IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 27, 2017, accepted May 25, 2017, date of publication July 11, 2017, date of current version August 14, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2725919

INVITED PAPER

Massive MIMO, Non-Orthogonal Multiple Access and Interleave Division Multiple Access

CHONGBIN XU[1,2], YANG HU[3], CHULONG LIANG[3], JUNJIE MA[4], AND LI PING[3], (Fellow, IEEE)
[1]Key Laboratory for Information Science of Electromagnetic Waves (MoE), the Department of Communication Science and Engineering, Fudan University, Shanghai 200433, China
[2]National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China
[3]Department of Electronic Engineering, City University of Hong Kong, Hong Kong
[4]Department of Statistics, Columbia University, New York, NY 10027-6902 USA

Corresponding author: Yang Hu (yhu228-c@my.cityu.edu.hk)

This work was supported in part by the National Natural Science Foundation of China under Grant 61501123, in part by the Open Research Fund of the National Mobile Communications Research Laboratory, Southeast University, under Grant 2015D06, in part by the University Grants Committee of the Hong Kong Special Administrative Region, China, under Project AoE/E-02/08, Project CityU 11208114, Project CityU 11217515, and Project CityU 11280216, and in part by the European Union through ICT projects under Project ICT-217033 WHERE and Project ICT-248894 WHERE2.

ABSTRACT This paper provides an overview on the rationales in incorporating massive multiple-input multiple-output (MIMO), non-orthogonal multiple access (NOMA), and interleave division multiple access (IDMA) in a unified framework. Our emphasis is on multi-user gain that refers to the advantage of allowing multi-user transmission in massive MIMO. Such a gain can potentially offer tens or even hundreds of times of rate increase. The main difficulty in achieving multi-user gain is the reliance on accurate channel state information (CSI) in the existing schemes. With accurate CSI, both OMA and NOMA can deliver performance not far away from capacity. Without accurate CSI, however, most of the existing schemes do not work well. We outline a solution to this difficulty based on IDMA and iterative data-aided channel estimation (DACE). This scheme can offer very high throughput and is robust against the pilot contamination problem. The receiver cost is low, since only maximum ratio combining (MRC) is involved and there is no matrix inversion or decomposition. Under time division duplex, accurate CSI acquired in the up-link can be used to support low-cost down-link solutions, such as zero forcing. These findings offer useful design considerations for future systems.

INDEX TERMS Massive MIMO, NOMA, IDMA, iterative MRC and DACE.

I. INTRODUCTION

Multiple-input multiple-output (MIMO) is a wireless technology employing multiple transmit and receive antennas [1]–[11]. Massive MIMO refers to the situation when the number of antennas involved is very large [12]–[18]. A typical massive MIMO setting is unbalanced, with far more antennas at a base station (BS) than those at a mobile terminal (MT) due to different physical sizes. Massive MIMO provides abundant spatial diversity. How to make best use of such diversity with low cost is a research topic of important practical implications. Channel state information (CSI) plays an important role in massive MIMO. CSI quality can be affected by, e.g., channel estimation error [19], channel variation and RF calibration error [20]. The correlation among different pilots also results in the so-called pilot contamination problem [12], [13], [21], [22]. CSI errors may seriously affect performance in massive MIMO.

Conventional orthogonal multiple access (OMA) schemes, such as frequency division multiple access (FDMA) and time division multiple access (TDMA), rely on orthogonality in either frequency or time to avoid interference. In MIMO, orthogonality can also be established in space via zero forcing (ZF) [2]. In the absence of interference, simple single-user detection (SUD) is sufficient in OMA.

Non-orthogonal multiple access (NOMA) refers to multi-user transmission schemes that allow interference among users. The concept of NOMA with power control (PC) and successive interference cancelation (SIC) was introduced in [23]. It was shown in [24] that PC-SIC is asymptotically capacity approaching in MIMO up-links when user number is sufficiently large. Recently, NOMA based on PC-SIC has been widely discussed to improve the fairness issue [25]–[32].

In general, MIMO capacity can only be achieved by NOMA but not by OMA [4]–[7]. Multi-user detection (MUD) [2], [24], [33]–[39] is required for this purpose. Therefore NOMA has a theoretical advantage. The difference, though, can be minor under perfect CSI in practical cellular environments when resource allocation over rate, power, time and frequency is applied in OMA. (See [2], [40] and Section III below.)

Direct-sequence code division multiple access (DS-CDMA) [41]–[43] and interleave division multiple access (IDMA) [44]–[60] are two realization techniques for NOMA. DS-CDMA relies on user-specific spreading sequences for multiple access. Spreading incurs rate loss, which is not preferred in high rate applications. IDMA overcomes the problem by employing user-specific interleaving for multiple access [44]–[60], which does not incur rate loss. Inspired by turbo and low-density parity-check (LDPC) codes [61]–[64], IDMA was originally introduced using a sparse graphic representation [65]. IDMA is applicable in very high rate applications [66], [67].

Various options involving the above mentioned concepts have been recently discussed for the 5th generation (5G) cellular systems [68]–[75]. It has been generally agreed that massive MIMO is promising in significantly enhancing throughput. Most works on NOMA are for small MIMO systems [25]–[32]. It is not yet clear how to efficiently integrate NOMA with massive MIMO. The advantages and disadvantages of various options are still heavily debated issues.

This paper provides an overview on the rationales in incorporating massive MIMO, NOMA and IDMA in a unified framework. We will first compare various options. We focus on multi-user gain [24] that refers to the advantage of allowing multiple users to transmit simultaneously at the same time-frequency resource block. Potentially, multi-user gain can lead to immense rate increase, but its reliance on CSI is an obstacle. With accurate CSI, both OMA and NOMA can perform not far away from capacity. Without accurate CSI, however, most existing schemes do not work well.

We will outline a solution based on time division duplex (TDD) and iterative processing. Only very coarse CSI for the up-link, that can be obtained using pilots, is required at the beginning. Such CSI may not be sufficient to establish reliable spatial orthogonality, and hence the up-link has to be NOMA. IDMA offers a simple option for this purpose. An iterative maximum ratio combining (MRC) and data aided channel estimation (DACE) [19], [22], [45], [76]–[78] technique is used to refine CSI and data estimates gradually. This I-MRC-DACE technique has the following features.

- It is robust against pilot contamination [22] and can achieve very high CSI accuracy [45].
- It can deliver drastically increased throughput.
- The cost involved is low. No matrix inversion or decomposition is involved.

Under TDD, CSI acquired from the up-link can be used in the down-link. Then ZF with resource allocation is a simple and efficient option for the down-link. NOMA may squeeze out more gain if higher complexity at MTs allows, but the extra benefit is limited in cellular environments.

In the above scheme, reliable up-link channel estimation holds the key to the overall performance, despite the fact that the down-link traffic is usually more demanding. We will provide numerical results to support the above claims. We will demonstrate the simplicity and efficiency of IDMA, compared with other more sophisticated signaling/detection techniques.

For convenience, we will use the following notations throughout this paper:

$N_{\rm BS}$ number of antennas at a BS,
$N_{\rm MT}$ number of antennas at an MT, and
K number of concurrently transmitting users at the same time and on the same frequency.[1]

We will mostly discuss the up-link and rely on duality for down-link performance [2], [5], [79]. We will discuss single-cell systems from Sections II to IV. Multi-cell systems will be discussed in Section V.

II. CHANNEL MODEL

Massive MIMO systems can be divided into millimeter-wave [80]–[86] and sub-millimeter wave ones [12]–[18]. Channel modelings are different in these two cases. Millimeter-wave systems are primarily for indoor applications [85], [86]. In this paper, for simplicity and aiming at general cellular applications, we will focus on conventional sub-millimeter modeling. However, most discussions below, in particular multi-user gain, can be extended to millimeter-wave systems.

A. UP-LINK CHANNEL

We start from the up-link. Assume underlying orthogonal frequency division multiplexing (OFDM) operations, so that inter-symbol interference (ISI) can be ignored. For simplicity we will assume $N_{\rm MT} = 1$.

We write the received signal (over a particular OFDM subcarrier) in a multi-user MIMO up-link system at time j as [2]

$$\boldsymbol{y}(j) = \sum_{k=1}^{K} \boldsymbol{h}_k x_k(j) + \boldsymbol{\eta}(j), \tag{1}$$

where $\boldsymbol{y}(j)$ is an $N_{\rm BS} \times 1$ signal vector received at BS antennas, $\boldsymbol{h}_k$ an $N_{\rm BS} \times 1$ channel coefficient vector, $x_k(j)$ a symbol transmitted from the kth user, and $\boldsymbol{\eta}(j)$ an $N_{\rm BS} \times 1$ vector of complex additive white Gaussian noise (AWGN) with mean 0 and variance $\sigma^2 = N_0/2$ per dimension. Eqn. (1) can be rewritten into a more compact form as

$$\boldsymbol{y}(j) = \boldsymbol{H}\boldsymbol{x}(j) + \boldsymbol{\eta}(j), \tag{2a}$$

with

$$\boldsymbol{H} \equiv [\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_k, \ldots, \boldsymbol{h}_K], \tag{2b}$$

$$\boldsymbol{x}(j) \equiv [x_1(j), x_2(j), \ldots, x_k(j), \ldots, x_K(j)]^T. \tag{2c}$$

The (n, k)th entry of $\boldsymbol{H}$ is denoted as $H_{n,k}$:

$$H_{n,k} = (\boldsymbol{h}_k)_n, \tag{2d}$$

[1] These K users may or may not interfere each other, depending on transmitter and receiver structures. ZF may eliminate interference. NOMA in general involves interference. Also note that, when resource allocation is applied, some users may be allocated to zero rate.

where $(\boldsymbol{h}_k)_n$ is the nth entry of $\boldsymbol{h}_k$. It is the channel coefficient between the nth BS antenna and MT k.

B. CHANNEL GAIN AND ANGLE

A mobile channel generally experiences both slow (including lognormal fading and path loss) and fast fading (i.e., Rayleigh fading). We model $H_{n,k}$ as [2]

$$H_{n,k} = H_k^{\text{slow}} \cdot H_{n,k}^{\text{fast}}, \tag{3a}$$

$$H_k^{\text{slow}} = H_k^{\text{lognormal}} \cdot H_k^{\text{path loss}}, \tag{3b}$$

where $H_k^{\text{lognormal}}$, $H_k^{\text{path loss}}$ and $H_{n,k}^{\text{fast}}$ are for, respectively, lognormal fading, path loss and Rayleigh fading. The followings are assumed.

- $H_{n,k}^{\text{fast}}$ is independent and identically distributed (i.i.d.) for every (n, k) pair.
- $H_k^{\text{lognormal}}$ and $H_k^{\text{path loss}}$ are i.i.d. over k and invariant over n for a fixed k.

