A MMSE Vector Precoding with Block Diagonalization for Multiuser MIMO Downlink

Jungyong Park, Byungju Lee and Byonghyo Shim
School of Information and Communication, Korea University
Anam-dong, Sungbuk-gu, Seoul, Korea, 136-713
Email: {jypark, bjlee}@isl.korea.ac.kr, bshim@korea.ac.kr

***Abstract*—Block diagonalization (BD) algorithm is a generalization of the channel inversion that converts multiuser multi-input multi-output (MIMO) broadcast channel into single-user MIMO channel without inter-user interference. In this paper, we combine the BD technique with a minimum mean square error vector precoding (MMSE-VP) for achieving further gain in performance with minimal computational overhead. Two key ingredients to make our approach effective are the QR decomposition based block diagonalization and joint optimization of transmitter and receiver parameters in the MMSE sense. In fact, by optimizing precoded signal vector and perturbation vector in the transmitter and receiver jointly, we pursue an optimal balance between the residual interference mitigation and the noise enhancement suppression. From the sum rate analysis as well as the bit error rate simulations (both uncoded and coded cases) in realistic multiuser MIMO downlink, we show that the proposed BD-MVP brings substantial performance gain over existing multiuser MIMO algorithms.**

***Index Terms*—Multiuser MIMO, Block diagonalization, QR decomposition, MMSE vector precoding, Cholesky factorization.**

I. Introduction

In the multiuser multiple-input multiple-output (MIMO) downlink system, a basestation with multiple antennas transmits data streams to multiple mobile users, each with one or more receive antennas. Due to the fact that the co-channel interference is hard to be managed by the receiver operation, and more importantly, capacity of interference channel can be made equivalent to capacity without interference, pre-cancellation of the interference via the precoding at the transmitter has received much attention in recent years. It is now well-known that the capacity region of the multiuser MIMO downlink can be achieved by dirty paper coding (DPC) [1]–[8]. However, the shortcoming of the DPC is that system implementation leads to the high computational cost of successive encodings and decodings. Recently, as a way to reduce the complexity of DPC while preserving a large fraction of DPC capacity, block diagonalization (BD) has been proposed [9]. The main advantage of the BD technique is that the transmit precoding matrix assures zero inter-user interference in the received signal. Extensions of the BD, employing waterfilling or vector perturbation on top of the BD [9] [10], have also been suggested to fill the gap between the DPC and BD.

In this paper, we propose a technique that combines the QR decomposition based BD and minimum mean square error vector precoding (MMSE-VP) [11] for pursuing further gain in performance. By combining the MMSE-VP with the BD, the proposed method, henceforth referred to as the BD-MVP, mitigates noise enhancement effect, thereby achieving considerable gain in performance (both capacity and bit error rate). Unlike the conventional VP scheme that controls parameters at the transmitter only, the MMSE-VP technique jointly optimizes parameters at the transmit and receiver to pursue a balance between the residual interference mitigation and noise enhancement suppression. Moreover, due to the QR decomposition based block diagonalization, dimension of the system being processed is reduced, and hence the proposed BD-MVP brings additional benefit in complexity over the SVD based BD. In a nutshell, the proposed approach improves the quality of service (QoS) of mobile users in multiuser MIMO environments with moderate computational cost. In fact, from the sum rate analysis and BER simulation results, we show that the proposed BD-MVP outperforms the BD algorithm and its variants (BD-VP and BD-WF) [10] as well as Tomlinson-Harashima precoding (THP) based schemes (BD-GMD and BD-UCD) [12].

The rest of this paper is organized as follows. In Section II, we describe the system model and the summary of the BD technique and its extensions (BD-WF and BD-VP). In Section III, we present the proposed BD-MVP method. We provide the simulation results in Section IV and conclude in Section V.

We briefly summarize notations used in this paper. We employ uppercase boldface letters for matrices and lowercase boldface for vectors. The superscripts $(\cdot)^H$ and $(\cdot)^T$ denote conjugate transpose and transpose, respectively. diag$(\cdot)$ is a diagonal matrix where nonzero elements exist only on the main diagonal of the matrix. $\|\cdot\|$ indicates an L_2-norm of a vector. $\mathcal{CN}(m,\sigma^2)$ denotes a complex Gaussian random variable with mean m and variance σ^2.

II. Multiuser MIMO downlink

In this section, we present the system model for the multiuser MIMO downlink and then review the BD and BD-VP techniques.

This work is supported by Basic Science Research Program through National Research Foundation (NRF) funded by the Ministry of Education, Science and Technology (MEST) (No. 2011-0012525) and second Brain Korea 21 project.

A. System model

In a multiuser MIMO downlink, a basestation equipped with N_T antennas transmits information to K mobile users. Each mobile user has $N_{R,j}$ receive antennas and thus the total number of receive antennas is $N_R = \sum_{j=1}^{K} N_{R,j}$. In the sequel, we use the notation $\{N_{R,1}, \cdots, N_{R,K}\} \times N_T$ to describe this multiuser system. For each channel realization, we assume that a block of data symbols with length N_B is transmitted.

Under the flat fading channel assumption, the total received signal vector $\mathbf{y}[n] = [\mathbf{y}_1[n]^T \ \ \mathbf{y}_2[n]^T \ \cdots \ \mathbf{y}_K[n]^T]^T$ is given by

$$\mathbf{y}[n] = \mathbf{H}\tilde{\mathbf{x}}[n] + \mathbf{w}[n], \quad n = 1, 2, \cdots, N_B, \tag{1}$$

where $\mathbf{H} = [\mathbf{H}_1^T \ \mathbf{H}_2^T \ \cdots \ \mathbf{H}_K^T]^T$ is the $N_R \times N_T$ composite channel matrix and $\tilde{\mathbf{x}}[n] = \mathbf{F}\mathbf{x}[n]$ is the precoded signal vector being transmitted at the basestation where $\mathbf{F} = [\mathbf{F}_1 \ \mathbf{F}_2 \ \cdots \ \mathbf{F}_K]$ and $\mathbf{x}[n] = [\mathbf{x}_1[n]^T \ \mathbf{x}_2[n]^T \ \cdots \ \mathbf{x}_K[n]^T]^T$ are the $N_T \times N_R$ precoding matrix and the $N_R \times 1$ transmit signal vector, respectively, and $\mathbf{w}[n] = [\mathbf{w}_1[n]^T \ \mathbf{w}_2[n]^T \ \cdots \ \mathbf{w}_K[n]^T]^T$ is the complex Gaussian noise vector $\left(\mathbf{w}[n] \sim \mathcal{CN}(0, \sigma_w^2\mathbf{I})\right)$. In this setup, the received signal of the user j can be expressed as

$$\mathbf{y}_j[n] = \mathbf{H}_j\mathbf{F}_j\mathbf{x}_j[n] + \mathbf{H}_j \sum_{l=1, l\neq j}^{K} \mathbf{F}_l\mathbf{x}_l[n] + \mathbf{w}_j[n]. \tag{2}$$

Note that the first and second terms in the right side of (2) are the desired signal and multiuser interferences, respectively.

B. BD algorithm

The BD algorithm is an extension of the channel inversion for receivers with multiple antennas [9]. The key feature of the BD is to employ the precoding matrix $\mathbf{F}$ annihilating all multiuser interference. The precoding constraint for achieving this goal is

$$\mathbf{H}_i\mathbf{F}_j = \mathbf{0} \quad \text{for} \quad i = 1, \cdots, j-1, j+1, \cdots, K. \tag{3}$$

Clearly, $\mathbf{F}_j$ should be a null space of $\tilde{\mathbf{H}}_j = [\mathbf{H}_1^T \cdots \mathbf{H}_{j-1}^T \ \mathbf{H}_{j+1}^T \cdots \mathbf{H}_K^T]^T$. From the SVD of $\tilde{\mathbf{H}}_j$

$$\tilde{\mathbf{H}}_j = \tilde{\mathbf{U}}_j\tilde{\mathbf{\Sigma}}_j[\tilde{\mathbf{V}}_j^{(1)} \ \tilde{\mathbf{V}}_j^{(0)}]^H \tag{4}$$

where $\tilde{\mathbf{U}}_j$ and $\tilde{\mathbf{\Sigma}}_j$ are matrices of the left singular vectors and the ordered singular values of $\tilde{\mathbf{H}}_j$, respectively, and $\tilde{\mathbf{V}}_j^{(1)}$ and $\tilde{\mathbf{V}}_j^{(0)}$ are right singular matrices corresponding to nonzero singular values and zero singular values, respectively. It is clear that $\tilde{\mathbf{V}}_j^{(0)}$, right singular matrix corresponding to zero singular values, forms an orthogonal basis for the null space of $\tilde{\mathbf{H}}_j$. Thus, $\tilde{\mathbf{V}}_j^{(0)}$ becomes the right choice for the precoding matrix $\mathbf{F}_j$ and hence (2) becomes

$$\mathbf{y}_j[n] = \mathbf{H}_j\mathbf{F}_j\mathbf{x}_j[n] + \mathbf{w}_j[n]. \tag{5}$$

