A Simple Precoding Scheme for Multi-User MIMO Transmission Over a Shared Channel in a TDD Cellular Network

Abheek Saha
Hughes Systique Corporation,
Gurgaon, India
Email: abheek.saha@hsc.com

***Abstract*—Multi-User Multiple-Input Multiple-Output (MIMO) transmission is one of the key technologies for achieving the ambitious targets for coverage and throughput in modern cellular networks. It allows us to take advantage of the large number of transmission elements possible in the radio-heads or eNodeB and allow users to share a channel. A key challenge in the deployment of multi-user MIMO is the problem of cross-user interference due to mutual non-orthogonality within the shared channel. The transmitter must select an optimal transmit precoding so as to eliminate this cross-user interference, since the receivers cannot coordinate and jointly decode the transmission. In this paper, we propose a novel algorithm for multi-user MIMO precoding over a shared channel. Our algorithm is a combination of ideas both from Dirty Paper Coding as well as the more recent Interference Alignment techniques. We demonstrate that we can achieve better performance than zero-forcing and that our algorithm is practical to implement within the framework of existing 4th and upcoming 5th generation systems, being realizable in linear time.**

Keywords*—*Multi-user MIMO; Dirty Paper Coding; Interference Alignment; Shared channel; Block Cholesky decomposition.

I. Introduction

Fourth and fifth generation cellular networks are distinguished by the rapid and widespread deployment of multiple antenna systems. These systems can be deployed in various ways; multiple antenna deployed at a single tower, multiple distributed antenna installations under the control of a single centralized Radio Access Node (cRAN), or even multiple cooperating eNodeBs (this is known as a Coordinated Multipoint or CoMP deployment). The initial deployment of multiple antenna transmission was in the single user MIMO systems [1], where a single eNodeB and a single User Terminal (UT) communicate over a dedicated channel using multiple receive and transmit elements. These have been commercially deployed for about a decade, with mixed results ; in practical environments, achieving more than 2 simultaneous streams per channel has proven to be difficult. Advances in radio technology are pushing multi-user MIMO (MuMIMO) as a replacement to single user MIMO. MuMIMO [2] consists of a single eNodeB with a large number of antennas to simultaneously transmit to multiple MIMO-capable UTs each equipped with a smaller number of antennas, over a common shared channel (Figure 1). The promise of MuMIMO is in the increase of the number of simultaneous streams per channel, hence increasing both the throughput and coverage at both cell-center and cell-edge. The primary advantage is that it is easier to put larger numbers of antenna elements on an eNodeB than on a UT (due to the form factor) and independent UTs with geographical separation have a higher spatial diversity than that of large number of antennas on a single receiver. This spatial diversity leads to a corresponding independence in channel matrices which is the basis of the gains of MIMO transmission.

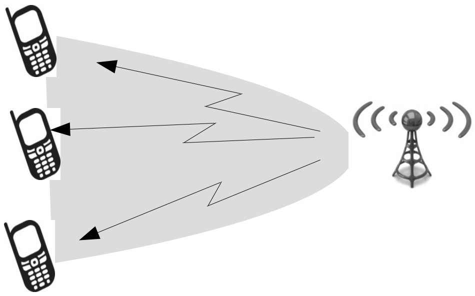

Figure 1. Deployment of MuMIMO

MuMIMO has been supported in 4th generation cellular networks (Long Term Evolution (LTE)) standards since Release 10 and is increasingly seeing commercial deployment. As we move towards the fifth generation, the number of antenna on the wireless network nodes (Remote Radio Heads or eNodeBs) is also increasing manifold (Massive MIMO). More complex modes of deployment, such as cooperative MuMIMO and multi-user CoMP are being proposed, especially to support the cell-edge (Figure 2) [3]. Over the last few years, the 3rd Generation Partnership Project has rapidly pushed the MuMIMO transmission modes into the mainstream of cellular access networks [4], standardizing the relevant operating modes, associated sounding and dedicated reference signals etc. The 5G New Radio standard (5G-NR) recognizes the importance of MuMIMO and has introduced further enhancements to the existing LTE standards to support more complex and efficient deployments [5]. This includes the use of comb-structures and cyclic shifts to support a higher number of orthogonal training signals, to support up to 32 simultaneous layers. The specifications also allow for flexible deployment of training signals, so as to support very low-latency decoding and adaptation for high-doppler environments.

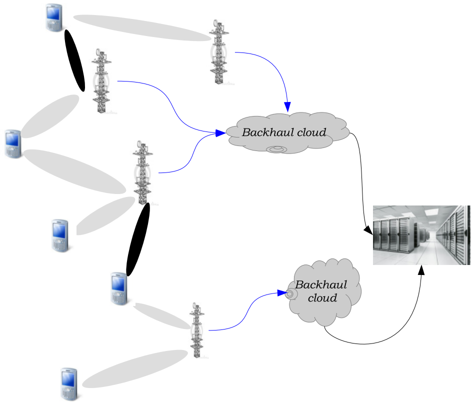

Figure 2. Cooperative MuMIMO

In contrast to the single-user MIMO, where the computation of optimal precoding matrix is well understood, optimal/near-optimal precoding/decoding for MuMIMO is relatively complex and has been the subject of much recent research. MuMIMO performance is limited by co-user interference over a complex shared medium; the individual channels of the individual UTs cause interference between each other in a manner which is tied to the mutual information in the channel. While the network can fully anticipate the cross-user interference, the UTs have only knowledge of their own channel and cannot coordinate with each other. This creates an interesting problem of optimal multi-user encoding at the network, which is the subject of this paper.

The contribution of this paper is as follows. We propose a novel MuMIMO transmit precoding scheme for a transmitter with $N * K, N, K > 1$ antenna transmitting to N receivers, each with K antenna elements. In the existing literature (Section III) there are two main approaches to the MuMIMO precoding problem. The earlier research was based on the technique of Dirty Paper Coding (DPC) [6], which is an elegant approach to solving the *known interference issue*, but is however, difficult to scale to a large number of users due to the need to solve a difficult joint optimization problem. On the other hand, in recent years, much work has taken place using the technique of Interference Alignment(IA) [7], which was primarily designed for the shared cross-channel environment, i.e., K transmitters and K receivers on a single channel. Our Successive interference Compensation (SiC) algorithm is a mixture of both approaches. We use the block diagonalization approach of DPC and join it with the sub-space reduction technique of interference alignment. We show that the resultant algorithm is fast, easy to implement and provides an intuitive outer bound of the K user MuMIMO bounds. We note that the K user MIMO case is relatively less addressed in the interference alignment literature as well as the older DPC literature, in terms of theoretical upper bounds for achievable rate. This shall be described in more detail in Subsection II-B.