The following normalizations are adopted

$$\mathrm{E}\left(\left|H_k^{\text{lognormal}}\right|^2\right) = \mathrm{E}\left(\left|H_k^{\text{path loss}}\right|^2\right) = \mathrm{E}\left(\left|H_{n,k}^{\text{fast}}\right|^2\right) = 1. \tag{4a}$$

For large N_{BS}, following the law of large numbers and from (4a), we have

$$\frac{1}{N_{\text{BS}}} \sum_{n=1}^{N_{\text{BS}}} \left|H_{n,k}^{\text{fast}}\right|^2 \approx 1. \tag{4b}$$

We call $||\boldsymbol{h}_k||^2$ channel gain and

$$\boldsymbol{\phi}_k = \boldsymbol{h}_k / ||\boldsymbol{h}_k|| \tag{5a}$$

channel angle. From (3a) and (4b),

$$||\boldsymbol{h}_k||^2 \approx \left|H_k^{\text{slow}}\right|^2 N_{\text{BS}}. \tag{5b}$$

Property 1: In a massive MIMO system with the assumptions stated earlier, channel gain $||\boldsymbol{h}_k||^2$ is approximately determined by slow fading and channel angle $\boldsymbol{\phi}_k$ by fast fading.

Consider a special case when only user k is allowed to transmit. Using (5b) and as $\det(\boldsymbol{I}_{m\times m}+\boldsymbol{AB}) = \det(\boldsymbol{I}_{n\times n}+\boldsymbol{BA})$ where $\boldsymbol{I}_{m\times m}$ and $\boldsymbol{I}_{n\times n}$ are unit matrices with proper sizes, we have the channel capacity for user k [2]

$$\begin{aligned} r_k^{\text{single-user}} &= \log_2\left(\det\left(\boldsymbol{I} + (N_0)^{-1}\boldsymbol{h}_k\boldsymbol{h}_k^H p_k\right)\right) \\ &= \log_2\left(1 + (N_0)^{-1}||\boldsymbol{h}_k||^2 p_k\right) \\ &\approx \log_2\left(1 + (N_0)^{-1}\left|H_k^{\text{slow}}\right|^2 N_{\text{BS}} p_k\right), \end{aligned} \tag{6}$$

when $N_{\text{BS}} \to \infty$. Here p_k is the transmission power of user k. We refer to (6) as single-user capacity. The approximation in (6) results from the channel hardening effect [87], [88] and greatly simplifies the analysis problem.

Property 2: When N_{BS} is large, single-user capacity is approximately independent of fast fading and is determined only by slow fading.

The above says that the effect of fast fading is averaged out in capacity evaluation. This should be distinguished from the fact that the knowledge on fast fading is required in the realization of MIMO capacity, as discussed below.

C. CONSTRAINTS

Denote by r_k and $p_k = \mathrm{E}(|x_k(j)|^2)$ the rate and transmission power of user k, respectively. The sum transmission rate and power of all users are denoted, respectively, by

$$R_{\text{sum}} = \sum_{k=1}^{K} r_k \quad \text{and} \quad P_{\text{sum}} = \sum_{k=1}^{K} p_k. \tag{7a}$$

For a complex channel, the sum signal to noise ratio (SNR) is defined as

$$SNR_{\text{sum}} = P_{\text{sum}}/N_0. \tag{7b}$$

We will consider the following two types of constraints for system design.

- Equal-power constraint (EPC): In this case, $p_k = P_{\text{sum}}/K$ for all k. We optimize r_k to maximize R_{sum}.
- Sum-power constraint (SPC): We optimize p_k and r_k together to maximize R_{sum}.

Alternatively, we can also maximize the proportional fairness criterion sum-log-rate $\sum \log(r_k)$ under SPC [2], although we will only briefly discuss this issue in this paper.

D. MULTI-CELL SYSTEMS

The above SNR_{sum} is for a single-cell system. A multi-cell system can be characterized by a signal to interference and noise ratio (SINR). Suppose that cross-cell interference behaves in the same way as additive noise. Then a multi-cell system with a given SINR is equivalent to a single-cell one with a matching SNR. Because of this equivalence, for simplicity of discussions, we will focus on single-cell systems below. (Also see footnote 2.)

In Section V-C, we will see that, due to the pilot contamination problem, cross-cell interference actually cannot be treated in the same way as additive noise. We will discuss the consequence of this problem and a potential solution using DACE.

E. DOWN-LINK

According to the duality principle [5], [79], the down-link performance can be predicted from the up-link counterpart in many cases. This greatly saves our effort, as will be explained later.

III. PERFORMANCE WITH MULTI-USER CONCURRENT TRANSMISSION

In this section, we compare different transmission strategies in massive MIMO. We will show that allowing more users to transmit concurrently can significantly enhance sum-rate. We will assess the impact of CSI errors on this issue for different realization techniques.

A. GAIN FROM MULTI-USER CONCURRENT TRANSMISSION

Massive MIMO potentially provides two types of gain:

- power gain from beamforming (BF), and
- rate gain through exploiting spatial diversity.

To benefit from BF, single-user transmission suffices. Power gain via BF can also be translated into rate gain. The latter grows logarithmically with N_{BS}, which is relatively slow.

The potential sum-rate gain offered by spatial diversity is much more impressive. This can be achieved by allowing more users to transmit simultaneously over the same time and same frequency. The related advantage is referred to as multi-user gain. Sum-rate increases with N_{BS} much faster in this way. It can be shown that, for a fixed K/N_{BS} ratio, sum-rate grows linearly with N_{BS}.

Multi-user spatial diversity has been discussed in [2]. The emphasis in [2] is on opportunistic beamforming and scheduling in SISO and small MIMO, where single-user capacity is affected by fast fading. Scheduling is an effective way to benefit from channel fluctuation in this case.

For massive MIMO, however, single-user capacity is approximately determined by slow fading only (See Property 2 in Section II-B). Scheduling over slow fading may result in a serious delay problem. Therefore, instead of scheduling, our emphasis is on multi-user concurrent transmission with a large K.

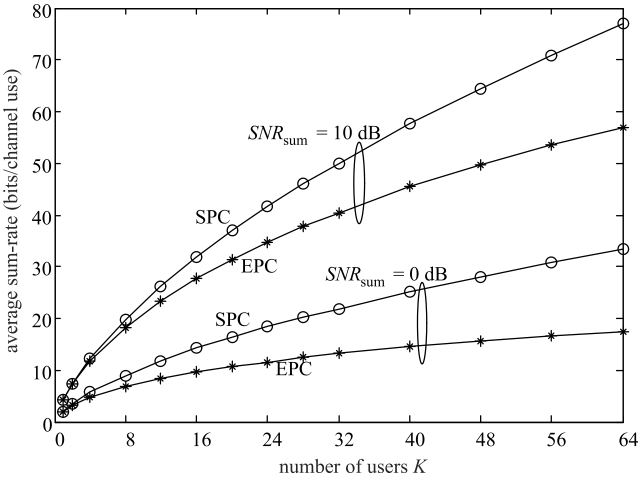


FIGURE 1. Sum-rate capacity with EPC and SPC. $N_{\mathrm{BS}} = 64$ and $N_{\mathrm{MT}} = 1$. Both slow and fast fading factors are included. Path loss is based on a hexagon cell with a normalized side length = 1. The minimum normalized distance between users and the BS is 35/289, corresponding to an unnormalized distance of 35m for a LTE cell with radius 289m. The loss factor = −3.76 and lognormal fading deviation = 8 dB. The channel samples are normalized such that the average power gain = 1. No power control over slow fading.

A large K implies high transmitter and receiver costs. We need to examine carefully whether the potential benefits can justify the cost. For this purpose, Fig. 1 shows the potential sum-rate capacity gain for a single-cell system. Perfect CSI at receiver (CSIR) and perfect CSI at transmitter (CSIT) are assumed in Fig. 1. The curves apply to both up- and down-links following the duality principle [2], [5], [79]. For capacity computation, see [8].

We can make the following observations from Fig. 1:

- Multi-user gain is seen from the growth of sum-rate with K. The potential gain is in the order of tens of times for both EPC and SPC.
- The difference between EPC and SPC is initially small but becomes noticeable when K is very large. This indicates the importance of resource allocation when K is large.

Fig. 1 shows that multi-user gain is very attractive. Diversifying power over more users, i.e., increasing K, is a very effective way to increase sum-rate. Motivated by this, in what follows, we will discuss efficient realization techniques for multi-user gain at affordable cost.

B. ANGLE AND NEAR-FAR DIVERSITIES

Spatial diversity can be further divided into angle and near-far diversities that refer to, respectively, the variations in $\{\boldsymbol{\phi}_k\}$ and $\{||\boldsymbol{h}_k||^2\}$. The related advantages are referred to as angle gain and near-far gain respectively.

Recall that the size of $\boldsymbol{H}$ in (2) is $N_{\mathrm{BS}} \times K$ and the angles $\{\boldsymbol{\phi}_k\}$ are normalized columns of $\boldsymbol{H}$. Assume that $\{\boldsymbol{\phi}_k\}$ are i.i.d. (See Section II-B). When $K \ll N_{\mathrm{BS}}$, $\{\boldsymbol{\phi}_k\}$ are almost orthogonal to each other. This effectively results in K parallel interference-free channels. Increasing K can thus increase system sum-rate, which leads to angle gain.

Angle diversity alone does not fully explain multi-user gain. For example, it is impossible to have more than N_{BS} orthogonal angles when $K > N_{\mathrm{BS}}$. The variation in $\{||\boldsymbol{h}_k||^2\}$ due to slow-fading also contributes. This is referred to as near-far diversity [2], [23], [24], [89], although it can be caused by both block fading and path loss in massive MIMO.

Roughly speaking, angle gain can be from the EPC curves and near-far gain from the difference between a pair of SPC and EPC curves with the same SNR_{sum} in Fig. 1. For SPC with high SNR, angle gain dominates the overall gain.

Angle gain is available only in MIMO and not in SISO. On the other hand, near-far gain is available in both MIMO and SISO. Angle and near-far gains can be achieved by both OMA and NOMA. We will discuss various options in the following subsections.

C. ZF, MRC AND MRC-SIC

Fig. 1 is for capacity analysis. We now turn attention to practical realizations. ZF and MRC are two common options.

A ZF estimator is given by [2]:

$$\hat{\boldsymbol{x}}(j) = \left(\boldsymbol{H}^H \boldsymbol{H}\right)^{-1} \boldsymbol{H}^H \boldsymbol{y}(j). \tag{8a}$$

Substituting (2a) into (8a) and letting $\boldsymbol{\xi}(j) = (\boldsymbol{H}^H\boldsymbol{H})^{-1}\boldsymbol{H}^H \boldsymbol{\eta}(j)$, we have

$$\hat{\boldsymbol{x}}(j) = \boldsymbol{x}(j) + \boldsymbol{\xi}(j). \tag{8b}$$

Clearly, $\hat{\boldsymbol{x}}(j)$ can be used to estimate $\boldsymbol{x}(j)$. With ZF, different users are divided into different orthogonal subspaces, which avoids interference and provides angle gain. When $\boldsymbol{H}^H\boldsymbol{H}$ is ill-conditioned, the noise term $\boldsymbol{\xi}(j) = (\boldsymbol{H}^H\boldsymbol{H})^{-1}\boldsymbol{H}^H\boldsymbol{\eta}(j)$ in (8) suffers from an amplification effect [2]. Water-filling over different orthogonal subspaces can be used to alleviate the problem, which also provides near-far gain.