In order to maximize the achievable sum rate of the BD, the waterfilling algorithm can be additionally incorporated [13]. Since $\mathbf{HF}$ has a block diagonal structure, it is possible to perform the SVD for each user's block matrix $\mathbf{H}_j\mathbf{F}_j$ ($j = 1, \cdots, K$). Denoting this effective channel matrix as $\mathbf{H}_{\text{eff},j} = \mathbf{H}_j\mathbf{F}_j = \mathbf{H}_j\tilde{\mathbf{V}}_j^{(0)}$, the SVD of $\mathbf{H}_{\text{eff},j}$ becomes

$$\mathbf{H}_{\text{eff},j} = \mathbf{U}_j[\mathbf{\Sigma}_j \ \mathbf{0}][\mathbf{V}_j^{(1)} \ \mathbf{V}_j^{(0)}]^H. \tag{6}$$

Since each column of $\mathbf{V}_j^{(1)}$ corresponds to the singular vector of nonzero singular value in $\mathbf{\Sigma}_j$, by multiplying $\mathbf{V}_j^{(1)}$ into the right side of $\mathbf{H}_{\text{eff},j}$, we obtain $\mathbf{H}_{\text{eff},j}\mathbf{V}_j^{(1)} = \mathbf{U}_j\mathbf{\Sigma}_j$. Further, after identifying and removing $\mathbf{U}_j$[1], we can find a power loading matrix $\mathbf{D}_j$ maximizing the information rate of the user j. In summary, the achievable sum rate of the BD-WF is [9]

$$R_{\text{BD-WF}} = \max_{\mathbf{D}:\text{Tr}(\mathbf{D})\leq P_T} \log\det\left(\mathbf{I} + \frac{\mathbf{\Sigma}^2\mathbf{D}}{\sigma_w^2}\right) \tag{7}$$

where $\mathbf{D} = \text{diag}(\mathbf{D}_1, \cdots, \mathbf{D}_K)$ is determined by the water-filling under the power constraint P_T.

C. BD-VP algorithm

Although the BD-WF is optimal in maximizing the sum rate of the BD algorithm, delivery of precoding matrix index or assignment of dedicated pilot for each user to estimate the effective channel $\mathbf{H}_{\text{eff},j}$ might be a bit burdensome. The BD-VP, introduced to avoid this load, employs the vector perturbation technique [14] [15] on top of the BD to compensate for the transmit power suppression caused by the BD. Specifically, the transmitted signal of the user j is

$$\mathbf{x}_j[n] = \frac{1}{\sqrt{\gamma_j}}\mathbf{H}_{\text{eff},j}^{-1}\tilde{\mathbf{s}}_j[n] = \frac{1}{\sqrt{\gamma_j}}\mathbf{H}_{\text{eff},j}^{-1}(\mathbf{s}_j[n] + \tau\mathbf{p}_j[n]) \tag{8}$$

where $\tilde{\mathbf{s}}_j[n] = \mathbf{s}_j[n] + \tau\mathbf{p}_j[n]$ where $\mathbf{s}_j[n]$ and τ are the transmit signal vector of the user j and the pre-defined positive real number, respectively, and $\gamma_j = \|\mathbf{H}_{\text{eff},j}^{-1}(\mathbf{s}_j[n]+\tau\mathbf{p}_j[n])\|^2$.[2] Note that a choice of $\mathbf{p}_j[n]$ minimizes γ_j so that

$$\mathbf{p}_j[n] = \arg\min_{\mathbf{p}'_j[n]\in\mathbb{Z}^{N_{R,j}}} \left\|\mathbf{H}_{\text{eff},j}^{-1}(\mathbf{s}_j[n] + \tau\mathbf{p}'_j[n])\right\|^2 \tag{9}$$

where $\mathbb{Z}$ is the integer set and $N_{R,j}$ is the number of transmit streams for each user. Since the perturbation vector $\mathbf{p}_j[n]$ is chosen from the set of $N_{R,j}$-dimensional lattice points, considerable reduction of γ_j can be achieved, which helps to improve the effective SNR of the received signal. In fact, it is shown that the gap of the sum rate between the BD-VP and BD-WF is negligible in high SNR regime [10]. Note that the achievable rate of the BD-VP is given by

$$\begin{aligned} R_{\text{BD-VP}} &= \sum_{j=1}^{K}\sum_{k=1}^{N_{R,j}} \log_2(1 + \text{SNR}_{j,k}) \\ &= \sum_{j=1}^{K}\sum_{k=1}^{N_{R,j}} \left(1 + \frac{\rho\lambda_k^2}{N_{R,j}}\right) \end{aligned} \tag{10}$$

[1]Note that $\mathbf{U}_j$ can be identified once the precoded effective channel $\mathbf{H}_{\text{eff},j}$ is estimated. $\mathbf{H}_{\text{eff},j}$ can be estimated from either the dedicated pilot signal so called demodulation reference signal (DM-RS) assigned for each user or combination of the precoding matrix $\mathbf{F}_j$ and the estimated channel $\mathbf{H}_j$.

[2]In principle γ_j should be delivered to the receiver in the BD-VP technique. However, since the performance difference between γ_j and $E\gamma_j$ is negligible [15], sporadic use of $E\gamma_j$ is more beneficial in practice.

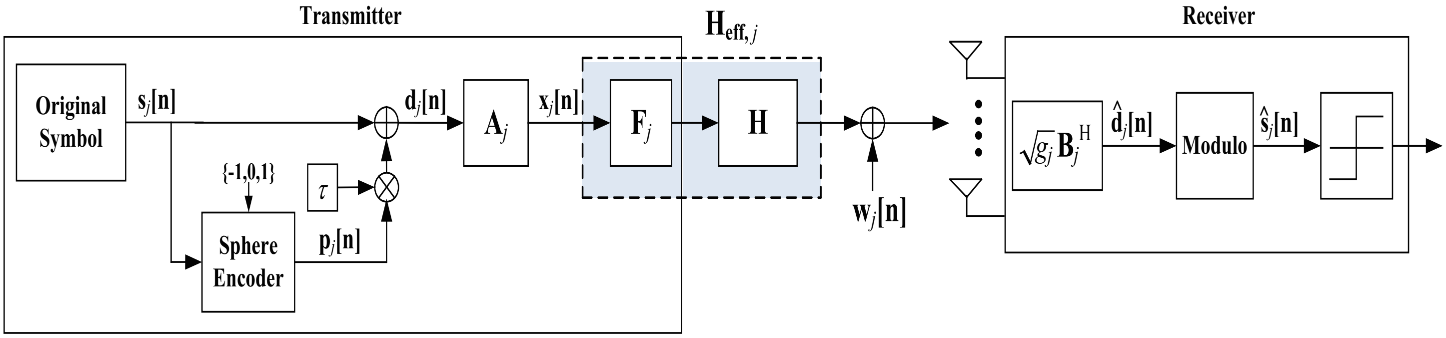

Fig. 1. The transceiver structure of the proposed BD-MVP technique.

where $\rho = \frac{P_T}{\sigma_w^2}$ and λ_k $(k = 1, \cdots, N_{R,j})$ is the singular value of $\mathbf{H}_{\text{eff},j}$.

In the implementation perspective, the integer least squares problem in (9) is processed at the transmitter and solved by a closest lattice point search, referred to as *sphere encoder* [15]. In the receiver, the received signal vector of the user j is modeled as

$$\begin{aligned}\mathbf{y}_j[n] &= \frac{1}{\sqrt{\gamma_j}}\tilde{\mathbf{s}}_j[n] + \mathbf{w}_j[n] \\ &= \frac{1}{\sqrt{\gamma_j}}(\mathbf{s}_j[n] + \tau\mathbf{p}_j[n]) + \mathbf{w}_j[n]. \qquad (11)\end{aligned}$$

In removing $\tau\mathbf{p}_j[n]$, we can employ the modulo operation and hence mod$(\sqrt{\gamma_j}\mathbf{y}_j[n], \tau)$ becomes a sufficient statistic for the symbol slicing.

III. BD-MVP

In this section, we present the proposed BD-MVP algorithm that combines the BD and MMSE-VP. While the main feature of the BD-VP is to annihilate inter-stream interference completely, the aim of the proposed BD-MVP is to allow a small inter-stream interference for optimal balance between the residual interference mitigation and noise enhancement suppression. We first provide the motivation of our work and then discuss details of the proposed algorithm.

A. The BD-MVP

Although the BD-VP outperforms the BD, due to the zero forcing nature of the algorithm, mobile suffers from the noise enhancement effect. An idea motivated by this observation is to reduce the noise enhancement effect by employing the QR decomposition based BD together with the MMSE-VP [11]. Towards this end, we pursue a tradeoff between the residual interference mitigation and noise enhancement suppression in the MMSE sense. In a nutshell, key distinctions of the proposed method over the BD-VP are that 1) elimination of multiuser interference by the QR decomposition based BD and 2) per user optimization of perturbation vector using the MMSE criterion.