The rest of this paper is organized as follows. In Section II, we describe the problem in context of the generic theory of interference channels and provide a survey of previous work, organized in terms of the two main approaches, Dirty Paper Coding (II-A) and Interference Alignment (II-B). In Section III, we give a detailed mathematical framework for our problem, followed by the description of the algorithm in Subsection III-A. In Section IV, we provide some simulation results comparing our approach against the baseline Zero-Forcing approach. Finally, in Section V, we conclude the paper by proposing future directions in our research.

II. Multi-user Interference Channels - theory and previous work

Our problem involves the multi-user shared channel in a generic LTE Time Division Duplex (TDD) cellular network. A single eNodeB with $N \times K$ antennas is controlling a cell in which there are N UTs with K receive antennas each. The eNodeB has to transmit simultaneously to all N UTs during each transmission frame over a shared good quality (high Signal to Noise Ratio (SNR)) Gaussian channel, varying stochastically from frame to frame. The eNodeB has perfect Channel State Information (CSI) for all UTs because it is a TDD network. It can also control the exact decoding matrix to be used by each UT, by embedding appropriate reference signals in each frame. The UTs have knowledge only of their own channels. Our aim is to design an algorithm by which the single eNodeB, given the knowledge of the complete set of channels for all N receivers, can near-optimally choose the transmission precoding matrix, so as to maximize the aggregate capacity for the channel. The aggregate capacity is a function of the number of layers transmitted, the number of UTs transmitted to and the Bit Error Rate (BER) for each UT. Because the system is operating in a high SNR environment, the system performance i.e. BER is constrained by the cross-user interference rather than the external noise.

The most generic case of a shared channel is the cross-channel case, which supports K transceiver pairs over a common channel. The broadcast channel case, on the other hand, has a single transmitter with K receivers. The common thread among both these cases is that the system performance (aggregate rate capacity) is limited by co-channel interference and the individual agents are cooperative, as pointed out by the authors in [8]. The problem of rate maximizing for selfish users is an open problem. In the most generic model of the distributed cross channel, neither the receivers nor the transmitters can coordinate with each other in realtime [9]. In other words, the individual receivers and transmitters have to individually process their own signals for precoding/ decoding. In such a system, the individual transmitters and receivers may agree on common parameters such as the precoding/decoding matrices and the operating codebook, but cannot share their processing in real-time; each will have to work on their own copy of the signal (transmit or receive) once the system is activated. Clearly, MuMIMO and Coordinated Multipoint systems are both special cases of the shared channel. MuMIMO is a broadcast channel case where there is a single transmitter, but multiple noncommunicating users. The CoMP case is a variant of the cross-channel, as there are multiple transmitters with limited ability to coordinate with each other.

Interest in the problem of shared channel transmission dates back at least 15 years starting from the two-user broadcast channel. The theoretical background is much older, dating back to the 1970s. The research question is as follows; how do we configure the transmitters and receivers in a shared channel so as to maximize the aggregate transmission rate. There are at least four parameters for optimization. The obvious ones are the transmit and receive filters (precoding/decoding matrices in MIMO terms). Alternately, one can design appropriate code-books, or decompose the code-book into separate subsets and reserve one for each transceiver pair. Finally, there exists user selection/scheduling. An associated problem is the need is to estimate the theoretical achievable rate capacity of such a channel in terms of the covariance between the channel matrices of individual users [10].

Over the years, there have been two major approaches to the shared channel rate optimization problem. The first is the DPC approach [6] as applied to the wireless channel, as is seen in the works of [11] and others. This approach works on the theory of *pre-compensation*; how to adjust the transmit signal so as to null out the effect of the known interference at the receiver. The second, which has garnered enormous interest of late, is the technique of interference alignment [12]. This technique (and its predecessor, Zero-Forcing(ZF)) works on the basis of one-time *optimal precoding*.

A. Dirty Paper Coding

Dirty Paper Coding originated in the work of Costa [6]. It solves the problem of transmitting a signal s to a receiver on top of a known (to the transmitter) interference vector z and a random noise term n. The problem is to construct an encoding operation $\mathcal{T}(z,n)$ based on the knowledge of z and r and a corresponding decoding operation $\mathcal{R}$, which can be used without knowledge of the interference term z. The original paper shows that the problem is solvable by proving the existence of an alphabet to encode s, z jointly and a corresponding pair of operations $\mathcal{T}, \mathcal{R}$, which can be used independently at either end of the channel for the encoding/decoding operations. A simple realization of DPC is Tomlinson Harashima Precoding (THP) [13][14], first introduced to solve a problem of self-interference due to cross-talk in cabled environments. In this work, $\mathcal{T}$ and $\mathcal{R}$ are modulo operations on the transmit symbol.

In the context of wireless and broadband MIMO, the early research in DPC focussed on the two receiver broadcast channel [15].The achievable rate for a two user broadcast channel have been extensively studied [16]–[18], culminating in the Marton's upper bound for two user broadcast channels.

In [15, Slepian Wolf Theorem], it is shown that the rate-capacity of the two user channel is limited only by the mutual information between the two signal-spaces and hence, achievable using a DPC method. The same result has been proven in different contexts by [19] and others.

Yu and Cioffi propose an alternate technique to achieve Marton's rate capacity in a two user broadcast system, using a decision feedback equalizer from the precoder output [11][20]. Published literature on practical DPC techniques for the shared wireless channel is relatively sparse, especially in the multiple user ($N > 2$) case. Much of the available literature uses Tomlinson Harashima precoding (or similar techniques) as a means of constrained interference suppression [21]. In [22], the authors pair the THP approach with a decision feedback filter to meet the power constraints on a per symbol basis. In [23], the authors implement a robust form of the THP for a decision feedback structure. Similar work is presented in [24]–[26].