An MRC estimator [2] is defined in a symbol-by-symbol form as

$$\hat{x}_k(j) = \boldsymbol{h}_k^H\boldsymbol{y}(j). \tag{9a}$$

Substituting (1) into (9a),

$$\hat{x}_k(j) = \lambda_k x_k(j) + \xi_k(j), \tag{9b}$$

where $\lambda_k \equiv \|\boldsymbol{h}_k\|^2 = \boldsymbol{h}_k^H\boldsymbol{h}_k$ is a scalar and

$$\xi_k(j) \equiv \sum_{k'=1,k'\neq k}^{K} \boldsymbol{h}_k^H\boldsymbol{h}_{k'}x_{k'}(j) + \boldsymbol{h}_k^H\boldsymbol{\eta}(j) \tag{9c}$$

is an interference (from $x_{k'}(j)$, $k' \neq k$ to $x_k(j)$) plus noise term. MRC does not involve matrix inversion and so has lower cost than ZF. However, interference is a problem for MRC, especially when K is large.

MRC-SIC [24] retains the low cost of MRC while suppresses interference. We illustrate its principle using $K = 2$. Assume that the signal of user 1 is decoded first. In this case, from (9) we have

$$\hat{x}_1(j) = \lambda_1 x_1(j) + \xi_1(j), \tag{10a}$$

$$\xi_1(j) = \boldsymbol{h}_1^H\boldsymbol{h}_2 x_2(j) + \boldsymbol{h}_1^H\boldsymbol{\eta}(j). \tag{10b}$$

We treat $\xi_1(j)$ as a Gaussian additive noise. The achievable rate of user 1 based on (10) is

$$r_1 = \log_2\left(1 + \frac{p_1\|\boldsymbol{h}_1\|^2}{p_2\left\|\boldsymbol{h}_1^H\boldsymbol{h}_2\right\|^2/\|\boldsymbol{h}_1\|^2 + N_0}\right). \tag{11a}$$

After successfully decoding $x_1(j)$, its interference is subtracted from $\boldsymbol{y}(j)$. We then decode $x_2(j)$ with achievable rate

$$r_2 = \log_2\left(1 + \frac{p_2\|\boldsymbol{h}_2\|^2}{N_0}\right). \tag{11b}$$

Based on the above, we have

$$R_{\text{sum}} = \log_2\left(1 + \frac{p_1\|\boldsymbol{h}_1\|^2}{p_2\left\|\boldsymbol{h}_1^H\boldsymbol{h}_2\right\|^2/\|\boldsymbol{h}_1\|^2 + N_0}\right) + \log_2\left(1 + \frac{p_2\|\boldsymbol{h}_2\|^2}{N_0}\right). \tag{11c}$$

Angle gain can be seen from the correlation term $\left\|\boldsymbol{h}_1^H\boldsymbol{h}_2\right\|^2$ in (11c) which represents interference. Its impact reduces statistically when N_{BS} increases.

Near-far gain stems from the difference in $\|\boldsymbol{h}_k\|^2$ in (11c) that allows room for optimization. For example, we can adjust p_1 and p_2 to minimize P_{sum} when r_1 and r_2 are fixed [24], [90], [91] or to maximize R_{sum} when r_1 and r_2 are adjustable [2], [5], [92], or alternately, to maximize sum log-rate $\log(r_1) + \log(r_2)$ for better fairness [2], [25], [26], [93].

It can be shown that MRC-SIC is asymptotically capacity approaching in a MIMO system with $K \to \infty$ [24].

We categorize MRC-SIC as NOMA since it requires MUD (in the form of SIC) and ZF as OMA since it requires SUD only. MRC is on the borderline: it employs SUD without spatial orthogonality. MRC works fine only when interference is negligible.

D. NOMA AND ITS LIMITATION IN SISO

The following discussion provides useful insights into NOMA and OMA. Let us consider a special case of MRC-SIC in SISO, for which every $\boldsymbol{h}_k$ reduces to a scalar h_k so (11c) becomes

$$R_{\text{sum}} = \log_2\left(1 + \frac{p_1|h_1|^2}{p_2|h_2|^2 + N_0}\right) + \log_2\left(1 + \frac{p_2|h_2|^2}{N_0}\right). \tag{12a}$$

This is a special case of the following SISO SIC scheme with K users:

$$R_{\text{sum}} = \sum_{k=1}^{K}\log_2\left(1 + \frac{p_k|h_k|^2}{\sum_{k'=k+1}^{K} p_{k'}|h_{k'}|^2 + N_0}\right). \tag{12b}$$

Power control (PC) can be applied over p_k. The achievable rate by this PC-SIC scheme coincides with the capacity of a K-user SISO multiple-access channel (MAC) given below [2], [94]

$$R_{\text{sum}} = \log_2\left(1 + \frac{1}{N_0}\sum_{k=1}^{K} p_k|h_k|^2\right). \tag{12c}$$

As comparison, we consider OMA under resource allocation. We adopt TDMA with flexible length (TDMA-FL), in which time slot lengths for different users are optimized to maximize sum-rate [2]. We first consider equal energy control (EEC), in which different users have different power levels but their total energy per frame is the same. (This is possible since their transmission durations can be different.) It is shown in [2, pp. 232–234] that TDMA-FL is capacity achieving under EEC. This is illustrated in Fig. 2. Clearly, NOMA and OMA have the same performance in this case. OMA is actually a simpler option since it involves SUD only.

The situation is slightly different if the total energy per frame per user can also be freely optimized under the sum-energy constraint (SEC). For example, consider maximizing sum-log-rate $\log(r_1) + \log(r_2)$ under the proportional fairness criterion [2]. Fig. 3 illustrates the related numerical

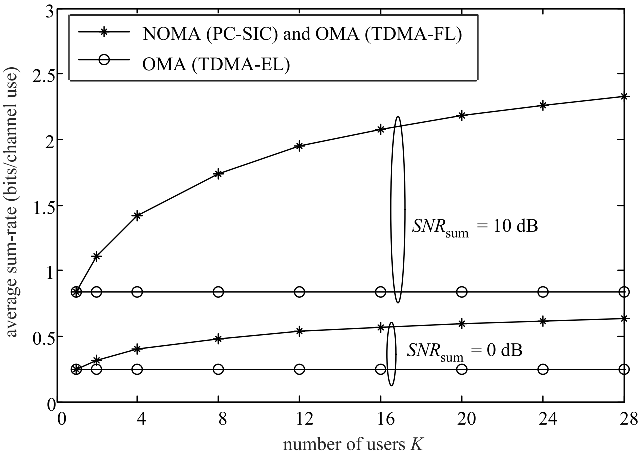


FIGURE 2. **Near-far gain by NOMA and OMA in SISO MAC under EEC. NOMA is based on SIC and OMA on TDMA-FL. TDMA with equal length (TDMA-EL) is used as reference. $N_{BS} = 1$ and $N_{MT} = 1$. Other system settings are the same as those in Fig 1.**

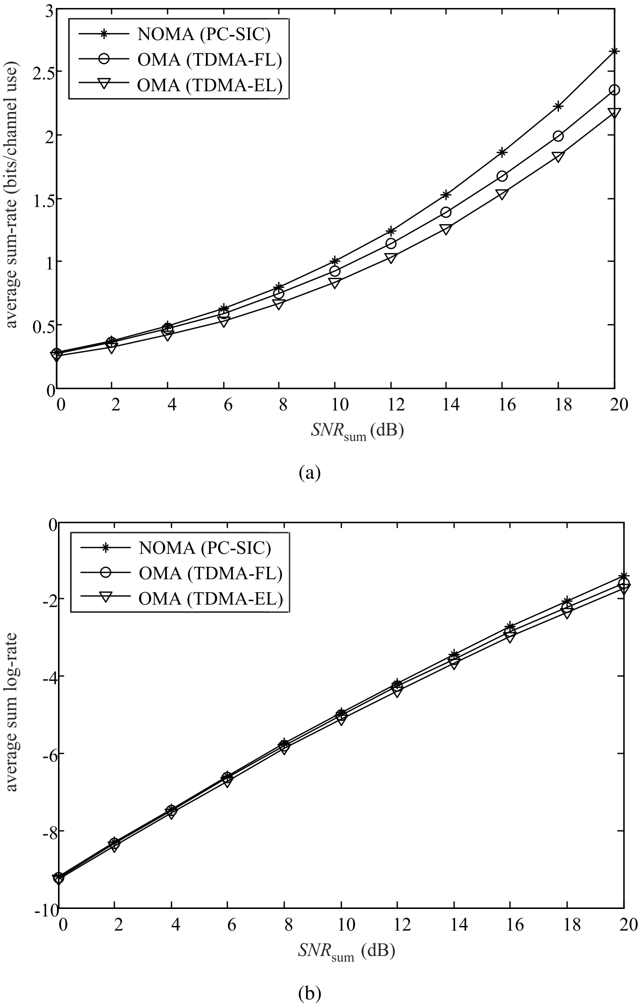
results for $K = 2$ in the typical SNR range.[2] The sum-rate curves in Fig. 3(a) show that PC-SIC is only slightly better than TDMA (about 8% over TDMA-FL and 20% over TDMA-EL at $SNR_{\text{sum}} = 10$ dB). The sum-log-rate curves in Fig. 3(b) indicate that all the schemes have almost the same fairness.

The multi-user gain in Figs. 2 and 3 solely comes from near-far diversity, since there is no angle diversity in SISO. Later in Section III-G, we will see the implications of the above observations in MIMO.

E. COMPARISONS OF ZF, MRC AND MRC-SIC

Fig. 4 compares different approaches under perfect CSIR with the same channel parameters as those in Fig. 1. For ZF under SPC, resource allocation is applied through user selection [90] and water-filling [2]. For MRC-SIC with SPC, to reduce complexity, we simply adopt the same PC levels as those obtained from capacity analysis. Decoding starts from the user with the highest channel gain. Unequal rate allocation [2] is assumed in all the schemes compared in Fig. 4.

Following the duality principles [2], [5], [79], Fig. 4 is applicable to both up- and down-links. Transmitter ZF and maximum ratio transmission (MRT) are, respectively, the down-link duals of receiver ZF and MRC [2]. The down-link dual of the up-link PC and MRC-SIC involves highly complicated dirty paper coding [95]. An exception is SISO, in which simple PC-SIC is capacity achieving for both links (using opposite decoding orders [2, p. 241]).