First, as a computationally efficient alternative of the SVD based BD, the proposed BD-MVP exploits the QR decomposition of the pseudo inverse of the composite channel matrix $\mathbf{H}$ [16]. To be specific, while the original BD employs the SVD of $\tilde{\mathbf{H}}_j$ to find out the null space of $\tilde{\mathbf{H}}_j$, the BD-MVP exploits the pseudo inverse of $\mathbf{H}$, i.e., $\hat{\mathbf{H}} = \mathbf{H}^H(\mathbf{H}\mathbf{H}^H)^{-1}$. Denoting $\hat{\mathbf{H}} = [\hat{\mathbf{H}}_1\ \hat{\mathbf{H}}_2\ \cdots\ \hat{\mathbf{H}}_K]$, we perform the QR decomposition for each submatrix of $\hat{\mathbf{H}}$ (i.e., $\hat{\mathbf{H}}_j = \hat{\mathbf{Q}}_j\hat{\mathbf{R}}_j$). Then, it is clear that $\hat{\mathbf{Q}}_j$ forms an orthogonal basis for the null space of $\tilde{\mathbf{H}}_j$ (see Appendix A).

Second, the MMSE-VP is employed to find the optimal perturbation vector and the power constraint factor. Denoting the desired symbol as $\mathbf{d}_j[n]$ and the received symbol as $\hat{\mathbf{d}}_j[n]$ (see Fig. 1), the total MSE for whole block can be expressed as

$$\begin{aligned}&\varepsilon(\mathbf{p}_j[n], \mathbf{x}_j[n], g_j) \\ &= \sum_{n=1}^{N_B} E\left(\|\hat{\mathbf{d}}_j[n] - \mathbf{d}_j[n]\|^2\right) \\ &= \sum_{n=1}^{N_B} E\left(\|g_j\mathbf{B}_j^H(\mathbf{H}_{\text{eff},j}\mathbf{x}_j[n] + \mathbf{w}_j[n]) - \mathbf{d}_j[n]\|^2\right) \\ &= \sum_{n=1}^{N_B} E((g_j\mathbf{B}_j^H(\mathbf{H}_{\text{eff},j}\mathbf{x}_j[n] + \mathbf{w}_j[n]) - \mathbf{d}_j[n])^H \\ &\quad\times\ (g_j\mathbf{B}_j^H(\mathbf{H}_{\text{eff},j}\mathbf{x}_j[n] + \mathbf{w}_j[n]) - \mathbf{d}_j[n])) \\ &= \sum_{n=1}^{N_B}(g_j^2\mathbf{x}_j^H[n]\mathbf{H}_{\text{eff},j}^H\mathbf{H}_{\text{eff},j}\mathbf{x}_j[n] - g_j\mathbf{d}_j^H[n]\mathbf{B}_j^H\mathbf{H}_{\text{eff},j} \\ &\quad\times\ \mathbf{x}_j[n] + g_j^2\text{tr}(\mathbf{R}_n) - g_j\mathbf{x}_j^H[n]\mathbf{H}_{\text{eff},j}^H\mathbf{B}_j\mathbf{d}_j[n] \\ &\quad+\ \mathbf{d}_j^H[n]\mathbf{d}_j[n]) \qquad (12)\end{aligned}$$

where $\mathbf{B}_j^H$ is the receive equalization matrix for the user j. Note that when $\mathbf{B}_j^H = \mathbf{I}_{N_{R,j}}$, this joint optimization problem turns into the transmitter optimization problem. Since $\mathbf{H}_{\text{eff},j}, \mathbf{d}_j[n]$, and tr$(\mathbf{R}_n)$ are given, our goal is to find $\mathbf{p}_j[n]$, $\mathbf{x}_j[n]$, and g_j minimizing the MSE under the power constraint. The corresponding optimization problem can be expressed as

$$\begin{aligned}(\mathbf{p}_j[n], \mathbf{x}_j[n], g_j) &= \arg\min \varepsilon(\mathbf{p}_j[n], \mathbf{x}_j[n], g_j) \\ \text{subject to} &\quad \sum_{n=1}^{N_B}\mathbf{x}_j^H[n]\mathbf{x}_j[n] = P_T. \qquad (13)\end{aligned}$$

This problem can be solved by Lagrangian multiplier [17]. One can show that the resulting transmit signal $\mathbf{x}_j[n]$ is given by (see Appendix B)

$$\begin{aligned}\mathbf{x}_j[n] &= \mathbf{A}_j\mathbf{d}_j[n] \\ &= \frac{1}{g_j}\mathbf{H}_{\text{eff},j}^H(\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H + \xi\mathbf{I})^{-1}\mathbf{B}_j\mathbf{d}_j[n]. \qquad (14)\end{aligned}$$

Further, using the matrix inversion lemma, $\mathbf{x}_j[n] = \frac{1}{g_j}(\mathbf{H}_{\text{eff},j}^H \mathbf{H}_{\text{eff},j} + \xi\mathbf{I})^{-1}\mathbf{H}_{\text{eff},j}^H \mathbf{B}_j \mathbf{d}_j[n]$ and hence the power constraint factor g_j is given by

$$g_j = \sqrt{\frac{1}{P_T}\sum_{n=1}^{N_B} \mathbf{d}_j^H[n]\mathbf{B}_j^H \mathbf{H}_{\text{eff},j}(\mathbf{H}_{\text{eff},j}^{\dagger})^{-2}\mathbf{H}_{\text{eff},j}^H \mathbf{B}_j \mathbf{d}_j[n]} \quad (15)$$

where $\mathbf{H}_{\text{eff},j}^{\dagger} = \mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H + \xi\mathbf{I}$, $\xi = N_B \text{tr}(\mathbf{R}_n)/P_T$. Once $\mathbf{x}_j[n]$ and g_j are available, what remains is to find the optimal perturbation vector $\mathbf{p}_j[n]$. By plugging $\mathbf{x}_j[n]$ and g_j into (12), one can show that (see Appendix C)

$$\mathbf{p}_j[n] = \arg\min_{\mathbf{p}_j^{'}[n]} \|\mathbf{L}_j \mathbf{B}_j(\mathbf{s}_j[n] + \tau\mathbf{p}_j^{'}[n])\|^2 \quad (16)$$

where we employ the Cholesky factorization $(\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H + \xi\mathbf{I})^{-1} = \mathbf{L}_j^H\mathbf{L}_j$ ($\mathbf{L}_j$ is the lower triangular matrix). Similar to the BD-VP, the perturbation vector $\mathbf{p}_j[n]$ can be found by the sphere encoder.

In the receiver, the signal vector $\mathbf{y}_j[n]$ is

$$\begin{aligned} \mathbf{y}_j[n] &= \mathbf{H}_{\text{eff},j}\mathbf{x}_j[n] + \mathbf{w}_j[n] \\ &= \frac{1}{g_j}\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H(\mathbf{H}_{\text{eff},j}^{\dagger})^{-1}\mathbf{B}_j(\mathbf{s}_j[n] + \tau\mathbf{p}_j[n]) \\ &+ \mathbf{w}_j[n] \\ &\approx \frac{1}{g_j}\mathbf{B}_j(\mathbf{s}_j[n] + \tau\mathbf{p}_j[n]) + \mathbf{w}_j'[n]. \end{aligned} \quad (17)$$

Note that $\mathbf{w}_j'$ contains the residual interference introduced by the regularization factor $\xi\mathbf{I}$ as well as the Gaussian noise $\mathbf{w}_j$. Since τ is known to the receiver, the effect of $\tau\mathbf{p}_j[n]$ can be removed using a simple matrix multiplication ($g_j\mathbf{B}_j^H$) followed by the modulo operation. The proposed algorithm is summarized in Appendix D.

B. BD-MVP with geometric mean decomposition

In this subsection, we consider the optimization problem to obtain unitary receive equalization matrix $\mathbf{B}_j^H$. The geometric mean decomposition (GMD) [18] [19], developed for this purpose, partitions the channel matrix $\mathbf{H}_{\text{eff},j}$ such that

$$\mathbf{H}_{\text{eff},j} = \mathbf{Q}_j\mathbf{R}_j\mathbf{P}_j^H \quad (18)$$

where $\mathbf{Q}_j$ and $\mathbf{P}_j$ are unitary matrices and $\mathbf{R}_j$ is a real upper triangular matrix with identical diagonal elements. We can choose the unitary matrix $\mathbf{B}_j = \mathbf{Q}_j$ to equalize the effective channel of the user j. Then, the transmit signal and power constraint factor are given by

$$\mathbf{x}_j[n] = \frac{1}{g_j}\mathbf{P}_j\mathbf{R}_j^H\left(\mathbf{R}_j\mathbf{R}_j^H + \xi\mathbf{I}\right)^{-1}\mathbf{d}_j[n], \quad (19)$$

$$g_j = \sqrt{\frac{1}{P_T}\sum_{n=1}^{N_B}\|\mathbf{R}_j^H(\mathbf{R}_j\mathbf{R}_j^H + \xi\mathbf{I})^{-1}\mathbf{d}_j[n]\|^2}, \quad (20)$$

and the total MSE becomes

$$\varepsilon = \xi\sum_{n=1}^{N_B}\mathbf{d}_j^H[n]\left(\mathbf{R}_j\mathbf{R}_j^H + \xi\mathbf{I}\right)^{-1}\mathbf{d}_j[n]. \quad (21)$$