B. Interference alignment

Interference alignment(IA) [9][12][27] works by decomposing a single channel into multiple sub-spaces, each corresponding to one of the degrees of freedom of each individual user. The key idea is that of trading degrees of freedom for interference [28]. In a standard IA realization, one of the subspaces is selected as the designated 'interference' subspace and all transmitters have to select an encoding such that the interference vector generated by that transmitter lies in the designated interference sub-space. This makes the other sub-spaces available for use for interference-free signal transmission. IA is a more efficient successor for the earlier zero-forcing (ZF) approach [29][30], as ZF requires each user to choose a separate interference sub-space, which is the null-space of the complement channel. The simplest case of IA is a K-user Multiple Input Single Output (MISO) interference channel [8], where K pairs of users sacrifice half the available degrees of freedom for interference free operation. Over the last ten years, an enormous corpus of literature has been created for interference alignment as a interference nulling technique in multiple contexts [7]. The theoretical work on interference alignment addresses the cross-channel case in two modes. In [12][27] we have two transmitters and two receivers sharing a single channel and both the transmitter and the receiver has multiple antenna. A specific subcase of the cross channel case is given in Section 10 of [12], which is the cognitive transmitter case; here, the two transmitters are able to share the transmit message that each intends to transmit to the other. In these enviroments, IA has been proposed as a distributed optimization problem [8][28] extended to the generic multi-antenna case in [31]. The other application of IA is in the K-user MISO cross-channel case [8], where we have a single transmitter transmitting to N users, each equipped with a single antenna. These IA techniques maybe adapted to the K user MuMIMO broadcast channel but are very complex to implement. This is both due to the full CSI requirement as well as the need to implement multiple matrix optimization passes for each frame. Practical algorithms mostly involve some version of ZF using rank-reduction technique [32][33]. Alternately, the somewhat more realizable alternating minimization algorithm in [34] can be used iteratively.

III. Successive Interference Compensation with Block Diagonalization

In this section, we present a simple to implement algorithm for precoding a MuMIMO transmission over a known broadcast channel with N receivers (UTs). The novelty of our solution is in that approaches interference compensation as in the DPC approach, using diagonalization to linearize the problem. It then uses the rank-reduction technique of interference alignment in order to compensate the interference vector without violating the transmit power norm. This idea of trading off between power and degrees of freedom is an adaptation of the theoretical work in [10].

As is standard in LTE cellular networks, the UTs are not aware that they are part of a MuMIMO cohort. They simply obey the eNodeB instructions as encoded in the reference signals to decode the transmitted symbols. In LTE Release 12 and onwards, this is achieved by an appropriately encoded Demodulation Reference Signals (DMRS) embedded in the transmission frame, which can be used both to control both the decoding matrix the UT will use and the number of streams each UT should decode. The UTs can only act as per the subset of information they have and cannot anticipate what the eNodeB is going to do. The eNodeB, however, requires full CSI information for all UTs. In a TDD network, this is directly available from the uplink. In an Frequency Division Duplex (FDD) network, this has to be signaled and has the additional complexity of quantization error. For the purpose of this paper, we assume that the CSI information for all user channels are available at the eNodeB with arbitrarily small error, as is achievable in a standard TDD network.

The SiC algorithm is computationally simple as it is single-pass and only involves matrix operations of size $K \times K$. The inversion of a Hermitian matrix required in the first block diagonalization stage is easily computed from the singular value decomposition. In contrast, the ZF approach requires us to implement Gram-Schmidt orthogonolization of an assymetric $N.(K-1) \times N.K$ matrix, this operation being repeated N times per frame. Standard IA algorithms are even more complex, because the optimal matrix search requires multiple iterative passes, each of which require quadratic operations on the entire $N.K \times N.K$ transmission matrix.

In the rest of the paper, we use the following conventions. We number matrix/vector rows and columns from 1 to N. Variables denoted by capital letters, i.e., $A, B, C...$ are considered to be elements of $\mathcal{M}_{N.K \times N.K}$ the set of matrices of $N.K$ rows and columns. Variables of the form $\tilde{A}_{m,n}, \tilde{B}_{m,n}, \tilde{C}_{m,n}$ represent the sub-matrix of size $K \times K$ of the corresponding matrix A, B, C etc, starting from the row position $(m-1) \times K$ and column position $(n-1) \times K$. Vectors are denoted by lower case letters x, y, etc. Uppercase greek letters (Υ, Γ, etc.) are used exclusively to denote diagonal or block-diagonal matrices and $\tilde{\Upsilon}_{m,n}, \tilde{\Gamma}_{m,n}$ their $K \times K$ size submatrices as defined above. Vector norms are denoted by $|x|$. For the equuivalent matrix norms, we utilize the trace function, which is given by $|X| = \text{Tr}(X)$. x^* and X^* represent the complex transpose of the vector x and the conjugate transpose of the matrix X respectively. The square root of a Hermitian semidefinite matrix S is obtained by taking the Singular Value Decomposition $S = V \Sigma V^*$ and then constructing the root $S^{1/2} = V \Sigma^{1/2} V^*$. S^{-1} is the matrix inverse.

A. Algorithm description

The SiC algorithm is implemented in two steps. In the first step (Subsection III-B), we block diagonalize the channel and hence separate the interference and signal space for each user. In the second step (Subsection III-C), we add a compensating vector for each user to cancel the effect of the causal noise and simultaneously reduce the number of transmitted streams so as to normalize the transmit power.

The interference compensation step can be interpreted both in the DPC sense and in the IA sense. In the DPC sense, we are successively modifying the space of code-words for each user to take into account the code-word transmitted by the previous user. If we see this in the sense of the formulation provided in [10], we are essentially choosing a transmit code-word from a modified dictionary $\mathcal{W}$, which maps to a subset of valid receiver code-words, but can cancel interference without violating the power transmit norm. In the IA sense, we can view the rank reduction step as a tradeoff between the degrees of freedom in the spatial sense to reduce the overlap between the users, without *fully orthonormalizing them*. This trade-off frees up some power so that we can add the additional compensating vector to cancel out interference. Thus, we are considering the combination of power and MIMO spatial sub-channels as a joint resource within which the optimal operating configuration has to be found.

Our algorithm improves upon the performance of standard MIMO IA algorithms, which are completely driven by the condition number of the aggregate channel matrix. If the condition number is large, i.e., the individual channel matrices are strongly correlated, IA algorithms provide poor results for all the UTs. This is because the act of subspace decomposition forces each user into a very poor channel, in order to achieve orthogonalization with respect to the common channel. Consider the worst case where there are two users, both with the exact same channel. The nullspace of one is the nullspace of the other, and neither will achieve any transmission in the IA case. In the SiC algorithm, at least one of the users will get through with no interference whatsover (the one which is encoded first), at the cost of the subsequent users.