(a)

(b)

FIGURE 3. **Achievable sum-rate under SEC. (a) Average sum rate and (b) average sum log-rate for NOMA and OMA. System parameters are the same as those in Fig. 2.**

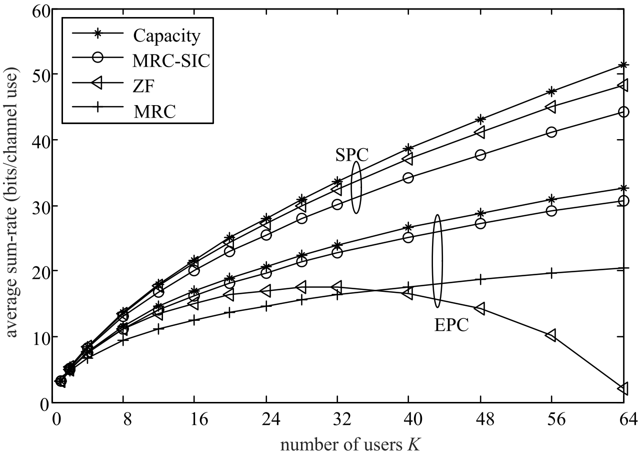


FIGURE 4. **Achievable sum-rate of different approaches under perfect CSIR. System parameters are the same as those in Fig. 1. $SNR_{\text{sum}} = 5$ dB.**

We are not aware of a good optimization method for MRC under SPC. Therefore there is no related result in Fig. 4.

[2] This range is used for the following reason. In a multi-cell system, we should consider interference from other cells. Define the SINR as

$$SINR_{\text{sum}} = P_{\text{sum}}/(N_0 + I_{\text{cross-cell}})$$

where $I_{\text{cross-cell}}$ is the total cross-cell interference from all neighbor cells. Based on the discussion in Section II-D, we assume that other-cell interference can be approximated as additive noise and then SNR_{sum} in a single-cell environment has the equivalent effect as $SINR_{\text{sum}}$ in a multi-cell one. At the end of Section V, we will show that 0-10 dB is a common range for $SINR_{\text{sum}}$ in a multi-cell system, which translates to 0-10 dB for SNR_{sum} in a single-cell system.

We make the following observations from Fig. 4.

- MRC works fine only when K is very small.
- When K is small, ZF can perform close to capacity. When K is large, however, ZF works well under SPC, but not under EPC. The problem is caused by noise amplification discussed below (8).
- MRC-SIC performs well for all K under both EPC and SPC.
- Angle gain can be seen from the EPC curves, which is around ten folds for MRC-SIC. Resource allocation offers further near-far gain. The latter can be achieved by both OMA via ZF and NOMA via MRC-SIC.

Thus, under perfect CSI, both OMA (e.g., ZF) and NOMA (e.g., MRC-SIC) can offer good multi-user gain. Using duality, we expect that ZF works well in the down-link. MRT is also an option when K is small.

We expect that MRC-SIC can potentially offer good fairness with proper rate and power optimization. We are still working on this problem.

F. NOMA VIA ZF-SIC

There are more sophisticated options. In the down-link ZF-SIC schemes in [28] and [30], ZF is used to create orthogonal subspaces. The users in the same sub-space form a group. SIC is applied to the users in each group. This ZF-SIC scheme improves fairness [28], [30] but suffers from some obstacles:

- Orthogonal grouping is difficult in practice. To see this, let Pr(g) be the probability of the event that the size of a group is g. It can be shown that $\Pr(g) \to 0$ quickly when g increases. The probability of $g \geq 2$ is usually very small, which implies that the benefit of SIC is small statistically.
- With CSI error, cross group interference may cause severe problem during SIC.
- The cost of SIC in the down-link can be a concern.

G. OMA VIA ZF-TDMA AND TDMA-ZF

Now suppose that orthogonal grouping has been done via ZF as above. We actually have a much simpler OMA alternative by applying TDMA to the users within each group. We call this approach ZF-TDMA. Similar to the discussions above, we have $\Pr(g \geq 2) \approx \Pr(g = 2)$. From Fig. 3, we expect that ZF-TDMA (with FL) underperforms ZF-SIC only slightly at $g = 2$.[3] The difference is insignificant compared with the overall multi-user gain in the order of tens of times, as seen in Fig. 4.

ZF-TDMA still involves orthogonal grouping. We can avoid the problem by swapping the order of ZF and TDMA as follows.

- Each time frame is divided into non-overlapping slots. Each time slot is assigned with several users.
- ZF is applied to the users in each slot.

We call the above TDMA-ZF. It can be categorized as OMA and requires only SUD. Judging from Fig. 3, we expect that flexible time allocation can offer good multi-user gain as well as improved fairness in TDMA-ZF.

H. IMPACT OF UNCERTAIN CSIR

From Fig. 4, channel capacity can be reasonably approached by ZF or MRC-SIC under perfect CSIR. The situation can be very different with CSIR error. Since the related capacity analysis is difficult, we will discuss the issue using numerical results below.

Denote by $\tilde{\boldsymbol{H}} = \{\tilde{H}_{n,k}\}$ and $\Delta\boldsymbol{H} = \{\Delta H_{n,k}\}$ two $N_{\text{BS}} \times K$ matrices of i.i.d. complex Gaussian entries with zero mean and unit variance. We adopt a simplified model to characterize CSI error [96], [97],

$$\boldsymbol{H} = \sqrt{\varepsilon}\tilde{\boldsymbol{H}} + \sqrt{1-\varepsilon}\Delta\boldsymbol{H}, \tag{13}$$

where $\sqrt{\varepsilon}\tilde{\boldsymbol{H}}$ and $\sqrt{1-\varepsilon}\Delta\boldsymbol{H}$ respectively represent the known and unknown parts of $\boldsymbol{H}$, and ε $(0 \leq \varepsilon \leq 1)$ is a confidence factor: $\varepsilon = 0$ for no CSI and $\varepsilon = 1$ for perfect CSI. The mean square error (MSE) of CSI is given by

$$\text{MSE} \equiv \text{E}\left[\left|H_{n,k} - \sqrt{\varepsilon}\tilde{H}_{n,k}\right|^2\right] = 1 - \varepsilon. \tag{14}$$

This relationship will be used in Section V-C.

Substituting (13) into (2a), we have

$$\boldsymbol{y}(j) = \sqrt{\varepsilon}\tilde{\boldsymbol{H}}\boldsymbol{x}(j) + \tilde{\boldsymbol{\eta}}(j), \tag{15a}$$

where

$$\tilde{\boldsymbol{\eta}}(j) = \sqrt{1-\varepsilon}\Delta\boldsymbol{H}\boldsymbol{x}(j) + \boldsymbol{\eta}(j) \tag{15b}$$

is an equivalent noise vector. Note the similarity between (2a) and (15). We can carry out simulation by treating $\sqrt{\varepsilon}\tilde{\boldsymbol{H}}$ and $\tilde{\boldsymbol{\eta}}(j)$ as if they are, respectively, the true channel matrix and additive noise. However, $\tilde{\boldsymbol{\eta}}(j)$ in (15b) is actually dependent on $\boldsymbol{x}(j)$. We observed considerable performance deterioration due to such dependency.

The impact of CSI error is demonstrated in Fig. 5 for different ε values. A rate-1/3 turbo code with two convolutional component codes of generator matrix $(1, 13/15)_8$ is used for each user. Detailed receiver principles related to Fig. 5 will be explained in Section IV. We can see the loss due to CSIR error for all the schemes. The loss becomes more noticeable when K increases.

MRC outperforms ZF in Fig. 5 when $\varepsilon = 1$ (i.e., perfect CSI) which is different from the results in Fig. 4. This is because, for convenience of simulation, equal rate allocation is used in Fig. 5. Related discussions can be found in [98]. The problem can be improved by unequal rate allocation, which is assumed in Fig. 4.

[3] Since the users in each group have the same angle, the operation within each group is equivalent to SISO SIC in (12). We can thus use Figs. 2 and 3 for performance assessment. This is only an approximate treatment since the distributions of channel gains are different for SISO and MIMO.

Also note that the difference among the curves in Fig. 3 increases with SNR_{sum}. In a cellular system, SNR_{sum} is limited by cross-cell interference. (See footnote 2.) For sum-rate maximization, each group has very small probability to be allocated with a very large sum SNR. Therefore we can focus on the range of SNR_{sum} in Fig. 3.

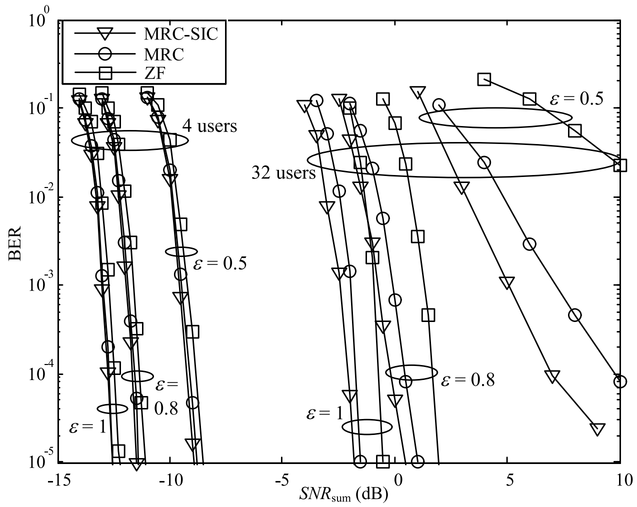


FIGURE 5. The impact of CSIR error. $N_{\text{BS}} = 64$ and $N_{\text{MT}} = 1$. Rayleigh fading. No slow fading. Rate-1/3 turbo coding. The length of information bits of each user $J_{\text{info}} = 1200$. A codeword is transmitted over 10 resource blocks. Each resource block contains 180 symbols experiencing the same fading conditions. $\varepsilon = 1$, 0.8 and 0.5 respectively for each scheme. Iterations process until convergence for the turbo decoder of each user.

We can also see from Fig. 5 that ZF is most sensitive to CSI error. This is expected, since CSI error destroys the interference free assumption in ZF. It appears that MRC-SIC is the more robust one against CSI error in Fig. 5.

Using duality principles, we can also show that CSI error has similar impact on ZF and MRT in the down-link.

I. SUMMARY: NOMA VS OMA

We now summarize Section III. Comparing Figs. 1, 3, 4 and 5, we can make the following observations:

- Multi-user gain from increasing K is potentially very large.
- With accurate CSI, a major part of multi-user gain can be achieved by OMA (i.e., ZF or TDMA-ZF) with proper resource allocation. The remaining gain by introducing NOMA is only incremental.
- Without accurate CSI, all existing methods perform poorly at large K.

Based on duality, the above arguments apply to both up- and down-links.

