the Grassmann manifold: A geometric approach to the noncoherent multiple-antenna channel," *IEEE Trans. Inf. Theory*, vol. 48, no. 2, pp. 359–383, Feb. 2002.
[24] J. H. Conway, R. H. Hardin, and N. J. A. Sloane, "Packing lines, planes, etc.: Packings in Grassmannian spaces," *Experim. Math.*, vol. 5, no. 2, pp. 139–159, 1996.
[25] D. Agrawal, T. Richardson, and R. Urbanke, "Packings in complex Grassmannnian space and their use as multiple-antenna signal constellations," Bell Labs, Lucent Technol., Murray Hill, NJ, USA, Tech. Rep., Jan. 1999.
[26] K. Schober, P. Jänis, and R. Wichman, "Geodesical codebook design for precoded MIMO systems," *IEEE Commun. Lett.*, vol. 13, no. 10, pp. 773–775, Oct. 2009.
[27] I. S. Dhillon, Jr., T. Stromher, R. W. Heath, Jr., and J. A. Tropp, "Constructing packings in Grassmannian manifolds via alternating projection," *Exper. Math.*, vol. 17, no. 1, pp. 9–35, 2008.
[28] P. Xia and G. B. Giannakis, "Design and analysis of transmit-beamforming based on limited-rate feedback," *IEEE Trans. Signal Process.*, vol. 54, no. 5, pp. 1853–1865, May 2006.
[29] X. Zheng, Y. Xie, J. Li, and P. Stoica, "MIMO transmit beamforming under uniform elemental power constraint," *IEEE Trans. Signal Process.*, vol. 55, no. 11, pp. 5395–5406, Nov. 2007.
[30] A. Shokrollahi, B. Hassibi, B. M. Hochwald, and W. Sweldens, "Representation theory for high-rate multiple-antenna code design," *IEEE Trans. Inf. Theory*, vol. 47, no. 6, pp. 2335–2367, Sep. 2001.
[31] G. Han and J. Rosenthal, "Unitary space–time constellation analysis: An upper bound for the diversity," *IEEE Trans. Inf. Theory*, vol. 52, no. 10, pp. 4713–4721, Oct. 2006.
[32] A. Edelman, T. A. Arias, and S. T. Smith, "The geometry of algorithms with orthogonality constraints," *SIAM J. Matrix Anal. Appl.*, vol. 20, no. 2, pp. 303–353, 1998.
[33] Z. Wen and W. Yin, "A feasible method for optimization with orthogonality constraints," *Math. Program.*, vol. 142, no. 1, pp. 397–434, 2013.
[34] M. D. Carmo, *Riemannian Geometry*. Cambridge, MA, USA: Birkhäuser, 1992.
[35] M. Brehler and M. K. Varanasi, "Asymptotic error probability analysis of quadratic receivers in Rayleigh-fading channels with applications to a unified analysis of coherent and noncoherent space-time receivers," *IEEE Trans. Inf. Theory*, vol. 47, no. 6, pp. 2383–2399, Sep. 2001.
[36] H. V. Trees, *Detection, Estimation, and Modulation Theory*. New York, NY, USA: Wiley, 1968.
[37] 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.
[38] A. Panagos, K. Kosbar, and A. I. Mohammad, "WLC21-4: Weak-group unitary space-time codes," in *Proc. IEEE Globecom*, Nov. 2006, pp. 1–4.
[39] X.-B. Liang and X.-G. Xia, "Unitary signal constellations for differential space-time modulation with two transmit antennas: Parametric codes, optimal designs, and bounds," *IEEE Trans. Inf. Theory*, vol. 48, no. 8, pp. 2291–2322, Aug. 2002.
[40] G. Han and J. Rosenthal, "Geometrical and numerical design of structured unitary space–time constellations," *IEEE Trans. Inf. Theory*, vol. 52, no. 8, pp. 3722–3735, Aug. 2006.
[41] K. Kreutz-Delgado, "The complex gradient operator and the $\mathbb{CR}$-calculus," Dept. Elect. Comput. Eng., Univ. California, San Diego, CA, USA, Jun. 2009. [Online]. Available: http://arxiv.org/abs/0906.4835
[42] A. Graham, *Kronecker Products and Matrix Calculus: With Applications*. New York, NY, USA: Elis Horwood Ltd., 1981.
[43] D. Goldfarb, Z. Wen, and W. Yin, "A curvilinear search method for p-harmonic flows on spheres," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 84–109, 2009.
[44] J. Nocedal and S. Wright, *Numerical Optimization*. New York, NY, USA: Springer, 2006.
[45] W. Sun and Y.-X. Yuan, *Optimization Theory and Methods: Nonlinear Programming*, vol. 1. Springer, 2006.
[46] K. B. Petersen and M. S. Pedersen, *The Matrix Cookbook*, vol. 7. Kongens Lyngby, Denmark: Tech. Univ. Denmark, 2008, p. 15.
[47] M. C. Jeruchim, P. Balaban, and K. S. Shanmugan, *Simulation of Communication Systems*. New York, NY, USA: Plenum Press, 1992.