Using the Cholesky factorization $\left(\mathbf{R}_j\mathbf{R}_j^H + \xi\mathbf{I}\right)^{-1} = \mathbf{T}_j^H\mathbf{T}_j$, (21) can be simplified to

$$\varepsilon = \xi\sum_{n=1}^{N_B}\|\mathbf{T}_j\left(\mathbf{s}_j[n] + \tau\mathbf{p}_j[n]\right)\|^2. \quad (22)$$

In order to minimize the total MSE, each independent term should be minimized and hence $\mathbf{p}_j[n]$ is chosen as

$$\mathbf{p}_j[n] = \arg\min_{\mathbf{p}_j^{'}[n]}\|\mathbf{T}_j(\mathbf{s}_j[n] + \tau\mathbf{p}_j^{'}[n])\|^2. \quad (23)$$

C. The Sum Rate of the BD-MVP

The sum rate of the proposed method is equivalent to the sum of single user rate by virtue of the fact that the precoding matrix $\mathbf{F}_j$ annihilates multiuser interference. Recalling that the transmit and received signal vectors of the user j are $\mathbf{x}_j[n] = \frac{1}{\sqrt{g_j}}\mathbf{H}_{\text{eff},j}^H(\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H + \xi\mathbf{I})^{-1}\mathbf{B}_j\mathbf{d}_j[n]$ and $\mathbf{y}_j[n] = \mathbf{H}_{\text{eff},j}\mathbf{x}_j[n] + \mathbf{w}_j[n]$, respectively, and also using the eigenvalue decomposition $\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H = \mathbf{Q}\Lambda\mathbf{Q}^H$, $\mathbf{H}_{\text{eff},j}\mathbf{x}_j[n]$ becomes

$$\mathbf{H}_{\text{eff},j}\mathbf{x}_j[n] = \frac{1}{\sqrt{g_j}}\mathbf{Q}\mathbf{\Phi}\mathbf{Q}\mathbf{B}_j\mathbf{d}_j[n] \quad (24)$$

where $\mathbf{Q}$ is the unitary matrix and $\mathbf{\Phi}$ is the diagonal matrix whose jth element is $\frac{\lambda_j}{\lambda_j+\xi}$ (λ_j is the jth diagonal element of Λ).

In (24), the kth entry of the user j is

$$[\mathbf{H}_{\text{eff},j}\mathbf{x}_j[n]]_k = \frac{1}{\sqrt{g_j}}\left[q_{k,1}\frac{\lambda_1}{\lambda_1+\xi} \cdots q_{k,N_{R,j}}\frac{\lambda_{N_{R,j}}}{\lambda_{N_{R,j}}+\xi}\right] \times \begin{bmatrix} q_{1,1}^H & \cdots & q_{N_{R,j},1}^H \\ \vdots & \cdots & \vdots \\ q_{1,N_{R,j}}^H & \cdots & q_{N_{R,j},N_{R,j}}^H \end{bmatrix}\begin{bmatrix} \mathbf{c}_{j,1}[n] \\ \vdots \\ \mathbf{c}_{j,N_{R,j}}[n] \end{bmatrix} \quad (25)$$

where $q_{m,n}$ is the (m,n)th entry of $\mathbf{Q}$ and $\mathbf{c}_{j,i}[n]$ is the ith entry of $\mathbf{B}_j\mathbf{d}_j[n]$. In (25), we see that $\frac{1}{\sqrt{g_j}}\left(\sum_{l=1}^{N_{R,j}}\frac{\lambda_l}{\lambda_l+\xi}|q_{k,l}|^2\right)\mathbf{c}_{j,k}[n]$ is the desired term and all of the rest (containing $\mathbf{c}_{j,l}[n], l \neq k$) are interferences. In this perspective, kth received signal for the user j can be expressed as

$$\mathbf{y}_{j,k}[n] = \frac{1}{\sqrt{g_j}}\left(\sum_{l=1}^{N_{R,j}}\frac{\lambda_l}{\lambda_l+\xi}|q_{k,l}|^2\right)\mathbf{c}_{j,k}[n] + \mathbf{z}_{j,k}[n] \quad (26)$$

where $\mathbf{z}_{j,k}[n] = \frac{1}{\sqrt{g_j}}\sum_{m=1,m\neq k}^{N_{R,j}}\left(\sum_{l=1}^{N_{R,j}}\frac{\lambda_l}{\lambda_l+\xi}q_{k,l}q_{m,l}^H\right) \times \mathbf{c}_{j,m}[n] + \mathbf{w}_{j,k}[n]$ contains the inter-stream interference and additive noise $\mathbf{w}_{j,k}$. Therefore, the SINR of each stream of the user j becomes

$$\text{SINR}_{j,k} = \sum_{n=1}^{N_B}\frac{\|s\mathbf{c}_{j,k}[n]\|^2}{\|t\mathbf{c}_{j,m}[n]\|^2 + \|\mathbf{w}_{j,k}[n]\|^2} \quad (27)$$

where $s = \frac{1}{\sqrt{g_j}}\left(\sum_{l=1}^{N_{R,j}}\frac{\lambda_l}{\lambda_l+\xi}|q_{k,l}|^2\right)$ and $t = \frac{1}{\sqrt{g_j}}\sum_{m=1,m\neq k}^{N_{R,j}}\left(\sum_{l=1}^{N_{R,j}}\frac{\lambda_l}{\lambda_l+\xi}q_{k,l}q_{m,l}^H\right)$ and the corresponding sum rate becomes

$$R_{\text{BD-MVP}} = \sum_{j=1}^{K}\sum_{k=1}^{N_{R,j}}\log_2\left(1 + \text{SINR}_{j,k}\right). \quad (28)$$

TABLE I
COMPLEXITY COMPARISON BETWEEN THE BD-VP AND BD-MVP

Operations	BD-VP	BD-MVP
Annihilation of multiuser interference	BD is done by the SVD $C_{\text{SVD}} = K\left(4N_T^2 N_R + 13(N_R - N_{R,j})^3\right)$	BD is achieved by the QR decomposition $C_{\text{QR}} = \frac{11}{3}N_T^3 + \frac{5}{3}N_T^2 + KN_{R,j}^2(N_T - \frac{1}{3}N_{R,j})$
Computation of perturbation vector	SE employing zero forcing criterion with complexity $C_{\text{VP-SE}}$	SE exploiting MMSE criterion with complexity $C_{\text{MVP-SE}}$ ($\approx C_{\text{VP-SE}}$)
Receiver equalizer operation	-	$\mathbf{B}_j^H$ multiplication is required per each user
Modulo operation	Remove $\tau\mathbf{p_j}$ to map the received signal into the original vector	Same as BD-VP

D. MSE Performance between the BD-VP and BD-MVP

In this subsection, we provide the MSE analysis of the BD-VP and BD-MVP algorithms. Since the optimization problem of the conventional VP can be interpreted as a zero forcing ($\xi = 0$) MMSE-VP problem [11], we can compare the MSE of two schemes. Moreover, due to the fact that the minimum MSE of the BD-VP and BD-MVP equals γ_j and g_j, respectively, the MSE of an instantaneous block for both γ_j and g_j can be compared without loss of generality. First, γ_j of the BD-VP algorithm is expressed as

$$\begin{aligned}\gamma_j &\triangleq \|\mathbf{H}_{\text{eff},j}^H(\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H)^{-1}(\mathbf{s}_j + \tau\mathbf{p}_j)\|^2\\ &= (\mathbf{s}_j + \tau\mathbf{p}_j)^H(\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H)^{-1}(\mathbf{s}_j + \tau\mathbf{p}_j)\\ &= (\mathbf{s}_j + \tau\mathbf{p}_j)^H\mathbf{Q}\Lambda^{-1}\mathbf{Q}^H(\mathbf{s}_j + \tau\mathbf{p}_j)\\ &= \sum_{l=1}^{N_{R,j}}\frac{1}{\lambda_j}|\mathbf{q}_l^H(\mathbf{s}_j + \tau\mathbf{p}_j)|^2. \end{aligned} \tag{29}$$

In a similar way, g_j of the BD-MVP scheme is

$$\begin{aligned}g_j &\triangleq \|\mathbf{L}_j(\mathbf{s}_j + \tau\mathbf{p}_j)\|^2\\ &= (\mathbf{s}_j + \tau\mathbf{p}_j)^H(\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H + \xi\mathbf{I})^{-1}(\mathbf{s}_j + \tau\mathbf{p}_j)\\ &= (\mathbf{s}_j + \tau\mathbf{p}_j)^H\mathbf{Q}\left(\Lambda + \xi\mathbf{I}\right)^{-1}\mathbf{Q}^H(\mathbf{s}_j + \tau\mathbf{p}_j)\\ &= \sum_{l=1}^{N_{R,j}}\frac{1}{\lambda_j + \xi}|\mathbf{q}_l^H(\mathbf{s}_j + \tau\mathbf{p}_j)|^2. \end{aligned} \tag{30}$$

Due to the inclusion of $\xi(>0)$ in the denominator, it is clear that

$$\gamma_j - g_j > 0. \tag{31}$$

Note that since the total MSE of the BD-VP is the sum of each individual block MSE ($\text{MSE}_{\text{BD-VP}} = \sum_{j=1}^K \gamma_j$) and the same is true for the BD-MVP ($\text{MSE}_{\text{BD-MVP}} = \sum_{j=1}^K g_j$), we conclude that the BD-MVP outperforms the BD-VP in the MSE sense. Note also that we only consider the case where $\mathbf{B}_j^H = \mathbf{I}_{N_{R,j}}$ in g_j computation so that further improvement in the MSE can be achieved using the GMD based receive equalization.