Assume TDD. CSIR acquired from the up-link can serve as CSIT for the down-link. Then reliable up-link channel estimation holds the key to massive MIMO systems under TDD, for both NOMA and OMA. In what follows, we will show that NOMA has an edge over OMA on this issue.

IV. IDMA AND ITERATIVE MUD

Fig. 4 shows that MRC-SIC can offer excellent performance at relatively low cost in the up-link. Ideal capacity-achieving coding and decoding are assumed there. A practical code incurs extra loss in each SIC step. Such loss accumulates during the SIC process, which can result in serious overall loss. Iterative detection outlined below can compensate for such loss. Iterative processing is also the core in the data aided channel estimation technique discussed in the next section. IDMA facilitates these iterative techniques.

A. SPARSE GRAPHIC MODEL FOR IDMA

IDMA was originally proposed through a sparse graph model [65]. Fig. 6(a) shows a SISO example of a 3-user IDMA system, in which a circle represents a variable and a square marked with "+" represents an addition. Interleaving is represented by random edge connections. When the frame length J increases, the graph becomes more sparse, which facilitates iterative decoding [99], [100]. More details are explained below.

B. IDMA TRANSMITTER

For simplicity, we first consider SISO systems. Let $\boldsymbol{c}_k = \{c_k(j)\}$ be a length-J codeword generated by users k. Assume that $c_k(j)$ is in the binary phase shift keying (BPSK) format: $c_k(j) \in \{-1, +1\}$. A transmitted symbol with BPSK signaling from user k at time j is given by [44]

$$x_k(j) = \sqrt{p_k} c_k(j'), \tag{16a}$$

where $\sqrt{p_k}$ is a power control factor and j' is determined by a user-specific interleaver $\pi_k(\cdot)$. Alternatively, a transmitted symbol with quadrature phase shift keying (QPSK) signaling from user k at time j is given by

$$\text{Re}[x_k(j)] = \sqrt{p_k/2} c_k(j'), \tag{16b}$$
$$\text{Im}[x_k(j)] = \sqrt{p_k/2} c_k(j''), \tag{16c}$$

where j' and j'' are determined by interleaving. The received symbol $y(j)$ at time j is given by

$$y(j) = \sum_{k=1}^{K} h_k x_k(j) + \eta(j), \tag{16d}$$

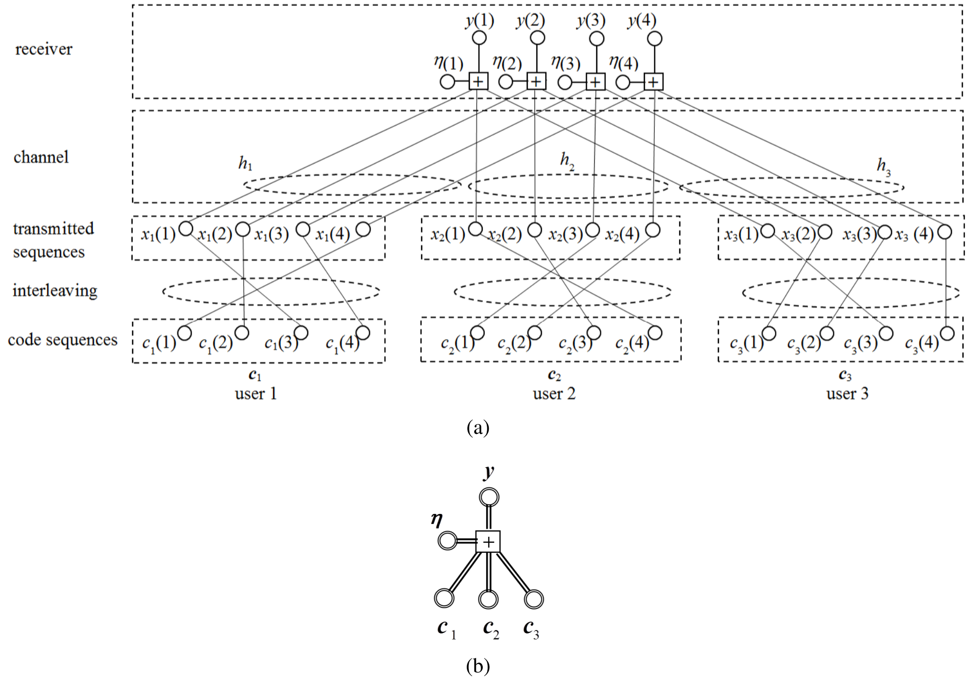
where h_k is the channel coefficient for user k and $\eta(j)$ an AWGN sample.

Fig. 6(b) is a protograph representation [101] of Fig. 6(a). Here each double circle represents a vector and each double line represents a vector connection. Random interleaving is implicitly assumed for each double edge. Denote by $\boldsymbol{y} = [y(1), y(2), \ldots, y(J)]$. Note the difference between $\boldsymbol{y}$ and $\boldsymbol{y}(j)$ in (1). The former is a temporal sequence received on one antenna and the latter a spatial sequence received on N_{BS} antennas at time j. Similarly, denote by $\boldsymbol{c}$ and $\boldsymbol{\eta}$ the coded and noise sequences over time, respectively.

Fig. 6 can be modified to represent the MIMO system in (1) by replacing $y(j)$ and $\eta(j)$ with their vector forms $\boldsymbol{y}(j)$ and $\boldsymbol{\eta}(j)$ for signals over multiple antennas.

C. ITERATIVE DETECTION

An IDMA receiver consists of two local processors, namely elementary signal estimator (ESE) and decoder (DEC) [44]



(a)

(b)

FIGURE 6. (a) Factor graph illustration of IDMA ($J = 4$ and $K = 3$). (b) Protograph representation of (a).

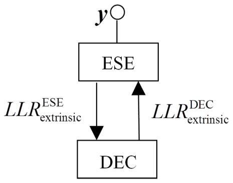

y
ESE
DEC
$LLR^{\text{ESE}}_{\text{extrinsic}}$
$LLR^{\text{DEC}}_{\text{extrinsic}}$

FIGURE 7. Illustration of iterative detector.

as shown in Fig. 7. Both ESE and DEC evaluate the extrinsic log-likelihood ratios (LLRs) below [44]

$$LLR_{\text{extrinsic}}\left(c_k(j)\right) = \log \frac{\Pr\left(c_k(j) = +1\right)}{\Pr\left(c_k(j) = -1\right)} - LLR_{\text{a priori}}\left(c_k(j)\right),\ \forall k, j. \quad (17)$$

The following two steps are iterated in the receiver.

- ESE: Evaluate (17) using (16d) only (i.e., ignoring the coding constraint). The results are fed to DEC as its *a priori* LLRs.
- DEC: Evaluate (17) using a bank of K local decoders, each for a user. The results are fed to ESE as its *a priori* LLRs.

DEC is a standard device. Maximum likelihood (ML) estimation [102] is the optimal realization for ESE. ML involves all 2^K bit combinations from K users, with complexity increasing exponentially with K. For other modulations, complexity can be even higher. For example, the complexity is $O(4^K)$ for QPSK.

Gaussian approximation (GA) is a low cost alternative. The principle of GA for SISO can be found in [44]. GA for MIMO is explained below.

D. ITERATIVE MRC (I-MRC)

We start from the MRC estimator in (9) that is repeated below:

$$\hat{x}_k(j) = \lambda_k x_k(j) + \xi_k(j), \quad (18a)$$

$$\xi_k(j) \equiv \sum_{k'=1,k'\neq k}^{K} \left(\boldsymbol{h}_k^H \boldsymbol{h}_{k'}\right) x_{k'}(j) + \boldsymbol{h}_k^H \boldsymbol{\eta}(j). \quad (18b)$$

Iterative MRC (I-MRC) with GA works as follows. We consider QPSK modulation. Using the extrinsic information from DEC, we compute mean and variance for $x_k(j)$ [44]:

$$\mu_k^x(j) = \mathrm{E}(x_k(j)), \quad (19a)$$

$$v_k^x = \mathrm{Var}(\mathrm{Re}(x_k(j))) = \mathrm{Var}(\mathrm{Im}(x_k(j))). \quad (19b)$$

Here we assume that the variances are the same for all j and also for both real and imaginary parts [103]. This is only an approximation to reduce complexity. It has been widely used in turbo decoding and turbo equalization [38], [44], [103]–[106]. In practice, we can take average if the variances are actually different. Using (18) and (19), we compute

$$\mathrm{E}\left(\xi_k(j)\right) = \boldsymbol{h}_k^H \sum_{k'=1,k'\neq k}^{K} \boldsymbol{h}_{k'} \mu_{k'}^x(j), \quad (20a)$$

$$\mathrm{Var}\left(\mathrm{Re}\left(\xi_k(j)\right)\right) = \sum_{k'=1,k'\neq k}^{K} \mathrm{Var}\left(\mathrm{Re}\left(\boldsymbol{h}_k^H \boldsymbol{h}_{k'} x_{k'}(j)\right)\right) + \|\boldsymbol{h}_k\|^2 N_0/2. \quad (20b)$$

Let $\mathrm{Re}[x_k(j)] = \sqrt{p_k/2}c_k(j')$ for a certain j' (See (16)). With GA, we approximate $\xi_k(j)$ in (18a) by a Gaussian random variable, so that (17) can be evaluated as [44]

$$LLR_{\mathrm{extrinsic}}\left(c_k(j')\right) = \frac{2\lambda_k\sqrt{p_k/2}}{\mathrm{Var}\left(\mathrm{Re}\left(\xi_k(j)\right)\right)}\mathrm{Re}\left(\hat{x}_k \text{ - } \mathrm{E}\left(\xi_k(j)\right)\right). \tag{21}$$

In (21), we effectively treat (18a) as a single user model so the detection complexity is negligible. The main complexity of I-MRC is the updating operations in (20). To reduce complexity, we rewrite (20a) in a sum-and-subtract form as

$$\mathrm{E}\left(\xi_k(j)\right) = \boldsymbol{h}_k^H\left(\sum_{k'=1}^{K}\boldsymbol{h}_{k'}\mu_{k'}^x(j) - \boldsymbol{h}_k\mu_k^x(j)\right). \tag{22a}$$

Note that $\left\|\boldsymbol{h}_k^H\boldsymbol{h}_{k'}\right\|^2 \rightarrow |H_k^{\mathrm{slow}}|^2|H_{k'}^{\mathrm{slow}}|^2N_{\mathrm{BS}}, k' \neq k$ when N_{BS} is large. We can rewrite (20b) as

$$\begin{aligned}&\mathrm{Var}\left(\mathrm{Re}\left(\xi_k(j)\right)\right)\\&\quad\approx |H_k^{\mathrm{slow}}|^2N_{\mathrm{BS}}\left(\sum_{k'=1}^{K}|H_{k'}^{\mathrm{slow}}|^2v_{k'}^x - |H_k^{\mathrm{slow}}|^2v_k^x\right)\\&\qquad+|H_k^{\mathrm{slow}}|^2N_{\mathrm{BS}}N_0/2.\end{aligned} \tag{22b}$$

The per user complexity in (22) does not grow with K since the summations in (22) can be shared by K users. Thus the overall complexity of I-MRC is significantly lower than that of ML. The latter grows exponentially with K.