B. Block Decomposition of a Composite Channel Matrix

The Block Cholesky decomposition (BLDL) technique has been used for DPC of the MuMIMO channel because it converts a multi-variate optimization problem to an stepwise optimization problem [35][36]. It allows us to decompose each UT's channel into a simple $K \times K$ *effective channel*, independent of the other UTs. For a symmetric matrix, the $K \times K$ block decomposition is computationally simple, because the diagonal matrices can be easily inverted. We let the channel matrix between the eNodeB and the M UTs, each with K antenna be written as a composite $\mathcal{H} \in \mathcal{M}_{MK\times MK}$ as in (1).

$$\mathcal{H} = \left[\begin{array}{c} \tilde{H}_1 \\ \tilde{H}_2 \\ \dots \\ \tilde{H}_M \end{array} \right] = \left[\begin{array}{cccc} \tilde{H}_{1,1} & \tilde{H}_{1,2} & \dots & \tilde{H}_{1,M} \\ \tilde{H}_{2,1} & \tilde{H}_{2,2} & \dots & \tilde{H}_{2,M} \\ \dots & \dots & \dots & \dots \\ \tilde{H}_{M,1} & \tilde{H}_{M,2} & \dots & \tilde{H}_{M,M} \end{array} \right] \quad (1)$$

The matrix $\tilde{H}_j$ represents the channel between the j UT and the eNodeB. Each $\tilde{H}_{j,k}$ is a $K \times K$ matrix within the composite matrix, where the diagonal terms represent the interference free channel and the off-diagonal terms represent the covariance between the different UTs. In the first step, we carry out Block Cholesky decomposition composite matrix $\mathcal{H}\mathcal{H}^*$ in the form given in (2), where the size of each sub-matrix is $K \times K$.

$$\begin{aligned} \mathcal{H}\mathcal{H}^* &= \mathcal{G}\Sigma\mathcal{G}^* \\ \Sigma &= \text{diag}[\tilde{S}_1, \tilde{S}_2, \dots, \tilde{S}_M] \\ \mathcal{G} &= \left[\begin{array}{ccccc} I & 0 & 0 & \dots & 0 \\ \tilde{G}_{2,1} & I & 0 & \dots & 0 \\ \dots & \dots & \dots & 0 & 0 \\ \tilde{G}_{M,1} & \tilde{G}_{M,2} & \dots & I & \end{array} \right] \\ \tilde{H}_{k,k} &= \sum_{p<k} \tilde{G}_{k,p}\tilde{S}_p\tilde{G}^*_{k,p} + S_{k,k} \\ \tilde{H}_{j,k<j} &= \sum_{p<k} \tilde{G}_{j,p}\tilde{S}_p\tilde{G}^*_{k,p} + \tilde{G}_{j,k}S_{k,k} \end{aligned} \quad (2)$$

Note that $S_k, 1 \leq k \leq N$ is a sequence of symmetric positive semi-definite matrices. We can write the singular value decomposition of S_k as in (3), where U is once again a unitary matrix of size $K \times K$

$$\tilde{S}_k = \tilde{U}_k\tilde{\Delta}_k\tilde{U}^*_k \quad (3)$$

The eNodeB precodes the transmission by the precoding matrix given in (4) choosing λ_k so as to meet the transmit norm $||P|| = 1$.

$$\mathcal{P} = \mathcal{H}^*\mathcal{G} \left[\begin{array}{cccc} \tilde{U}_1\lambda_1 & 0 & \dots & 0 \\ 0 & \tilde{U}_2\lambda_2 & 0 & \dots \\ \dots & \dots & \tilde{U}_{M-1}\lambda_{M-1} & 0 \\ \dots & \dots & 0 & \tilde{U}_M\lambda_M \end{array} \right] \quad (4)$$

The precoding is implemented on an appropriately chosen transmit vector z comprising of a block of K size transmit vectors $\tilde{z}_k$, each k-th block targetted to the k-th receiver (5). λ_k is the power loading term.

$$z = \left[\begin{array}{cccc} \tilde{z}_1 & \tilde{z}_2 & \dots & \tilde{z}_M \end{array} \right] \quad (5)$$

In the rest of this paper, we have assumed that $\lambda_k = \Delta_k^{-1/2}$ which essentially ensures that the eNodeB has a fixed power output $(NK)^2|z|$. The composite signal after passing through the channel is given as the vector r in (8).

A particularly useful feature of the BLDL decomposition is that the amount of interference at each stage (the power norm of the interference vector) is computable step-wise from the sub-matrices of the block diagonalized channel matrix. Further, the total co-channel interference in the Block Diagonalization is upper bounded by $\sum_p |\tilde{\mathcal{H}}_{p,p} - \tilde{S}_{p,p}|$. We can verify this as follows. Assume that the individual transmit blocks $\tilde{z}_k$ are of unit norm. The interference vector for the k-th user is given as $\mathfrak{i}_k$ from (7). By triangle inequality, we get upper bound of $\mathfrak{i}_k$ as in (6).

$$\begin{aligned} & |\mathfrak{i}_k| \leq \sum_{p\leq k} |G_{k,p-1}\Delta^{1/2}_{p-1}\tilde{z}_p| \leq \sum_{p\leq k} |G_{k,p-1}S_{p-1}G^*_{k,p-1}| \\ & |\sum_{p\leq k} G_{p,p-1}S_{p-1}G_{p,p-1}| = |\tilde{H}_{k,k} - \tilde{S}_{k,k}| \\ & \Rightarrow |\mathfrak{i}_k| \leq |\tilde{H}_{k,k} - \tilde{S}_{k,k}| \end{aligned} \quad (6)$$

Intuitively, we can check the result from the fact that each receiver has K antennas and can thus coherently decode K streams. This means that the energy of the K streams can be removed from the interference seen by the system as a whole. Each submatrix $G_{k,j}, k \neq j$ then captures the mutual information between the k-th and the j-th user. If the sum of the off-diagonal terms of $\mathcal{H}\mathcal{H}^*$ were negligible, (i.e., $\sum_p \tilde{H}_{j,p}\tilde{H}_{p,j} \equiv 0$) in (2), then the inter-receiver co-channel interference terms would also vanish.