Karim G. Seddik (S'04–M'08–SM'14) received the B.S. degree (Hons.) and the M.S. degree in electrical engineering from Alexandria University, Alexandria, Egypt, in 2001 and 2004, respectively, and the Ph.D. degree from the Electrical and Computer Engineering Department, University of Maryland at College Park, College Park, MD, USA, in 2008. He was an Assistant Professor with the Electrical Engineering Department, Alexandria University. He is currently an Associate Professor with the Electronics and Communications Engineering Department, The American University in Cairo, Egypt. His research interests include cooperative communications and networking, MIMO–OFDM systems, cognitive radio, layered channel coding, and distributed detection in wireless sensor networks.

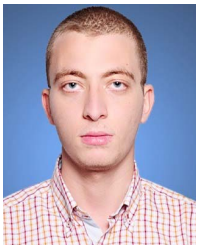
Dr. Seddik was a recipient of the Certificate of Honor from the Egyptian President for being ranked first among all departments, College of Engineering, Alexandria University, in 2002, the Graduate School Fellowship from the University of Maryland at College Park in 2004 and 2005, and the Future Faculty Program Fellowship from the University of Maryland at College park in 2007. He has served on the technical program committees of numerous IEEE conferences in wireless networks and mobile computing.

Ramy H. Gohary (S'02–M'06–SM'13) received the B.Eng. degree (Hons.) from Assiut University, Egypt, in 1996, the M.Sc. degree from Cairo University, Egypt, in 2000, and the Ph.D. degree from McMaster University, Hamilton, ON, Canada, in 2006, all in electronics and communications engineering. From 2006 to 2007, he was a Post-Doctoral Fellow with McMaster University. In 2008, he was a Visiting Scholar with the Electrical and Computer Engineering Department, University of Minnesota, MN, USA. From 2008 to 2010, he was a Visiting Scientist with the Terrestrial Wireless Systems Branch, Communications Research Center, Canada. From 2010 to 2012, he was the Project Manager of the Carleton-BlackBerry [formerly Research in Motion (RIM)] research project. He is currently the Project Manager of the Carleton-Huawei collaborative research project. He is currently an Assistant Professor with the Department of Systems and Computer Engineering. He is a member of the College of Reviewers and the Panel Review Board of the Ontario Center of Excellence. He received the Natural Sciences and Engineering Research Council Visiting Fellowship Award in 2007.

Dr. Gohary has co-authored over 80 IEEE journal and conference papers, and co-supervised over ten Ph.D. and Master's students. He is the co-inventor of five U.S. patents. He was also a Technical Program Committee Member for the 2013, 2014, 2015, and 2016 IEEE Wireless Communications and Networking Conference, the 2014 and 2015 IEEE Conference on Communications, the 2014 IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, the 2014 IEEE Global Telecommunications Conference, the 2015 Spring and Fall IEEE Vehicular Technology Conference (VTC), and the 2014 IEEE International Symposium on Signal Processing and Information Technology. He was the Volunteer Team Co-Chair of the IEEE VTC in 2010 and the Local Arrangement Co-Chair of the IEEE Signal Processing Advances for Wireless Communications, in 2014.

His research interests include the design of embedded systems, applications of embedded systems in machine-to-machine communications, Internet-of-Things, cross-layer design of wireless networks, analysis and design of MIMO wireless communication systems, applications of optimization and geometry in signal processing and communications, information theoretic aspects of multiuser communication systems, and applications of iterative detection and decoding techniques in multiple antenna and multiuser systems.