E. RELATED SCHEMES

The following are some schemes related to IDMA.

1) POWER CONTROL

Similar to SIC in (11), we can optimize p_k in (16). The users with higher arrival powers will converge first during iterative detection, which reduces interference to other users. Detailed discussions can be found in [66] and [67].

2) INTERLEAVER DESIGN

IDMA with random interleaving usually works well. Carefully designed interleaving may also help [107]–[110]. Fig. 6 shows an example. We apply an extra rate-1/2 repetition coding on top of the IDMA scheme in Fig. 6(b) with 3 users. The results are interleaved, partitioned and transmitted in two time slots, as shown in Fig. 8(a). Fig. 8(b) shows a slightly different variation of 6 users and 4 time slots, in which each user transmits in two selected time slots. The latter scheme is named as LDS-CDMA in [109] and [110]. Although its name contains CDMA, LDS-CDMA actually does not employ user-specific spreading sequences as in classic DS-CDMA. Instead, it relies on user specific interleaving for this purpose, similar to IDMA. Interleaving leads to sparsity in both IDMA and LDS-CDMA, which facilitates iterative detection.

The detection complexity in the schemes in Fig. 8 is determined by K_{MUD}, where K_{MUD} is the number of users involved in each slot. Note that K_{MUD} is different from K. For example, $K_{\mathrm{MUD}} = 3$ in both Figs. 8(a) and 8(b) so the complexity is $O(4^3)$ with ML and QPSK.

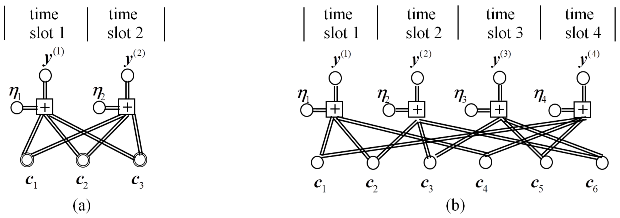


FIGURE 8. **Protographs for (a) IDMA with $K = 3$ and (b) LDS-CDMA with $K = 6$.**

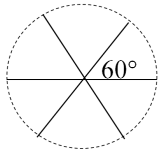


FIGURE 9. **Illustrations of phase control for $K = 3$.**

3) PHASE CONTROL

We can control the phases of the arrival signals [111], [112] to reduce interference. For example, when $K = 3$, the phases of three users can be controlled as illustrated in Fig. 9. This method has the following limitations.

- In phase control, we need to compensate for the phase shift of the channel. Recall from (1) that the received signal for user k is $\boldsymbol{h}_kx_k(j)$. For SISO, h_k is a scalar and the phase of $x_k(j)$ can be adjusted for this purpose. In MIMO, it is usually not possible to compensate for the phases of all the elements (corresponding to different antennas) in $\boldsymbol{h}_k$. This means that we cannot realize the required phase distribution in Fig. 9 on all the antennas.
- Phase control requires accurate CSIT, which incurs extra cost on the feedback link. Such cost is higher than that for power control, since typically phase changes faster than power.

4) SIGNAL SHAPING AND LABELING

Fig. 10(a) shows a standard BPSK constellation with two different labelings B_1 and B_2. Fig. 10(b) shows two constellations resulted from superposition coded modulation (SCM) [106], [113]. M_1 is the superposition of two B_1's. M_2 is the superposition of B_1 and B_2.

Let signals $+1$ and -1 in Fig. 10(a) have probability 0.5 each. Then signal 0 in Fig. 10(b) has probability of 0.5 and signals 2 and -2 have probability 0.25 each. Such unequal transmission probabilities make Fig. 10(b) more Gaussian like. Such a signal shaping technique is studied in [106] and [113]–[116] . The optimality of SCM labeling is proved in [106].

For the SCMA scheme developed in [112] and [116], one replicas of each c_k uses M_1 and the other uses M_2.

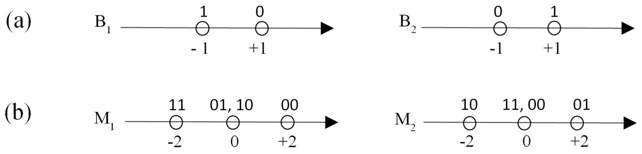


FIGURE 10. **(a) BPSK constellation with two different labelings B_1 and B_2. (b) SCM constellation obtained after superimposing two BPSK signals in (a). For M_1, both BPSK signals are with labelling B_1. For M_2, one is with B_1 and the other with B_2.**

This scheme is effective when it is used together with phase control.

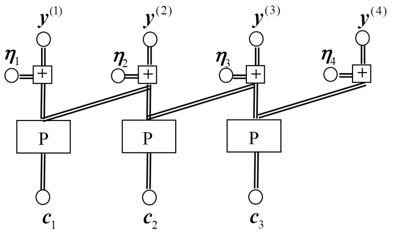


FIGURE 11. **Protograph for an SC-IDMA system with $K = 3$ and coupling width $W = 2$. A block marked "P" partitions the transmitted sequence from a user into two segments that are then transmitted in two consecutive time slots. Random user-specific interleaving is applied before partitioning. No compression is used.**

5) SPATIAL-COUPLING

Fig. 11 shows a protograph of spatially coupled IDMA (SC-IDMA) that is related to the schemes discussed in [117]–[124]. In Fig. 11, a block marked "P" partitions each $\boldsymbol{c}_k$ into two segments that are transmitted in two adjacent time slots. Each received signal $\boldsymbol{y}^{(k)}$ in Fig. 11 is the superposition of the signals from two users except at the two terminations. In Fig. 11, K_{MUD} is at most 2 (for $\boldsymbol{y}^{(2)}$ and $\boldsymbol{y}^{(3)}$) and so the complexity is $O(4^2)$ with QPSK and ML, which is much lower than the schemes in Fig. 8. Compression has been discussed for spatially coupled schemes in [123] and [124] for universal coding. No compression is used in the simulation results for SC-IDMA in this paper. Empirically, we observed that SC-IDMA performance can be improved by properly increasing the powers of the two users at the two ends (i.e., users 1 and 3 in Fig. 11). Such unequal power control enhances the termination effect as discussed in [122]. The simulation results below are based on power allocation ratios 1:0.75:1.

F. NUMERICAL COMPARISONS

The turbo codes used below all consist of two component convolutional codes with generator matrix $(1, 13/15)_8$ (same as that in Section III-H). The basic rate is 1/3. Puncturing is used to obtain rate $= 1/2$. Each iteration involves one ESE operation and one decoding operation per component code. The input information of each decoder is the combination of the feedback from ESE and the output of the appositive component code in the last iteration. There is no internal iteration in the turbo code of each user. The length of information bits of each user is denoted as J_{info}.

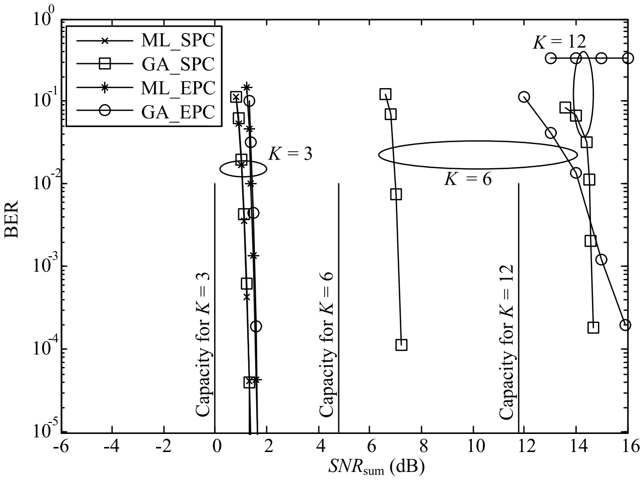


FIGURE 12. **IDMA systems with EPC and SPC over AWGN channels. ML detection for $K = 3$ and GA detection for $K = 3, 6, 12$. Rate-1/3 turbo coding followed by rate-1/2 repetition coding is used for each user. $J_{\text{info}} = 4800$. QPSK modulation. Sum-rate = 1, 2, and 4 for $K = 3, 6$ and 12, respectively.**

We first consider SISO systems. Fig. 12 compares EPC and SPC for IDMA with QPSK signaling. The advantage of SPC is marginal for $K = 3$ but becomes significant for $K = 6$ and 12. In particular, for $K = 12$, EPC performs very poorly but SPC still works well. The performances of ML and GA are almost the same at $K = 3$, even though GA has much lower complexity. The complexity of $O(4^K)$ for ML becomes unbearably high at $K = 6$ and 12. We therefore can only effort to provide the simulation results for ML at $K = 3$.

The power factors in Fig. 12 are as follows: {1, 1.23, 1.54} for $K = 3$; {1, 1, 1, 1, 1.73, 2.24} for $K = 6$; {1, 1, 1, 1, 1.73, 2.24, 2.91, 3.79, 4.92, 6.40, 8.32, 10.82} for $K = 12$. The detailed power optimization algorithm can be found in [66] and [67].

Fig. 13 shows the effect of interleaver design, phase and modulation control and spatial-coupling. All schemes have sum-rate = 3/2 with rate-1/2 turbo coding for each user. The followings are some details.

- IDMA follows Fig. 8(a) with 3 users, rate-1/2 repetition coding, QPSK and power allocation ratios 1:0.7:0.5.
- SC-IDMA follows Fig. 11 with 3 users, QPSK and power allocation ratios 1:0.75:1.
- LDS-CDMA follows Fig. 8(b) with 6 users, rate-1/2 repetition coding, QPSK and equal power allocation.
- SCMA follows Fig. 8(b) with 6 users, rate-1/2 repetition coding, 60° phase control (Fig. 9), SCMA modulation (Fig. 10(b)) and equal power allocation.

Power control may be possible for LDS-CDMA and SCMA, but there is no known method for this purpose. Exhaustive search over 6 users is too costly. We therefore only consider equal power allocation for LDS-CDMA and SCMA.