C. Cancelling the co-channel interference vector

Because of the nature of the precoding, the co-channel interference also takes a particular form, in that each i-th user is only affected by the interference generated by the previous users. We can verify this by formally deriving the interference vector $\mathfrak{i}_k$ from the structure of the receive vector given in (7).

$$\begin{aligned} y_1 &= \Delta_1^{1/2}\tilde{z}_1 \\ y_2 &= U_2^*G_{2,1}U_1\Delta_1^{1/2}\tilde{z}_1 + \Delta_2^{1/2}\tilde{z}_2 \\ & \dots \\ y_k &= \tilde{U}_k^* \sum_{j<k} G_{k,j}\tilde{\Delta}_j^{1/2}\tilde{U}_j\tilde{z}_j + \tilde{\Delta_k}^{1/2}\tilde{z}_k \\ &= \mathfrak{i}_k + \tilde{\Delta_k}^{1/2}\tilde{z}_k \end{aligned} \quad (7)$$

We will now compensate for this interference. To each transmit vector z_m, we shall add an additional compensating vector ζ_m, so that the combination, after precoding will counteract the effect of $\mathfrak{i}_m$, the known interference vector *for this, the m-th user.* While this step will cancel the interference vector completely, it may cause the combined output vector $z_m + \zeta_m$ to exceed the power norm. To take care of this, we shall truncate the transmit block as shown in (9). The output vector will have zero co-user interference, but some of the streams will be nulled out. We interpret this as a reduction in the degrees of freedom available for this particular channel. The only impact on the receiver is that it has to discard the last L_k symbols it receives. We repeat

$$
\begin{aligned}
r=& \mathcal{HP}\hat{z} = \mathcal{HH}^*\mathcal{G} \begin{bmatrix} \tilde{U}_1\Delta_1^{-1/2} & 0 & \dots & 0 \\ 0 & \tilde{U}_2\Delta_2^{-1/2} & 0 & \dots \\ \dots & \dots & \tilde{U}_{M-1}\Delta_{M-1}^{-1/2}\lambda_{M-1} & 0 \\ \dots & \dots & 0 & \tilde{U}_M\Delta_M^{-1/2}\lambda_M \end{bmatrix} \begin{bmatrix} \tilde{z}_1 \\ \tilde{z}_2 \\ \dots \\ \tilde{z}_M \end{bmatrix} \\
&= \lambda \begin{bmatrix} 1 & 0 & \dots & 0 \\ \tilde{G}_{2,1} & 1 & 0 & \dots \\ \dots & 1 & 0 & \dots \\ \tilde{G}_{M,1} & \tilde{G}_{M-1,1} & 1 & \dots \end{bmatrix} \begin{bmatrix} \tilde{U}_1\tilde{\Delta}_1^{1/2} & 0 & \dots & 0 \\ 0 & \tilde{U}_2\tilde{\Delta}_2^{1/2} & 0 & \dots \\ \dots & \dots & \tilde{U}_{M-1}\tilde{\Delta}_{M-1}^{1/2} & 0 \\ \dots & \dots & 0 & \tilde{U}_M\tilde{\Delta}_M^{1/2} \end{bmatrix} \begin{bmatrix} \tilde{z}_1 \\ \tilde{z}_2 \\ \dots \\ \tilde{z}_M \end{bmatrix} \qquad (8)
\end{aligned}
$$

this step successively for the $(k+1)$-th and then the $(k+2)$-th user, as long as the impact of the interference is more than the reduction of throughput due to truncation. We note that the interference vector $(i)_k$ has to be updated after the compensation is completed for the $(k-1)$-th user, because it is dependent on the compensated output as well. For each user, the crucial task, hence, is to minimize the value of L_k, the number of streams that the k-th user has to reduce.

$$
\begin{bmatrix} z_{k,1} \\ z_{k,2} \\ \dots \\ z_{k,K-L_k} \\ \dots \\ z_{k,K} \end{bmatrix} \rightarrow \begin{bmatrix} z_{k,1} \\ z_{k,2}+\zeta_{k,2} \\ \dots \\ z_{k,K-L_k}+\zeta_{k,K-L_k} \\ \dots \\ 0 \end{bmatrix} \qquad (9)
$$

1) Truncation step: We consider the problem of transmission of a vector $\tilde{z}_k$ through a channel with matrix $H_k \in \mathcal{M}_K$ with a known interference vector $\mathfrak{i}_k$ with norm ε_k. We can find an interference compensating vector ζ_k which we can add to $\tilde{z}_k$ to get $\hat{z}_k$ with $K-L$ non-zero spatial streams, such that $T(\tilde{z}_k, K-L) = T(\hat{z}_k + \mathfrak{i}_k), K-L)$, where $T(A,n)$ is the truncation operator. The receive vector y is given by the equation (10), where the eNodeB uses the precoding matrix $V\tilde{\Delta}^{-1}$, where V comes from the SVD of H, $H = U\Delta V^*$.

$$
y = U^*\mathfrak{i}_k\tilde{\Delta}^{-1} + (x+\zeta) \qquad (10)
$$

We know that U is a unitary matrix, so $|U^*\mathfrak{i}_k| = \varepsilon_k$. If we set $\zeta = \left(-U^*\mathfrak{i}_k\tilde{\Delta}^{-1}\right)$ then $y = x$. However, by triangle inequality

$$
|x+\zeta| \leq |x| + \frac{\varepsilon_k}{|\tilde{\Delta}|} \qquad (11)
$$

Hence, our modified transmit vector $x+\zeta$ may violate the transmit power norm by an amount up to $\frac{\varepsilon_k}{|\tilde{\Delta}|}$. To solve this problem, we reduce the number of spatial streams from K to $K-L$. Hence we only wish to find a ζ whereby the first $K-L$ entries in y match $K-L$ entries in x. The remaining entries in ζ are set to zero. The equation in (10) is then modified to that of (12), where $T(S,k)$ is the truncation operator when truncates the matrix S to its first k rows and columns.

$$
\begin{aligned}
\zeta=& -U^*\eta\tilde{\Delta}^{-1} \\
y=& U^*\mathfrak{i}_k\tilde{\Delta}^{-1} + \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_{K-L} \\ 0 \\ \dots \\ 0 \end{bmatrix} + \begin{bmatrix} \zeta_1 \\ \zeta_2 \\ \dots \\ \zeta_{K-L} \\ 0 \\ \dots \\ 0 \end{bmatrix}
\end{aligned} \qquad (12)
$$

D. Aggregate Rate

We now come to the problem of estimating the aggregate rate we achieve by the SiC algorithm. We recall that for a MIMO transmission, the aggregate rate for a given user is the sum of the eigen-values of the effective channel matrix for that user, corresponding to each stream or layer chosen for transmission. In our case, we have deliberately truncated the effective precoding matrix; this is the cost incurred for mitigating cross-user interference. From the expression in (12) we can estimate the number of streams L_k which the k-th user has to sacrifice in order to achieve the null interference condition. The aggregate simply becomes the throughput of remaining streams, as in (13)

$$
\sum_{i\in\mathcal{K}} \log_2\left(1+\alpha\lambda_i/\,\rho\right) \qquad (13)
$$

$\mathcal{K}$ is the set of streams which are retained for transmission and ρ is the wide-band Gaussian (non-causal) noise in the system. For a given k-th user facing the interference vector $\mathfrak{i}_k$ as given in (6), the number of streams which have to be reduced is given by L in the equation (14). Note that if $|\mathfrak{i}| \approx |\tilde{\Delta_k}^{1/2}|$, we get $L = K/2$ which is the MISO case.

$$
L = K\frac{|\mathfrak{i}|}{|\Sigma_k| + |\mathfrak{i}|} \qquad (14)
$$

Consider the vector z_k with known interference vector $\mathfrak{t}$. For each stream that we reduce, we reduce the norm of the interference vector by at least $1/K$. We also allow for the addition of a compensating vector ζ_k of norm $1/K$. To achieve null interference condition, we have to allow the norm of the compensating vector to match the worst case norm of the interference vector. Hence, we get $|\mathbb{i}|(K - L/K) = L/K$. Simplifying for L we get the result above.