E. Comments on Complexity

In this subsection, we discuss the complexity of the BD-MVP and BD-VP. Since the receiver operation of the BD-VP and BD-MVP (the modulo operation and receive equalization) is very simple to analyze and in fact trivial, we focus only on the computations associated with transmitter precoding. In our analysis, we measure the complexity of the algorithm by the number of floating point operations (flops). The major operations of the proposed method and BD-VP include 1) annihilation of multiuser interference using $\mathbf{F}_j$ and 2) per user optimization of perturbation vector $\mathbf{p}_j$.

The computational complexity of the SVD for the BD-VP C_{SVD} and QR decomposition for the BD-MVP C_{QR} are [20]

- $C_{\text{SVD}} : K \times \left(4N_T^2 N_R + 13(N_R - N_{R,j})^3\right)$.
- $C_{\text{QR}} : \frac{11}{3}N_T^3 + \frac{5}{3}N_T^2$ for channel inversion and $K \times N_{R,j}^2 \times (N_T - \frac{1}{3}N_{R,j})$ for orthogonalization.

Comparing to the SVD based BD, the QR based BD has smaller computational complexity since the QR based orthogonalization of $N_T \times N_{R,j}$ matrix $\hat{\mathbf{H}}_j$ is simpler than the SVD operation on $(N_R - N_{R,j}) \times N_T$ matrix $\tilde{\mathbf{H}}_j$. For example, the required flops of the BD-VP and BD-MVP for the $\{4, 4\} \times 8$ system are 2688 and 2137, respectively. In fact, computational benefit of the BD-MVP over the BD-VP grows with N_T and K mainly because the SVD is operated in a higher dimension.

Next, we investigate the complexity of perturbation vector computation. Due to the fact that the vector precoding is implemented using the sphere encoding, the computational complexity associated with this operation is nondeterministic [21]. The lower bound on the complexity of the sphere encoding C_{SE} is given by [22]

$$C_{\text{SE}} \geq \frac{\beta^{\nu N_s} - 1}{\sqrt{\beta} - 1} \tag{32}$$

where β is the modulation order, N_s is the number of search dimension, and ν is the complexity component given by $\nu = \frac{1}{2}\left(1 + \frac{4(\beta-1)}{3\omega^2}\text{SNR}\right)$ where ω is a constant. Noting that the dominant factor of C_{SE} is N_s and both schemes have the same N_s, we expect that the complexity of the BD-VP and BD-MVP would be more or less similar. We summarize the complexity analysis of this subsection in Table I.

IV. SIMULATION RESULTS AND DISCUSSIONS

In this section, we compare the performance of the proposed BD-MVP method with the BD, BD-VP, BD-WF, as well as THP based schemes (BD-GMD and BD-UCD). While the BD-GMD generates subchannels with identical SNRs using zero forcing based the GMD, the BD-UCD employs MMSE extension of uniform channel decomposition (UCD) for obtaining further gain in performance (readers are referred to [12] for more details).

The simulation setup is based on the 16-QAM transmission. We assume that elements of each user's channel matrix are i.i.d. zero mean complex Gaussian random variables with unit variance, which almost surely ensures that each user's channel

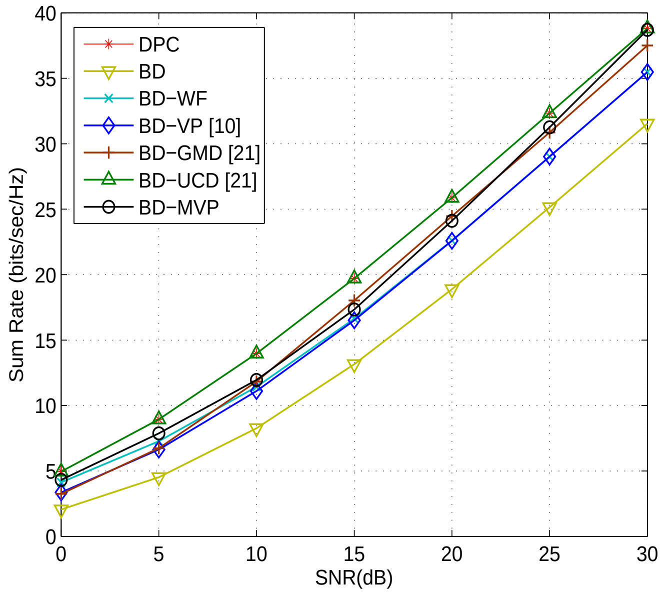

Fig. 2. Achievable sum rates of DPC, conventional BD algorithms (BD, BD-WF and BD-VP), THP based schemes (BD-GMD and BD-UCD), and the proposed BD-MVP.

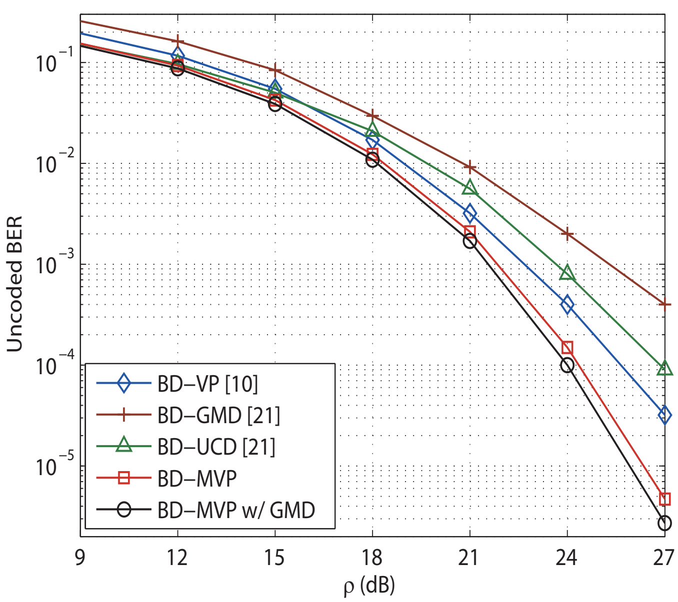

Fig. 3. Uncoded BER performance for $\{4,4\}\times 8$ multiuser MIMO systems.

matrix has full rank ($\text{rank}(\mathbf{H}_j) = \min(N_{R,j}, N_T)$). In order to observe the comprehensive picture of theoretical and practical performance, we measure the sum rate and bit error rate (BER) (both uncoded and coded BER), respectively.

In Fig. 2, we plot the sum rate for the $\{2,2\}\times 4$ system as a function of SNR. Note that the achievable rate of the DPC is [3]–[5]

$$R_{\text{DPC}} = \sup_{\mathbf{D}\in\mathcal{A}} \log|\mathbf{I} + \mathbf{H}^H\mathbf{D}\mathbf{H}| \tag{33}$$

where $\mathcal{A}$ is the set of nonnegative diagonal matrices $\mathbf{D}$ satisfying $\text{tr}(\mathbf{D}) \le P_T$. Generally, we observe that the sum rate of the BD-UCD, BD-GMD, BD-WF, and BD-MVP is not much different from the sum rate of the DPC. In particular, the sum rate of the BD-MVP becomes close to the sum rate of the DPC as the SNR increases. However, the BD algorithm leaves nonnegligible gap to the sum rate of the DPC.

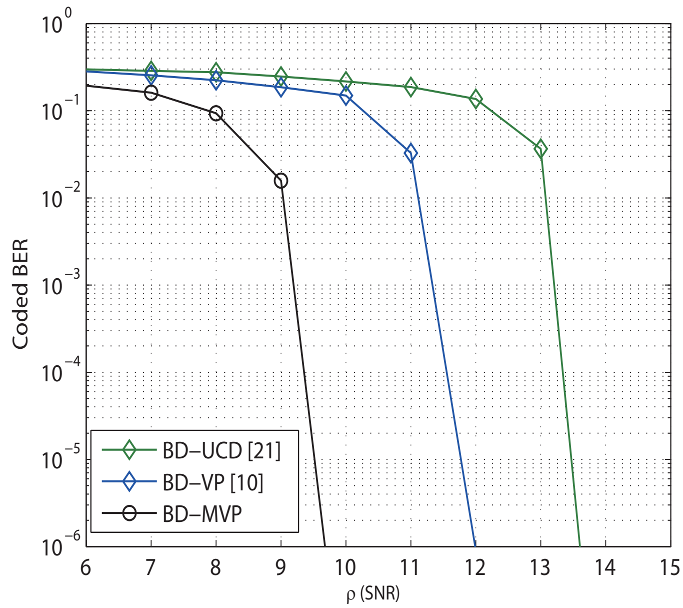

Fig. 4. Coded BER performance for $\{4,4\}\times 8$ systems with 16-QAM modulation (Turbo encoding with rate r = 1/2).