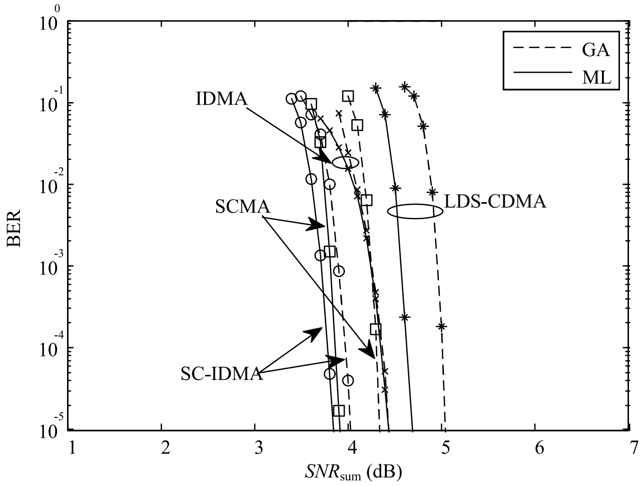


FIGURE 13. Comparison over AWGN channels. For all schemes, $J_{\text{info}} = 4800$, rate-1/2 turbo coding and sum-rate = 3/2. Iterations process until convergence.

We make the following observations from Fig. 13.

- SC-IDMA offers the best performance as well as the lowest decoding complexity of $O(4^2)$ with ML (as $K_{\text{MUD}} = 2$ in Fig. 11).
- Except for SC-IDMA, all other schemes have $K_{\text{MUD}} = 3$ with ML complexity $O(4^3)$.
- The use of GA can greatly reduce complexity. GA incurs certain performance loss. Such loss is marginal in IDMA and SC-IDMA and slightly more noticeable in LDS-CDMA and SCMA.
- Sparsity is not unique for SCMA since all the schemes compared in Fig. 13 rely on sparsity to facilitate iterative detection. The unique features of SCMA are actually the special phase control, signal-shaping and labelling method in Figs. 9 and 10.
- The 3-user settings of IDMA and SC-IDMA are more flexible than the 6-user settings of LDS-CDMA and SCMA.

Based on the above observations, from now on we will only discuss IDMA with GA and power control, which are simple as well as of excellent performance. Our work on spatial-coupling is still preliminary. We will report the related results later.

Next we proceed to massive MIMO systems. Fig. 14 shows the convergence behavior of I-MRC for a 16-user IDMA system with fast fading only. Single-user interference-free performance is included as reference. As sum-rates are different for $K = 1$ and $K = 16$, SNR_{sum} is not a fair criterion for comparison and so E_b/N_0 is used instead in Fig. 14. We can see that, after 30 iterations, the 16-user system performs almost the same as the single-user one. The performance is sufficiently good after 10 iterations.

Fig. 15 illustrates the multi-user gain for $K = 8$ with I-MRC. Multiple signal streams are used for each user for rate adjustment [106], [115]. We consider three different settings:

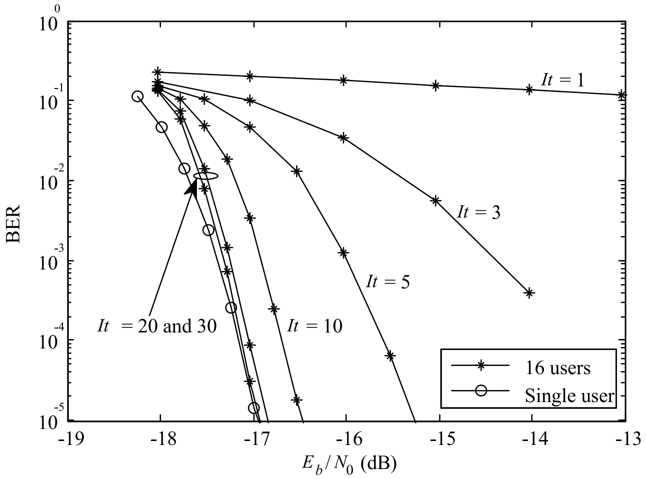


FIGURE 14. I-MRC for $K = 16$. $N_{\text{BS}} = 64$ and $N_{\text{MT}} = 1$. Rayleigh fading. No slow fading. Rate-1/3 turbo coding and $J_{\text{info}} = 1200$. A codeword is transmitted over 10 resource blocks. Each resource block contains 180 symbols experiencing the same fading conditions. Maximum iteration numbers (denoted by *It* in the figure) are 1, 3, 5, 10, 20 and 30, respectively. Different interleavers are applied to the users based on IDMA. Single user interference-free performance is included as reference.

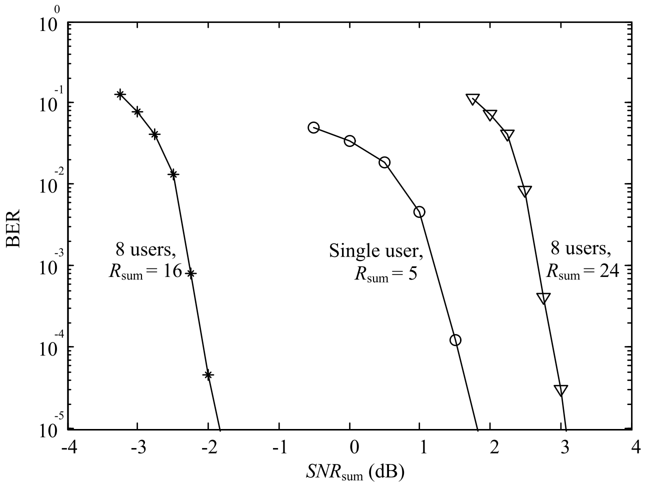


FIGURE 15. Multi-user gain for $K = 8$ with I-MRC. Maximum iteration number is 30. Equal transmit power is assumed for different users. Power control is used for the streams assigned to the same user. Rate-1/2 turbo coding and $J_{\text{info}} = 1200$ for each stream. Other parameters and settings are the same as those in Fig. 14.

- $K = 1$ and $R_{\text{sum}} = 5$ with five signal streams assigned to the sole user,
- $K = 8$ and $R_{\text{sum}} = 16$ with two streams per user, and
- $K = 8$ and $R_{\text{sum}} = 24$ with three streams per user.

For $K = 1$, all the signal streams see the same channel so there is no spatial diversity among them, which results in poor performance. Increasing K from 1 to 8 results in drastically enhanced rate or reduced power or both in Fig. 15. Fig. 15 is a compelling evidence for multi-user gain: allowing more concurrent transmitting users is more efficient than increasing single user rate. The power allocation levels used in Fig. 15 are obtained through heuristic search. We will discuss the detailed search algorithm elsewhere.

Though not shown in the figure, it is observed that interleaver design, phase control, modulation control and spatial-coupling can only offer marginal benefit in massive MIMO. Only power control is effective and it is applied to the system simulated in Fig. 15. For more details on power control algorithms, see [66], [67], [125], [126].

V. ITERATIVE DATA-AIDED CHANNEL ESTIMATION

It was shown in Section III that CSI quality is crucial in massive MIMO systems. We now discuss a data-aided channel estimation (DACE) technique [19], [22], [45], [76]–[78] to improve CSI acquisition in the up-link. DACE can be naturally combined with MRC under the NOMA framework, which provides an efficient solution to massive MIMO.

A. DACE

The correlation among the pilots used by different users can lead to considerable performance degradation. This is referred to as the pilot contamination problem [16]. DACE is analyzed in [22] for this problem. DACE can be used jointly with MRC, which involves the iteration of following two operations:

- using partially decoded data as pilots to refine channel estimation and
- using improved channel estimates to refine data estimation using MRC.

Data sequences are typically much longer than pilots, so correlation is low among them. DACE increases the effective pilot power as well as reduces pilot contamination [22].

The principle of DACE is outlined as follows. We divide a transmitted signal frame of length J into blocks, each of length J', and add K pilot symbols into each block. The total length of each block becomes $J' + K$. Assume that channel coefficients remain unchanged in each block. For each block, the received signals of data symbols at the nth antenna can be rewritten in form of time sequences according to (2) as below.

$$\boldsymbol{y}_n = \sum_{k=1}^{K} H_{n,k}\boldsymbol{x}_k + \boldsymbol{\eta}_n, \tag{23a}$$

where

$$\boldsymbol{y}_n = [y_n(1), \cdots, y_n(j), \cdots, y_n(J')]^T, \tag{23b}$$

$$\boldsymbol{x}_k = [x_k(1), \cdots, x_k(j), \cdots, x_k(J')]^T, \tag{23c}$$

$$\boldsymbol{\eta}_n = [\eta_n(1), \cdots, \eta_n(j), \cdots, \eta_n(J')]^T. \tag{23d}$$

Here $y_n(j)$, $x_k(j)$ and $\eta_n(j)$ are, respectively, entries of $\boldsymbol{y}(j)$, $\boldsymbol{x}(j)$ and $\boldsymbol{\eta}(j)$ in (2). Let $\{\bar{\boldsymbol{x}}_k = \mathrm{E}(\boldsymbol{x}_k), \forall k\}$ be the DEC feedbacks and $\left\{\bar{H}_{n,k} = \mathrm{E}(H_{n,k}), \forall n, k\right\}$ be obtained from the previous iteration that are known to the receiver. For user k, we can thus compute

$$\boldsymbol{z}_{n,k} = \boldsymbol{y}_n - \sum_{k'=1,k'\neq k}^{K} \bar{H}_{n,k'}\bar{\boldsymbol{x}}_{k'}. \tag{24}$$

Substituting (23a) into (24), we have

$$\boldsymbol{z}_{n,k} = \bar{\boldsymbol{x}_k} H_{n,k} + \boldsymbol{\xi}_{n,k}, \tag{25a}$$

where

$$\boldsymbol{\xi}_{n,k} = H_{n,k} \cdot (\boldsymbol{x}_k - \bar{\boldsymbol{x}}_k) + \sum_{k'=1,k'\neq k}^{K} \left(H_{n,k'}\boldsymbol{x}_{k'} - \bar{H}_{n,k'}\bar{\boldsymbol{x}}_{k'}\right) + \boldsymbol{\eta}_n \tag{25b}$$

is an equivalent noise term. We can see that the operation in (24) is to minimize the power of $\boldsymbol{\xi}_{n,k}$ by canceling the interference from other users. Assume that $\left\{\xi_{n,k}(j), \forall j\right\}$ (the entries of $\boldsymbol{\xi}_{n,k}$) are i.i.d. with zero mean and variance $V^{\xi}_{n,k} = \mathrm{E}(|\xi_{n,k}(j)|^2)$. For each block, user k is assigned with a unique length-K pilot sequence $\boldsymbol{p}_k$ that is orthogonal to the pilot sequences of other users in the same cell. Thus users are free from same-cell pilot interference. Let $\boldsymbol{z}^{\text{pilot}}_{n,k}$ be the pilot sequence observation of user k at the nth antenna. The LMMSE estimation of $H_{n,k}$ based on (25) is given by

$$\hat{H}_{n,k} = \frac{\dfrac{\boldsymbol{p}_k^H \boldsymbol{z}^{\text{pilot}}_{n,k}}{N_0} + \dfrac{\bar{\boldsymbol{x}}_k^H \boldsymbol{z}_{n,k}}{V^{\xi}_{n,k}}}{\dfrac{1}{|H^{\text{slow}}_k|^2} + \dfrac{\boldsymbol{p}_k^H \boldsymbol{p}_k}{N_0} + \dfrac{\bar{\boldsymbol{x}}_k^H \bar{\boldsymbol{x}}_k}{V^{\xi}_{n,k}}}. \tag{26}$$

Some details on (26) can be found in the Appendix.