From the above result and the overall upper bound on the interference given in (7), we get an upper bound on the total number of streams that have to be sacrificed to achieve the null interference condition. The total number of streams to be reduced over the entire cohort of M receivers is given by

$$\sum_{k=2}^{M} L_k = \sum_{k} \frac{|\tilde{\mathcal{H}}_k - \tilde{S}_k|}{|\tilde{\mathcal{H}}_k - \tilde{S}_k| + |\tilde{S}_k|} \qquad (15)$$

E. Optimal ordering

The expression in (2) also gives us a useful heuristic for ordering the UTs prior to the block-diagonalization stage. Let us assume that we are scheduling a set of UTs whose indices are given in U. If we compute the relative orthogonality of the k-th channel to the rest in terms of $\mathcal{R}(k, U) = \min_{k \neq j, j \in U} |H_{j,k} H^{*}_{j,k}|$, then ordering the UTs in descending order of $\mathcal{R}(k, U)$ improves the aggregate capacity. The initial UTs get the best transmission rate, since their interference vectors are relatively low. The UTs which may interfere with the others are further down the list. We can demonstrate this by a simple example. Consider a 3 UT system, where the UTs have channel matrices H_1, H_2 and $\alpha H_1 + (1-\alpha) H_2$, where H_1 and H_2 are perfectly orthogonal to each other. If we organize the composite channel matrix is $\tilde{H} = \begin{bmatrix} H_1 & H_2 & \alpha H_1 + (1-\alpha) H_2 \end{bmatrix}$, then the cross user interference vectors are 0, 0 and $\alpha^2 + (1-\alpha)^2$. On the other hand, if we flip the positions of the 2nd and the 3rd UTs, i.e., $\tilde{H} = \begin{bmatrix} H_1 & \alpha H_1 + (1-\alpha) H_2 & H_2 \end{bmatrix}$, then the cross user interference vectors are 0, α^2 and 1 respectively. As we can see, the second ordering has lower rate capacity though the first one is less fair. In general, we find the ordering in terms of descending $\mathcal{R}(k, u)$ gives good results as we shall see in Section IV below.

A scheduling algorithm to balance ordering and capacity is currently under study.

IV. Simulation Results and Discussion

In this section, we present some simulation results. We have simulated a system comprising of a single eNodeB with 64 transmit antenna (an 8x8 antenna configuration) and multiple UTs; only the basic downlink shared channel is implemented and the DMRS and other reference signals are communicated directly to the UTs. The entire simulation code is written in C and the key elements of the transmit and receive chain are implemented using the Gnu Scientific Library (GSL). The channel matrices are randomly generated (using the GSL random number generator) with full rank and condition number 0.5. At each frame, the eNodeB creates a transmit vector for transmission to the N users simultaneously, with K symbols per UT, from a standard 16-QAM constellation. All transmit vectors are normalized to a unit norm; hence, the power saved by decimating any of the spatial streams is distributed over the rest of the spatial streams. The channel matrices are then randomly generated and then the algorithm given in is implemented at the simulated eNodeB. The resultant precoded transmit vector is passed through the random composite channel with AWGN noise added to it and handed over to the UTs. At each UT, the decoding chain is implemented using the signaled decoding matrix and SNR computed individually. Figure 3 show the combined bit-rate achieved for four cases; 2 UTs of 32 antenna each, 4 UTs of 16 antenna each, 8 UTs of 8 antenna each, 16 UTs of 4 antenna each and finally, 32 UTs of 2 antenna each. For each configuration, we have run the simulation 500 times. As we can see in Figure 3, as the number of antenna increase, the total bitrate asymptotically approaches the best case performance. The gap in between is equivalent to the *coordination penalty* described by [37].

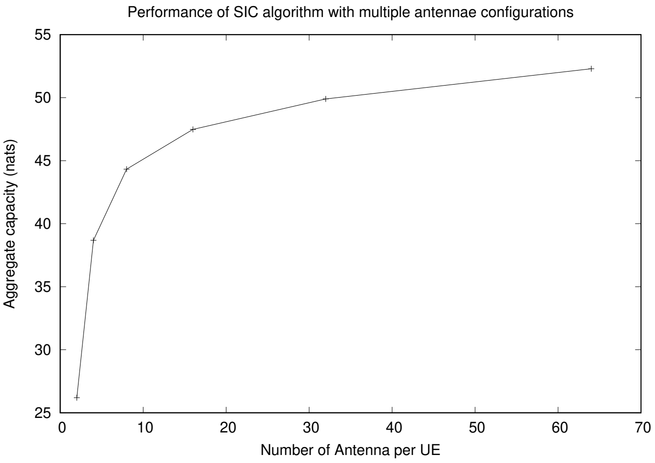

Figure 3. Relative performance of the successive interference cancellation algorithm vs full coordination

In Figure 4, we compare the performance of the system against a reference case. The reference for us is the zero-forcing algorithm as implemented in [29]. Zero-forcing worsk by setting the precoding matrix for each user to $P_i = H_{ii} \tilde{V}_{ii}$, where $\tilde{V}_{i,i}$ lies in the nullspace of the complementary vector ((16).

$$\tilde{H}_{i,i} = \begin{bmatrix} H_{i,1} & H_{i,2} & \dots & H_{i,i-1} & H_{i,i+1} & \dots \end{bmatrix} \qquad (16)$$

We chose ZF as the baseline algorithm, because as of now it remains the most practical algorithm in the $N >> 1, K > 2$ case, being implementable in approximately linear time. As mentioned earlier, MuMIMO implementations of existing IA or DPC algorithms, or more sophisticated lattice coding algorithms remain prohibitively expensive to implement since they scale super-linearly in N for large values of K. Further, multi-pass algorithms as suggested in literature are not realizable in current cellular networks, given that the corresponding signaling mechanisms don't exist. As we have previously indicated, the SiC algorithm should substantially outperform the ZF algorithm when the condition number of the aggregate channel matrix is high. The chart in Figure 5 shows the relative performance of the two algorithms for different channel condition numbers.