In Fig. 3, we plot the uncoded BER for the $\{4,4\}\times 8$ systems. We observe that the BD-MVP outperforms the BD-VP and BD-UCD resulting in about 0.8 dB and 1.5 dB gain at BER $= 10^{-3}$, respectively. We also observe that the BD-MVP with receive equalization provides additional gain (around 0.3 dB at BER $= 10^{-3}$) over the BD without equalization.

In order to assess the computational complexity of algorithms under test, we examine the running time of each scheme using same set of symbols under same environments. The program is coded with the Matlab program and the results are obtained as an average time (in seconds) for 10^4 symbols running under a duo-core processor with Windows 7 environment. As summarized in Table II, the BD-GMD achieves the lowest complexity among all candidates by virtue of the fact that small number of computations is needed for the THP modulo operation. Due to the considerable number of computations associated with the sphere encoding, we see that the running time of the BD-MVP and BD-VP is 3.3 and 3.9 times higher than that of the BD-GMD. Since number of iterations for performing waterfilling is substantial, BD-UCD achieves the highest computational complexity.

Finally, in order to observe if this gain can be transferred to the gain after decoding, we check the coded BER performance. Fig. 4 illustrates the coded BER performance of the proposed method with half-rate ($r = 1/2$) turbo encoded data. The standard Gray mapping is used to map the information bits to 16-QAM symbols. As an encoder, we use a universal mobile telecommunications systems (UMTS) standard with feedforward polynomial $1+D+D^3$ and feedback polynomial

TABLE II
TIME COMPLEXITY FOR $\{4,4\} \times 8$ SYSTEM (RUNNING TIME FOR 10^4 SYMBOLS)

Algorithm	BD-MVP	BD-VP	BD-GMD	BD-UCD
Running time (sec)	4.0531	4.8308	1.2125	19.2539

$1+D^2+D^3$ [23]. The size of codebook is set to 4000. Similar to [15], $\mathbf{y}'_{j,re}[n] = \mathbf{y}_{j,re}[n] - (\tau/g_j)\mathbf{p}_{j,re}[n]$ is employed to compute the likelihood function

$$\begin{aligned} p(\mathbf{y}'_{j,re}[n]|\mathbf{s}_{j,re}) &= \sum_{m=-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \\ &\times \exp\{-(\mathbf{y}'_{j,re}[n] - (\mathbf{s}_{j,re} - m\tau)/g_j)^2/2\sigma^2\} \\ & \mathbf{y}'_{j,re}[n] \in \left[-\frac{\tau}{2g_j}, \frac{\tau}{2g_j}\right). \end{aligned} \tag{34}$$

The imaginary part is handled in the same way. In our simulation, we replace the infinite sum in (34) with a sum of a few terms on both sides of $m = 0$ since the performance loss over the setup using infinite sum is negligible. From the Fig. 4, we observe that the proposed method outperforms the BD-VP and BD-UCD. In particular, gain of the BD-MVP over the BD-VP and BD-UCD at waterfall region is around 2 dB and 4 dB, respectively. Since the performance requirement of coded system typically lies on the waterfall regime, the proposed method can bring a substantial reduction on the transmitter power over the BD-UCD and BD-VP. Clearly, we conclude that the proposed scheme is competitive in both practical and theoretical perspectives.

V. CONCLUSION

In this paper, we investigated an approach based on block diagonalization (BD) achieving robustness of multiuser MIMO downlink. Motivated by the fact that mobile users suffer from noise enhancement effect due to the zero forcing nature of the BD algorithm, we combined the BD with the MMSE-VP to achieve the balance between residual interference mitigation and noise enhancement suppression. We could observe from the sum rate analysis as well as practical BER performance (both coded and uncoded) that the proposed BD-MVP approach uniformly outperforms the BD algorithm and its variants. Future study needs to be directed towards the investigation of the performance when the channel state information at the transmitter (CSIT) is imperfect.

APPENDIX A
QR DECOMPOSITION BASED BD

The pseudo inverse of $N_{R,j} \times N_T$ composite channel matrix $\mathbf{H} = \left[\mathbf{H}_1^T \ \mathbf{H}_2^T \ \cdots \ \mathbf{H}_K^T\right]^T$ is $\hat{\mathbf{H}} = \mathbf{H}^H \left(\mathbf{H}\mathbf{H}^H\right)^{-1} = \left[\hat{\mathbf{H}}_1 \ \hat{\mathbf{H}}_2 \ \cdots \ \hat{\mathbf{H}}_K\right]$ where $\hat{\mathbf{H}}_j$ is $N_T \times N_{R,j}$ matrix. Then, one can show that

$$\begin{aligned} \mathbf{H}\hat{\mathbf{H}} &= \begin{bmatrix} \mathbf{H}_1 \\ \vdots \\ \mathbf{H}_K \end{bmatrix} \begin{bmatrix} \hat{\mathbf{H}}_1 & \cdots & \hat{\mathbf{H}}_K \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{H}_1\hat{\mathbf{H}}_1 & \cdots & \mathbf{H}_1\hat{\mathbf{H}}_K \\ & \vdots & \\ \mathbf{H}_K\hat{\mathbf{H}}_1 & \cdots & \mathbf{H}_K\hat{\mathbf{H}}_K \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{I}_{N_{R,1}} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{I}_{N_{R,K}} \end{bmatrix} = \mathbf{I}_{N_R}. \end{aligned} \tag{35, 36}$$

Clearly, $\mathbf{H}_j\hat{\mathbf{H}}_k = 0$ when $j \neq k$. By defining $\tilde{\mathbf{H}}_j = \left[\mathbf{H}_1^T \ \cdots \ \mathbf{H}_{j-1}^T \ \mathbf{H}_{j+1}^T \cdots \ \mathbf{H}_K^T\right]^T$, the zero inter-user interference constraint is satisfied as $\tilde{\mathbf{H}}_j\hat{\mathbf{H}}_j = \mathbf{0}$. Now consider the QR decomposition of $\hat{\mathbf{H}}_j$

$$\hat{\mathbf{H}}_j = \hat{\mathbf{Q}}_j\hat{\mathbf{R}}_j \quad \text{for} \quad j = 1, \cdots, K \tag{37}$$

where $\hat{\mathbf{R}}_j$ is $N_{R,j} \times N_{R,j}$ upper triangular matrix and $\hat{\mathbf{Q}}_j$ is $N_T \times N_{R,j}$ matrix whose columns form an orthonormal basis for $\hat{\mathbf{H}}_j$. From the zero inter-user interference constraint, we have $\tilde{\mathbf{H}}_j\hat{\mathbf{Q}}_j\hat{\mathbf{R}}_j = \mathbf{0}$. Since $\hat{\mathbf{R}}_j$ is invertible, it follows that $\tilde{\mathbf{H}}_j\hat{\mathbf{Q}}_j = \mathbf{0}$.

APPENDIX B
PROOF OF JOINT OPTIMIZATION PROBLEM

The Lagrangian function to solve the joint optimization problem is

$$\begin{aligned} & f(\mathbf{p}_j[n], \mathbf{x}_j[n], g_j, \lambda) \\ &= \varepsilon(\mathbf{p}_j[n], \mathbf{x}_j[n], g_j) + \lambda(\sum_{n=1}^{N_B} \mathbf{x}_j^H[n]\mathbf{x}_j[n] - P_T). \end{aligned} \tag{38}$$

By equating the derivatives of $f(\cdot)$ with respect to $\mathbf{x}_j[n]$, g_j and λ to zero, we can obtain the following equation

$$\begin{aligned} \frac{\partial f(\cdot)}{\partial \mathbf{x}_j[n]} &= g_j^2\mathbf{x}_j^H[n]\mathbf{H}_{\text{eff},j}^H\mathbf{H}_{\text{eff},j} - g_j\mathbf{d}_j^H[n]\mathbf{B}_j^H\mathbf{H}_{\text{eff},j} \\ &+ \lambda\mathbf{x}_j^H[n] = 0, \end{aligned} \tag{39}$$

$$\begin{aligned} \frac{\partial f(\cdot)}{\partial g_j} &= \sum_{n=1}^{N_B}\{2g_j\mathbf{x}_j^H[n]\mathbf{H}_{\text{eff},j}^H\mathbf{H}_{\text{eff},j}\mathbf{x}_j[n] - \mathbf{d}_j^H[n]\mathbf{B}_j^H \\ & \times \mathbf{H}_{\text{eff},j}\mathbf{x}_j[n] + 2\mathbf{g}_j\text{tr}(\mathbf{R}_n) - \mathbf{x}_j^H[n]\mathbf{H}_{\text{eff},j}^H\mathbf{B}_j \\ & \times \mathbf{d}_j[n]\} = 0, \end{aligned} \tag{40}$$

$$\frac{\partial f(\cdot)}{\partial \lambda} = \sum_{n=1}^{N_B}\mathbf{x}_j^H[n]\mathbf{x}_j[n] - P_T = 0. \tag{41}$$