He is also interested in the analysis, design and hardware implementation of computer networks. He has been named the Best Professor by the Carleton Student Engineering Society for 2015. He is a registered Limited Engineering Licensee in the province of Ontario, Canada.



Mohammad T. Hussien (S'17) received the B.Sc. degree (Hons.) and the M.Sc. degree in electrical engineering from Alexandria University, Alexandria, Egypt, in 2012 and 2015, respectively. His M.Sc. thesis was layered channel coding in wireless communication systems. He is currently pursuing the Ph.D. degree in electrical engineering with the School of Electrical Engineering and Computer Science, Pennsylvania State University, PA, USA. He is currently a Research Assistant with the School of Electrical Engineering and Computer Science, Pennsylvania State University. His research interests include wireless communications and network coding.

Mohammad Shaqfeh (S'07–M'09) received the B.Sc. degree in electrical engineering (communications stream) from United Arab Emirates University, Al Ain, United Arab Emirates, in 2003, the M.Sc. degree in communications technology from Ulm University, Ulm, Germany, in 2005, and the Ph.D. degree from The University of Edinburgh, Edinburgh, U.K., in 2009. In 2009, he joined the Department of Electrical and Computer Engineering, Texas A&M University at Qatar, Doha, Qatar, where he is currently an Associate Research Scientist. His research interests include wireless communications systems, information theory, and smart transportation networks.

Hussein Alnuweiri (S'81–M'83–SM'17) received the Ph.D. degree in electrical and computer engineering from the University of Southern California at Los Angeles, Los Angeles, CA, USA, in 1989. He is currently a Professor with the Department of Electrical and Computer Engineering, Texas A&M University at Qatar, Doha, Qatar.

From 1991 to 2007, he was a Professor with the Department of Electrical and Computer Engineering, The University of British Columbia. From 1996 to 1998, he represented the University of British Columbia, Vancouver, BC, Canada, at the ATM Forum. From 2000 to 2006, he served as a Canadian Delegate with the ISO/IEC JTC1/SC29 Standards Committee (MPEG-4 Multimedia Delivery), where he was within the MPEG-4 standardization JTC1-SC29WG11 Group to develop the first client-server MPEG4 video streaming reference software.

He has a long record of industrial collaborations with several major companies worldwide. He has authored or co-authored over 200 refereed journal and conference papers in various areas of computer and communications research. He is also an inventor, and holds four U.S. patents. His research interests include mobile Internet technologies, cyber security and cyber systems, mobile cloud computing, wireless communications, routing and information dissemination algorithms in mobile networking, and quality-of-service provisioning and resource allocation in wireless networks.

Halim Yanikomeroglu (F'17) received the B.Sc. degree in electrical and electronics engineering from the Middle East Technical University, Ankara, Turkey, and the M.A.Sc. and the Ph.D. degrees in electrical and computer engineering from the University of Toronto, Canada. Since 1998, he has been with the Department of Systems and Computer Engineering, Carleton University, Ottawa, Canada, where he is currently a Full Professor. His research interests include wireless networks and technologies. In recent years, his research has been funded by Huawei, Telus, Allen Vanguard, Blackberry, Samsung, Industry Canada, Communications Research Center of Canada, DragonWave, Mapsted, and Nortel. This collaborative research resulted in about 25 patents.

Dr. Yanikomeroglu is a Distinguished Lecturer of the IEEE Communications Society and a Distinguished Speaker for the IEEE Vehicular Technology Society in 5G wireless technologies. He has been involved, in various capacities, in the organization of the IEEE Wireless Communications and Networking Conference (WCNC) since its inception in 1998, including serving as a Steering Committee Member, an Executive Committee Member, and the Technical Program Chair or Co-Chair of WCNC 2004, Atlanta, GA, USA, WCNC 2008, Las Vegas, NV, USA, and WCNC 2014, Istanbul, Turkey. He was the General Co-Chair of the IEEE 72nd Vehicular Technology Conference (VTC2010-Fall) held in Ottawa. He is currently serving as the General Chair of the IEEE 86th Vehicular Technology Conference (VTC2017-Fall) to be held in Toronto. He has served on the editorial boards of several IEEE journals. He was the Chair of the IEEE Technical Committee on Personal Communications (now called Wireless Technical Committee).