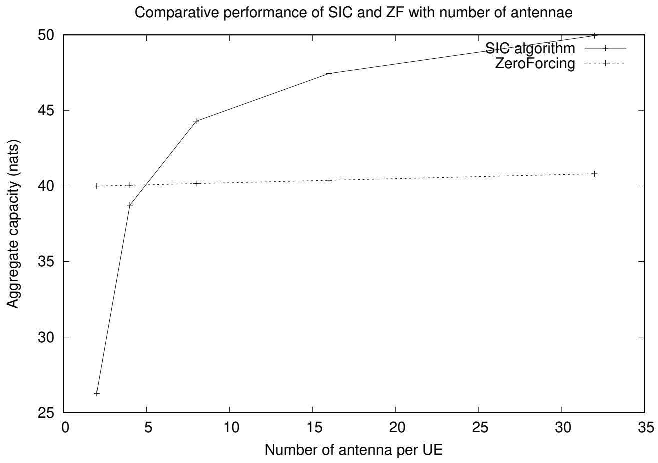

Figure 4. Performance of SIC Algorithm versus ZF

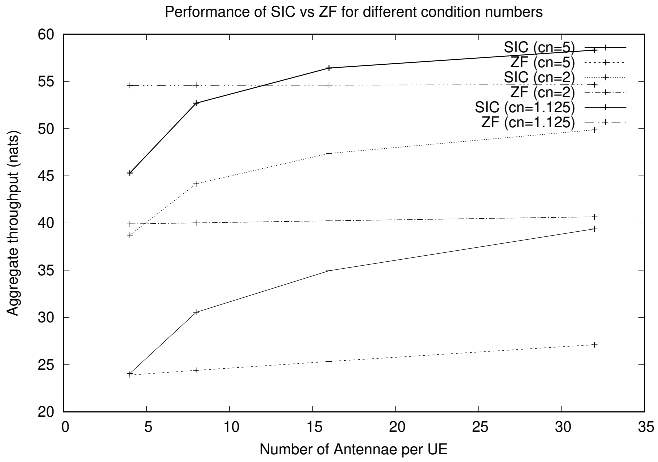

Figure 5. Performance of SIC Algorithm versus ZF with different channel conditions

Our first Figure 3 shows the expected performance of the SIC as the number of users reduce and the number of antenna per user goes up. As the users have larger and larger numbers of receive antenna, the amount of signal energy which gets converted to interference drops off asymptotically. Further, the interference per user is now spread over a larger number of streams and hence is easier to eliminate. Figure 4 shows that there is a substantial gain of the Successive Interference Compensation algorithm versus the standard zero-forcing case. As argued earlier, this is because of the very large penalty in the zero-forcing case due to full orthonormalization of all channels, whereas the SiC algorithm trades off transmit power for overlap. In the last chart, we can see that performance in the ZF case flat-lines at low

V. Conclusion

In this paper, we have proposed an algorithm for Mu-MIMO transmission over the shared channel from a single transmitter (eNodeB) to multiple UTs. The SiC algorithm is implementable at the eNodeB of a standard LTE cellular network, operating in TDD mode, using standard linear operations. We have implemented it in cloud RAN settings relatively easily, because the matrix operations are straightforward to implement and all the algorithms are linear, with no requirement for complex iterative optimization procedures. We have demonstrated its performance with respect to the existing standards of zero-forcing. The future extension of the SiC algorithm is to the CoMP case, which has to take into account the limitations of how much information can be shared between the cooperating eNodeBs. We have considered one case where we have two eNodeBs (configured in master-slave mode) and N UTs, where the slave eNodeB is dedicated to generating the interference compensation vector for the block diagonalized transmission of the master . In this case, we have to share just the interference vector (as known to the master) between the two eNodeBs. The slave, based on its own knowledge of the channel can use the interference vector can do interference cancellation. This allows us to extend the successive interference compensation to multiple eNodeBs, without requiring full CSI. This will be further explored in the future. Another area which we are pursuing is the optimal scheduling algorithm for all users, so as to guarantee minimum guaranteed QoS rates, while maintaining maximum aggregate rate capacity. This shall be published in future work.

References

[1] A. J. Paulraj, D. A. Gore, R. U. Nabar, and H. Bolcskei, “An overview of mimo communications-a key to gigabit wireless,” Proceedings of the IEEE, vol. 92, no. 2, 2004, pp. 198–218.

[2] R. W. H. Jr., T. Wu, Y. H. Kwon, and A. C. K. Soong, “Multiuser mimo in distributed antenna systems with out-of-cell interference,” IEEE Transactions on Signal Processing, vol. 59, no. 10, Oct 2011, pp. 4885–4899.

[3] Y.-N. R. Li, J. Li, W. Li, Y. Xue, and H. Wu, “Comp and interference coordination in heterogeneous network for lte-advanced,” in Globecom Workshops (GC Wkshps), 2012 IEEE, Dec 2012, pp. 1107–1111.

[4] J. Lee, J.-K. Han, and J. Zhang, “Mimo technologies in 3gpp lte and lte-advanced,” EURASIP J. Wirel. Commun. Netw., vol. 2009, Mar. 2009, pp. 3:1–3:10. [Online]. Available: http://dx.doi.org/10.1155/2009/302092

[5] M. Shafi, A. F. Molisch, P. J. Smith, T. Haustein, P. Zhu, P. De Silva, F. Tufvesson, A. Benjebbour, and G. Wunder, “5g: A tutorial overview of standards, trials, challenges, deployment, and practice,” IEEE journal on selected areas in communications, vol. 35, no. 6, 2017, pp. 1201–1221.

[6] M. H. M. Costa, “Writing on dirty paper (corresp.),” Information Theory, IEEE Transactions on, vol. 29, no. 3, May 1983, pp. 439–441.

[7] N. Zhao, F. R. Yu, M. Jin, Q. Yan, and V. C. M. Leung, “Interference alignment and its applications: A survey, research issues, and challenges,” IEEE Communications Surveys Tutorials, vol. 18, no. 3, thirdquarter 2016, pp. 1779–1803.

[8] V. R. Cadambe and S. A. Jafar, “Interference alignment and degrees of freedom of the$k$-user interference channel,” IEEE Transactions on Information Theory, vol. 54, no. 8, Aug 2008, pp. 3425–3441.