Using the matrix inversion lemma, (39) becomes

$$\begin{aligned} \mathbf{x}_j[n] &= \frac{1}{g_j}(\mathbf{H}_{\text{eff},j}^H\mathbf{H}_{\text{eff},j} + \frac{\lambda}{g_j^2}\mathbf{I})^{-1}\mathbf{H}_{\text{eff},j}^H\mathbf{B}_j\mathbf{d}_j[n] \\ &= \frac{1}{g_j}\mathbf{H}_{\text{eff},j}^H(\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H + \frac{\lambda}{g_j^2}\mathbf{I})^{-1}\mathbf{B}_j\mathbf{d}_j[n]. \end{aligned} \tag{42}$$

By multiplying $\frac{\mathbf{x}_j[n]}{g_j}$ at the right side of (39), we have

$$\begin{aligned} & \mathbf{d}_j^H[n]\mathbf{B}_j^H\mathbf{H}_{\text{eff},j}\mathbf{x}_j[n] \\ &= g_j\mathbf{x}_j^H[n]\mathbf{H}_{\text{eff},j}^H\mathbf{H}_{\text{eff},j}\mathbf{x}_j[n] + \frac{\lambda}{g_j}\mathbf{x}_j^H[n]\mathbf{x}_j[n]. \end{aligned} \tag{43}$$

Plugging (43) into (40), we have

$$\sum_{n=1}^{N_B}\big(-2\frac{\lambda}{g_j}\mathbf{x}_j^H[n]\mathbf{x}_j[n]+2g_j\mathrm{tr}(\mathbf{R}_n)\big)=0. \tag{44}$$

After some manipulation, we have

$$\frac{\lambda}{g_j^2}=\frac{\mathrm{tr}(\mathbf{R}_n)}{\frac{1}{N_B}\sum_{n=1}^{N_B}\mathbf{x}_j^H[n]\mathbf{x}_j[n]}=\frac{N_B\mathrm{tr}(\mathbf{R}_n)}{P_T}. \tag{45}$$

Let $\xi=\frac{N_B tr(\mathbf{R}_n)}{P_T}$, then (42) can be rewritten as

$$\begin{aligned}\mathbf{x}_j[n] &= \frac{1}{g_j}(\mathbf{H}_{\text{eff},j}^H\mathbf{H}_{\text{eff},j}+\xi\mathbf{I})^{-1}\mathbf{H}_{\text{eff},j}^H\mathbf{B}_j\mathbf{d}_j[n]\\ &= \frac{1}{g_j}\mathbf{H}_{\text{eff},j}^H(\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H+\xi\mathbf{I})^{-1}\mathbf{B}_j\mathbf{d}_j[n]. \end{aligned}\tag{46}$$

Appendix C
Proof of (16)

Plugging $\mathbf{x}_j[n]$ in (14) and g_j in (15) into (12) and after some manipulations, we have

$$\begin{aligned}&\varepsilon(\mathbf{p}_j[n]|\mathbf{x}_j[n],g_j)\\
=\ &\sum_{n=1}^{N_B}(g_j^2\mathbf{x}_j^H[n]\mathbf{H}_{\text{eff},j}^H\mathbf{H}_{\text{eff},j}\mathbf{x}_j[n]-g_j\mathbf{d}_j^H[n]\mathbf{B}_j^H\mathbf{H}_{\text{eff},j}\mathbf{x}_j[n]\\
&\quad+\ g_j^2\mathrm{tr}(\mathbf{R}_n)-g_j\mathbf{x}_j^H[n]\mathbf{H}_{\text{eff},j}^H\mathbf{B}_j\mathbf{d}_j[n]+\mathbf{d}_j^H[n]\mathbf{d}_j[n])\\
=\ &\sum_{n=1}^{N_B}(\mathbf{d}_j^H[n]\mathbf{B}_j^H(\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H+\xi\mathbf{I})^{-1}\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H\\
&\quad\times(\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H+\xi\mathbf{I})^{-1}\mathbf{B}_j\mathbf{d}_j[n]+\mathbf{d}_j^H[n]\mathbf{d}_j[n]\\
&\quad-\mathbf{d}_j^H[n]\mathbf{B}_j^H\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H(\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H+\xi\mathbf{I})^{-1}\mathbf{B}_j\mathbf{d}_j[n]\\
&\quad-\mathbf{d}_j^H[n]\mathbf{B}_j^H(\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H+\xi\mathbf{I})^{-1}\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H\mathbf{B}_j\mathbf{d}_j[n]\\
&\quad+\xi\mathbf{d}_j^H[n]\mathbf{B}_j^H(\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H+\xi\mathbf{I})^{-1}\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H(\mathbf{H}_{\text{eff},j}\\
&\quad\times\mathbf{H}_{\text{eff},j}^H+\xi\mathbf{I})^{-1}\mathbf{B}_j\mathbf{d}_j[n])\\
=\ &\sum_{n=1}^{N_B}\mathbf{d}_j^H[n]\mathbf{Q}_j^H((\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H+\xi\mathbf{I})^{-1}\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H\mathbf{H}_{\text{eff},j}\\
&\quad\times\mathbf{H}_{\text{eff},j}^H(\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H+\xi\mathbf{I})^{-1}+\mathbf{I}-\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H\\
&\quad\times(\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H+\xi\mathbf{I})^{-1}-(\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H+\xi\mathbf{I})^{-1}\mathbf{H}_{\text{eff},j}\\
&\quad\times\mathbf{H}_{\text{eff},j}^H+\xi(\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H+\xi\mathbf{I})^{-1}\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H(\mathbf{H}_{\text{eff},j}\\
&\quad\times\mathbf{H}_{\text{eff},j}^H+\xi\mathbf{I})^{-1})\mathbf{B}_j\mathbf{d}_j[n]\\
=\ &\sum_{n=1}^{N_B}\mathbf{d}_j^H[n]\mathbf{B}_j^H((\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H+\xi\mathbf{I})^{-1}\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H\\
&\quad\times(\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H+\xi\mathbf{I})^{-1}-\mathbf{I}+\mathbf{I}-\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H\\
&\quad\times(\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H+\xi\mathbf{I})^{-1}+\mathbf{I}-(\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H+\xi\mathbf{I})^{-1}\mathbf{H}_{\text{eff},j}\\
&\quad\times\mathbf{H}_{\text{eff},j}^H+\xi(\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H+\xi\mathbf{I})^{-1}\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H(\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H\\
&\quad+\xi\mathbf{I})^{-1})\mathbf{B}_j\mathbf{d}_j[n]\\
=\ &\sum_{n=1}^{N_B}\mathbf{d}_j^H[n]\mathbf{B}_j^H((\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H+\xi\mathbf{I})^{-1}\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H\\
&\quad\times(\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H+\xi\mathbf{I})^{-1}-\mathbf{I}+2\xi(\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H+\xi\mathbf{I})^{-1}\\
&\quad+\xi(\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H+\xi\mathbf{I})^{-1}\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H(\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H\\
&\quad+\xi\mathbf{I})^{-1})\mathbf{B}_j\mathbf{d}_j[n]\\
=\ &\sum_{n=1}^{N_B}\mathbf{d}_j^H[n]\mathbf{B}_j^H(\xi(\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H+\xi\mathbf{I})^{-1})\mathbf{B}_j\mathbf{d}_j[n],\end{aligned}\tag{47}$$

where $\mathbf{I}-(\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H+\xi\mathbf{I})^{-1}\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H=\xi(\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H+\xi\mathbf{I})^{-1}$ follows from the matrix inversion lemma. Employing the Cholesky factorization $(\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H+\xi\mathbf{I})^{-1}=\mathbf{L}_j^H\mathbf{L}_j$ and also noting that $\mathbf{d}_j[n]=\mathbf{s}_j[n]+\tau\mathbf{p}_j[n]$, (47) becomes

$$\begin{aligned}\varepsilon(\mathbf{p}_j[n]|\mathbf{x}_j[n],g_j) &= \xi\sum_{n=1}^{N_B}\mathbf{d}_j^H[n]\mathbf{B}_j^H\mathbf{L}_j^H\mathbf{L}_j\mathbf{B}_j\mathbf{d}_j[n]\\ &= \xi\sum_{n=1}^{N_B}\|\mathbf{L}_j\mathbf{B}_j(\mathbf{s}_j[n]+\tau\mathbf{p}_j[n])\|^2.\end{aligned}\tag{48}$$

In summary, the optimization problem to find the perturbation vector $\mathbf{p}_j$ is expressed as

$$\mathbf{p}_j[n]=\arg\min_{\mathbf{p}_j^{'}[n]}\|\mathbf{L}_j\mathbf{B}_j(\mathbf{s}_j[n]+\tau\mathbf{p}_j^{'}[n])\|^2. \tag{49}$$

Appendix D
Summary of the BD-MVP Algorithm

- Input : $\mathbf{H},\mathbf{s}[n]$
- Output : $\mathbf{p}[n]$
- Variable
  1) j denotes the jth user being examined
  2) n denotes the nth block being examined

Step 1 : (QR based BD step)

In order to compute the nullspace of $\tilde{\mathbf{H}}_j$, we define the pseudo-inverse of the channel matrix $\mathbf{H}$ as $\hat{\mathbf{H}}=\mathbf{H}^H(\mathbf{H}\mathbf{H}^H)^{-1}=[\hat{\mathbf{H}}_1\cdots\hat{\mathbf{H}}_j\cdots\hat{\mathbf{H}}_K]$

for $j=1:K$

$\hat{\mathbf{H}}_j=\hat{\mathbf{Q}}_j\hat{\mathbf{R}}_j$

$\mathbf{F}_j=\hat{\mathbf{Q}}_j$

$\mathbf{H}_{\text{eff},j}=\mathbf{H}_j\mathbf{F}_j$

end

Step 2 : (MMSE-VP step) for $j=1:K$

factorize $(\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H+\xi\mathbf{I}_{n_j})^{-1}=\mathbf{L}_j^H\mathbf{L}_j$

for $n=1,\ldots,N_B$

$\mathbf{p}_j[n]=\arg\min_{\mathbf{p}_j^{'}[n]}\|\mathbf{L}_j\mathbf{B}_j(\mathbf{s}_j[n]+\tau\mathbf{p}_j^{'}[n])\|^2$

$\mathbf{x}_j[n]=\mathbf{H}_{\text{eff},j}^H(\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H+\xi\mathbf{I}_{n_j})^{-1}\mathbf{B}_j\mathbf{d}_j[n]$

$=\mathbf{H}_{\text{eff},j}^H(\mathbf{H}_{\text{eff},j}\mathbf{H}_{\text{eff},j}^H+\xi\mathbf{I}_{n_j})^{-1}\mathbf{B}_j(\mathbf{s}_j[n]+\tau\mathbf{p}_j[n])$

end

$g_j=\sqrt{\frac{1}{P_T}\sum_{n=1}^{N_B}\mathbf{d}_j^H[n]\mathbf{B}_j^H\mathbf{H}_{\text{eff},j}(\mathbf{H}_{\text{eff},j}^{\dagger})^{-2}\mathbf{H}_{\text{eff},j}^H\mathbf{B}_j\mathbf{d}_j[n]}$

for $n=1,\ldots,N_B$

$\mathbf{x}_j[n]=\frac{1}{g_j}\mathbf{x}_j[n]$

end

end

References

[1] I. E. Telatar, "Capacity of multi-antenna Gaussian channels," *Eur. Trans. Telecommun.*, vol. 10, pp. 585-595, Nov./Dec. 1999.

[2] H. Weingarten, Y. Steinberg, and S. Shamai, "The capacity region of the Gaussian MIMO broadcast channel," *IEEE Trans. Inf. Theory*, vol. 52, pp. 3936-3964, Sept. 2006.

[3] S. Vishwanath, N. Jindal, and A. Goldsmith, "Duality, achievable rates, and sum-rate capacity of Gaussian MIMO broadcast channels," *IEEE Trans. Inf. Theory*, vol. 49, pp. 2658-2668, Oct. 2003.

[4] W. Yu and J. M. Cioffi, "Sum capacity of Gaussian vector broadcast channels," *IEEE Trans. Inf. Theory*, vol. 50, pp. 1875-1892, Sept. 2004.

[5] P. Viswanath and D. N. C. Tse, "Sum capacity of the vector gaussian broadcast channel and uplink downlink duality," *IEEE Trans. Inf. Theory*, vol. 49, pp. 1912-1921, Aug. 2003.

[6] G. Caire and S. Shamai, “On the achievable throughput of a multiantenna Gaussian broadcast channel,” *IEEE Trans. Inf. Theory*, vol. 49, pp. 1691-1706, July 2003.

[7] S. I. Gelfand and M. S. Pinsker, “Coding for channel with side information,” *Problemi Peredachi Informatsii*, vol. 9, no. 1, pp. 19.31, 1980.

[8] M. Costa, “Writing on dirty paper,” *IEEE Trans. Inf. Theory*, vol. IT-29, pp. 439-441, May 1983.

[9] Q. H. Spencer, A. L. Swindlehurst, and M. Haardt, “Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels,” *IEEE Trans. Signal Process.*, vol. 52, pp. 461-471, Feb. 2004.

[10] C. B. Chae, S. Shim, and R. W. Heath Jr., “Block Diagonalized Vector Perturbation for Multiuser MIMO Systems,” *IEEE Trans. Wireless Comm.*, vol. 7, pp. 4051-4057, Nov. 2008.

[11] D. Schmidt, M. Joham, W. Utschick, “Minimum mean square error vector precoding,” *Eur. Trans. Telecommun.*, vol. 19, pp. 219-231, March 2007.

[12] S. Lin, W. L. Ho, and Y. C. Liang, “Block diagonal geometric mean decomposition (BD-GMD) for MIMO broadcast channels,” *IEEE Trans. Wireless Comm.*, vol. 7, no. 7, pp. 2778-2789, Jul. 2008.

[13] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, 1991.

[14] C. B. Peel, B. M. Hochwald, and A. L. Swindlehurst, “A vector-perturbation technique for near-capacity multiantenna multiuser communication - part I: channel inversion and regularization,” *IEEE Trans. Comm.*, vol. 53, pp. 195-202, Jan. 2005.

[15] B. M. Hochwald, C. B. Peel, and A. L. Swindlehurst, “A vector-perturbation technique for near-capacity multiantenna multiuser communication - part II: perturbation,” *IEEE Trans. Comm.*, vol. 53, pp. 537-544, March 2005.

[16] H. Sung, S. Lee, and I. Lee, “Generalized Channel Inversion Methods for Mutliuser MIMO Systems,” *IEEE Trans. Comm.*, vol. 57, pp. 3489-3994, Nov. 2009.

[17] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge Univ. Press, 2004.

[18] Y. Jiang, J. Li and W. W. Hager, “Joint transceiver design for MIMO communications using geometric mean decomposition,” *IEEE Trans. Signal Process.*, vol. 53, no. 10, pp. 3791-3803, Oct. 2005.

[19] F. Liu, L. Jiang, and C. He, “Joint MMSE vector precoding based on GMD method for MIMO systems,” *IEICE Trans. Comm.*, vol. 90, no. 9, pp.2617-2620, Sept. 2007.

[20] G. H. Golub and C. F. V. Loan, *Matrix Computations*, 3rd ed. The Johns Hopkins University Press, 1989.

[21] B. Shim and I. Kang, “Sphere decoding with a probabilistic tree pruning,” *IEEE Trans. Signal Process.*, vol. 56, no. 10, pp. 4867-4878, Oct. 2008.

[22] J. Jalden and B. Ottersten, “On the complexity of sphere decoding in digital communication,” *IEEE Trans. Signal Process.*, vol. 53, no. 4, pp. 1474-1484, Apr. 2005.

[23] “3rd Generation Partnership Project: Technical specification group radio access network; physical layer (Release 7),” Tech. Spec. 25.201 V7.2.0, 2007. (see http://www.3gpp.org/ftp/Specs/archive/25_series/25.201/25201-720.zip).

PLACE PHOTO HERE

Jungyong Park (S’ 09) received the B.S. degree in information and communication engineering from Dongguk University, Seoul, Korea in 2008, and the M.S. degree in the School of Information and Communication from Korea University, Seoul, Korea in 2010, where he is currently working toward the Ph.D. degree. His research interests include wireless communications and information theory.

PLACE PHOTO HERE

Byungju Lee (S’ 09) received the B.S. and M.S. degrees in the School of Information and Communication from Korea University, Seoul, Korea, in 2008 and 2010, where he is currently working toward the Ph.D. degree. His research interests include wireless communications and information theory.

PLACE PHOTO HERE

Byonghyo Shim (SM’ 09) received the B.S. and M.S. degrees in control and instrumentation engineering (currently electrical engineering) from Seoul National University, Korea, in 1995 and 1997, respectively and the M.S. degree in Mathematics and the Ph.D. degree in Electrical and Computer Engineering from the University of Illinois at Urbana-Champaign, in 2004 and 2005, respectively.

From 1997 and 2000, he was with the Department of Electronics Engineering at the Korean Air Force Academy as an Officer (First Lieutenant) and an Academic Full-time Instructor. From 2005 to 2007, he was with the Qualcomm Inc., San Diego, CA, working on CDMA systems with the emphasis on the next-generation UMTS receiver design. Since September 2007, he has been with the School of Information and Communication, Korea University, where he is now an Associate Professor. His research interests include signal processing for wireless communications, compressive sensing, applied linear algebra, and information theory.

Dr. Shim was the recipient of 2005 M. E. Van Valkenburg research award from ECE department of University of Illinois and 2010 Haedong award from IEEK.