[9] M. A. Maddah-Ali, A. S. Motahari, and A. K. Khandani, “Communication over mimo x channels: Interference alignment, decomposition, and performance analysis,” IEEE Transactions on Information Theory, vol. 54, no. 8, Aug 2008, pp. 3457–3470.

[10] S. Vishwanath, N. Jindal, and A. Goldsmith, “Duality, achievable rates, and sum-rate capacity of gaussian mimo broadcast channels,” Information Theory, IEEE Transactions on, vol. 49, no. 10, Oct 2003, pp. 2658–2668.

[11] W. Yu and J. Cioffi, “Sum capacity of gaussian vector broadcast channels,” Information Theory, IEEE Transactions on, vol. 50, no. 9, Sept 2004, pp. 1875–1892.

[12] S. A. Jafar and S. Shamai, “Degrees of freedom region of the mimo$x$channel,” IEEE Transactions on Information Theory, vol. 54, no. 1, Jan 2008, pp. 151–170.

[13] M. Tomlinson, “New automatic equaliser employing modulo arithmetic,” Electronics letters, vol. 7, no. 5, 1971, pp. 138–139.

[14] H. Miyakawa and H. Harashima, “Information transmission rate in matched transmission systems with peak transmitting power limitation,” in Nat. Conf. Rec., Inst. Electron., Inform., Commun. Eng. of Japan, 1969, pp. 138–139.

[15] T. Cover, “Comments on broadcast channels,” IEEE Transactions on Information Theory, vol. 44, no. 6, Oct 1998, pp. 2524–2530.

[16] ——, “Broadcast channels,” IEEE Transactions on Information Theory, vol. 18, no. 1, Jan 1972, pp. 2–14.

[17] ——, “An achievable rate region for the broadcast channel,” IEEE Transactions on Information Theory, vol. 21, no. 4, Jul 1975, pp. 399–404.

[18] K. Marton, “A coding theorem for the discrete memoryless broadcast channel,” IEEE Transactions on Information Theory, vol. 25, no. 3, May 1979, pp. 306–311.

[19] A. E. Gamal and E. van der Meulen, “A proof of marton’s coding theorem for the discrete memoryless broadcast channel (corresp.),” IEEE Transactions on Information Theory, vol. 27, no. 1, Jan 1981, pp. 120–122.

[20] W. Yu and J. Cioffi, “Trellis precoding for the broadcast channel,” in Global Telecommunications Conference, 2001. GLOBECOM ’01. IEEE, 2001, pp. 1344–1348 vol.2.

[21] R. Wesel and J. Cioffi, “Achievable rates for tomlinson-harashima precoding,” Information Theory, IEEE Transactions on, vol. 44, no. 2, Mar 1998, pp. 824–831.

[22] K. Kusume, M. Joham, W. Utschick, and G. Bauch, “Efficient tomlinson-harashima precoding for spatial multiplexing on flat mimo channel,” in Communications, 2005. ICC 2005. 2005 IEEE International Conference on, May 2005, pp. 2021–2025 Vol. 3.

[23] M. Shenouda and T. Davidson, “A framework for designing mimo systems with decision feedback equalization or tomlinson-harashima precoding,” Selected Areas in Communications, IEEE Journal on, vol. 26, no. 2, February 2008, pp. 401–411.

[24] A. Liavas, “Tomlinson-harashima precoding with partial channel knowledge,” Communications, IEEE Transactions on, vol. 53, no. 1, Jan 2005, pp. 5–9.

[25] L. Sun and M. Lei, “Quantized csi-based tomlinson-harashima precoding in multiuser mimo systems,” Wireless Communications, IEEE Transactions on, vol. 12, no. 3, March 2013, pp. 1118–1126.

[26] W. Yu, D. Varodayan, and J. Cioffi, “Trellis and convolutional precoding for transmitter-based interference presubtraction,” Communications, IEEE Transactions on, vol. 53, no. 7, July 2005, pp. 1220–1230.

[27] S. A. Jafar and M. J. Fakhereddin, “Degrees of freedom for the mimo interference channel,” IEEE Transactions on Information Theory, vol. 53, no. 7, July 2007, pp. 2637–2642.

[28] K. Gomadam, V. R. Cadambe, and S. A. Jafar, “Approaching the capacity of wireless networks through distributed interference alignment,” in IEEE GLOBECOM 2008 - 2008 IEEE Global Telecommunications Conference, Nov 2008, pp. 1–6.

[29] Q. Spencer, A. Swindlehurst, and M. Haardt, “Zero-forcing methods for downlink spatial multiplexing in multiuser mimo channels,” Signal Processing, IEEE Transactions on, vol. 52, no. 2, Feb 2004, pp. 461–471.

[30] Q. Spencer, C. Peel, A. Swindlehurst, and M. Haardt, “An introduction to the multi-user mimo downlink,” Communications Magazine, IEEE, vol. 42, no. 10, Oct 2004, pp. 60–67.

[31] C. M. Yetis, T. Gou, S. A. Jafar, and A. H. Kayran, “On feasibility of interference alignment in mimo interference networks,” IEEE Transactions on Signal Processing, vol. 58, no. 9, Sept 2010, pp. 4771–4782.

[32] L.-U. Choi and R. D. Murch, “A transmit preprocessing technique for multiuser mimo systems using a decomposition approach,” IEEE Transactions on Wireless Communications, vol. 3, no. 1, Jan 2004, pp. 20–24.

[33] C. Wang and R. Murch, “Adaptive downlink multi-user mimo wireless systems for correlated channels with imperfect csi,” Wireless Communications, IEEE Transactions on, vol. 5, no. 9, September 2006, pp. 2435–2446.

[34] S. W. Peters and R. W. Heath, “Interference alignment via alternating minimization,” in 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, April 2009, pp. 2445–2448.

[35] V. Stankovic and M. Haardt, “Generalized design of multi-user mimo precoding matrices,” Wireless Communications, IEEE Transactions on, vol. 7, no. 3, 2008, pp. 953–961.

[36] R. Zhang, “Cooperative multi-cell block diagonalization with per-base-station power constraints,” Selected Areas of Communication, IEEE Journal on, vol. 28, no. 9, 2010, pp. 1435–1445.

[37] K.Gomadam, V. R. Cadambe, and S. A. Jafar, “A distributed numerical approach to interference alignment and applications to wireless interference networks,” IEEE Transactions on Information Theory, vol. 57, no. 6, June 2011, pp. 3309–3322.