The advantages of DACE are two folds:

(i) With DACE, the estimated data are gradually used to help pilot for channel estimation. Pilot energy can be greatly reduced since only very coarse CSI is required initially.

(ii) DACE is robust against pilot contamination that is caused by the correlation among the pilot sequences used by different users [22]. Without DACE, longer pilot sequences will be required to reduce such correlation. Thus DACE also reduces the time overhead related to pilots.

B. SYMMETRIC AND ASYMMETRIC TRAFFIC FLOWS

In some services, such as speech, each user occupies both-up and down-links with symmetric traffic flows. In this case, under TDD, the CSI estimated at the BS is shared for both links, so is the advantage of DACE.[4]

Many applications have asymmetric traffics in the two links. For example, a user downloading a file may not upload anything at the same time. In this case, uploading-only users

[4] Here are some details. Consider estimating CSI using a pilot together an up-link codeword U. The estimated CSI is used to transmit a down-link codeword D. Assume that each codeword is transmitted over multiple coherent resource blocks with different fading realizations. Also assume that channel estimation and decoding are carried out after collecting all the observations of U. For casuality, U must be transmitted entirely before D in time. This can be ensured by arranging all the blocks of U on different subcarriers at the same time, followed by all the blocks of D.

From Property 2, there is no need for resource allocation over frequency in massive MIMO. (Spatial water-filling is still useful in ZF.) Thus transmitting each codeword with equal power across subcarriers at the same time does not compromise efficiency.

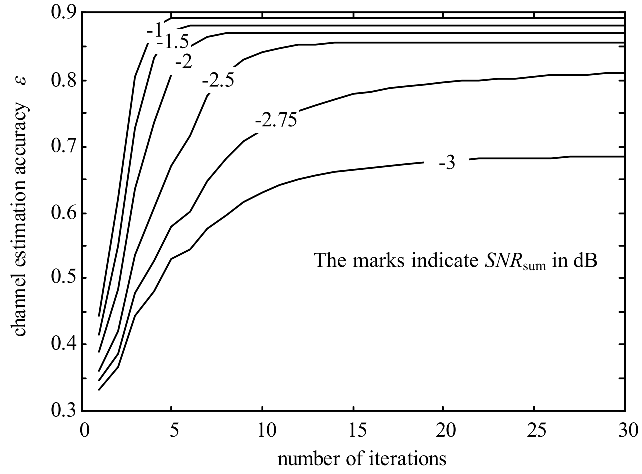


FIGURE 16. Iterative refinement of the channel estimation accuracy without inter-cell interference ($\beta = 0$). Fast fading only. No slow fading. $N_{\text{BS}} = 64$, $K = 16$, $J_{\text{info}} = 1312$. Rate-1/3 turbo coding with QPSK modulation. Codeword length = 1968 QPSK symbols. Each codeword is divided into 12 sections, each with 164 symbols. Each section is transmitted over a coherent resource block of 180 symbols (including 16 pilot symbols). The pilots of the sixteen users form an set of orthogonal bases. The pilot and data symbols have the same average power. All users have the same received power.

can take advantage of DACE but downloading-only users cannot.

Thus downloading-only users have to use dedicated pilots in the up-link for channel estimation. The related overheads will be high since there is no aid from data. Pilot contamination is an open problem in this case [12]. (In a way, DACE for uploading-only users still help, as it leaves more resource to downloading-only users.)

C. NUMERICAL RESULTS

In the following simulations, we focus on the up-link with symmetric traffic.

A turbo code with two convolutional component codes of generator matrix $(1, 13/15)_8$ (same as that in Sections III-H and IV-F) is used for each user below. Coding rate is 1/3. For the proposed iterative MRC and DACE (I-MRC-DACE) scheme, each iteration involves an MRC-DACE operation per user and one decoding operation per component code. The input information of each component decoder is the combination of the feedback from the symbol detector and the output of the appositive component code in the last iteration. There is no internal iteration for the turbo code of each user for DACE. DACE is not performed in MRC, ZF, and I-MRC. For I-MRC, each iteration involves one MRC operation per user and one decoding operation per component code. For MRC and ZF, iterations are performed only in the turbo decoder for each user. All pilot and data symbols have the same average power.

Fig. 16 shows channel estimation quality ε as a function of iteration number via I-MRC-DACE for different SNR_{sum}'s. In simulation, MSE can be measured numerically. Then ε can be calculated based on (14). We can see that iterative processing can offer significant improvement on channel estimation. For example, with $SNR_{\text{sum}} = -1$ dB, ε can be increased from 45% to nearly 90% within only 5 iterations.

We now consider cross-cell interference. Denote by $I_{\text{cross-cell}}$ the total cross-cell interference power from all neighboring cells and by P_{sum} the sum-power per cell (see footnote 2). For simplicity, we assume that $I_{\text{cross-cell}}$ is linearly proportional to P_{sum} via a factor β:

$$I_{\text{cross - cell}} = \beta P_{\text{sum}}. \tag{27}$$

A single cell system has $\beta = 0$. A typical value for a multi-cell system is $\beta = 0.6$ [2]. In practice, the value of β can be affected by many factors. A smaller path loss exponent will lead to a larger β value, which happens, e.g., in the environments with line of sight.

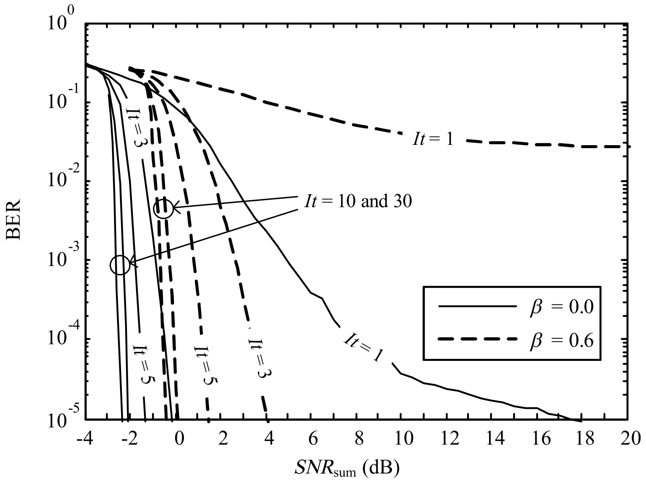


FIGURE 17. Performance of I-MRC-DACE with and without inter-cell interference at different iteration numbers. Iteration number is denoted as *It*. All cells use the same set of pilot sequences. Other parameters and settings are the same as those in Fig. 16.

Fig. 17 shows the impact of β on I-MRC-DACE. We can see that I-MRC-DACE converges quite fast with and without cross-cell interference. Most iterative gain is achieved within about 10 iterations.

Fig. 18 compares the impact of β on different schemes. We can see that I-MRC-DACE noticeably outperforms other options. The difference becomes very significant when β is large (e.g., $\beta \geq 0.6$).

We now examine the effect of pilot contamination. Recall that average interference power is βP_{sum}. The required SNR_{sum} to achieve a certain BER should increase with β. We thus write SNR_{sum} and P_{sum} as functions of β,

$$SNR_{\text{sum}}(\beta) = \frac{P_{\text{sum}}(\beta)}{N_0}. \tag{28}$$

Without pilot contamination, cross-cell interference can be treated as independent noise. In this case, the required SINR for a fixed BER should not change with β. This implies that

$$SINR_{\text{sum}} = \frac{P_{\text{sum}}(\beta)}{\beta P_{\text{sum}}(\beta) + N_0} = \frac{SNR_{\text{sum}}(\beta)}{\beta SNR_{\text{sum}}(\beta) + 1} \tag{29}$$

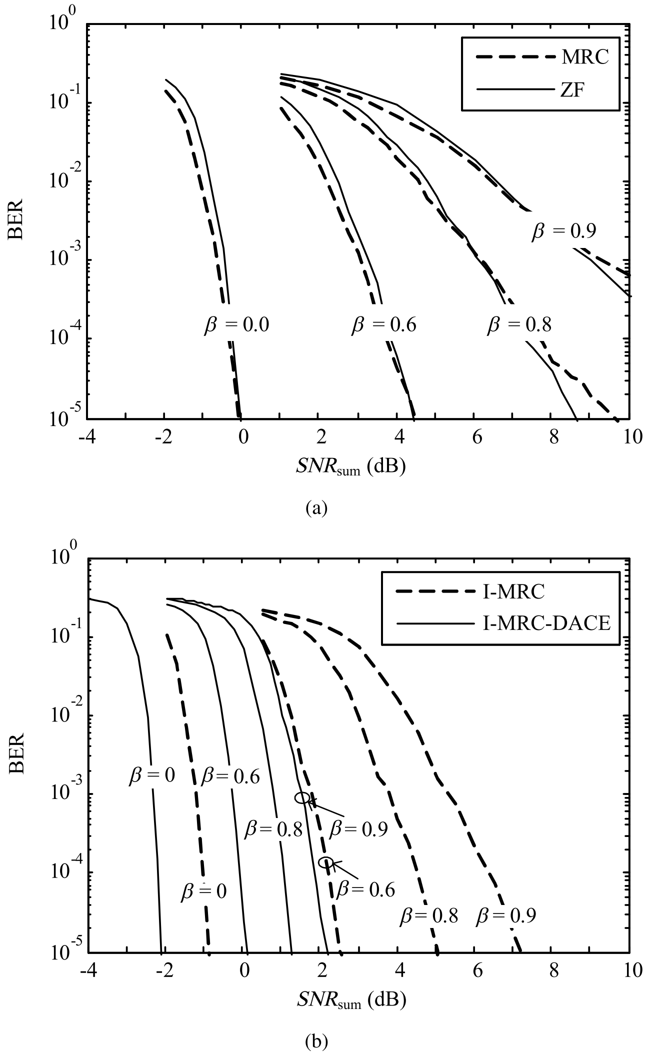


FIGURE 18. Performance comparison of different schemes with channel estimation at different β values: (a) MRC and ZF and (b) I-MRC and I-MRC-DACE. Iteration number = 10. Other parameters and settings are the same as those in Fig. 17.

remains unchanged for all β. In particular, setting $\beta = 0$ in (29), we have $SINR_{\rm sum} = SNR_{\rm sum}(0)$. Substituting this back to (29), we obtain

$$SNR_{\rm sum}(\beta) = \left(\frac{1}{SNR_{\rm sum}(0)} - \beta\right)^{-1}. \tag{30}$$

However, due to pilot contamination, cross-cell interference actually cannot be treated as independent noise and so (30) does not hold. We can examine the discrepancy by comparing two $SNR_{\rm sum}(\beta)$ values: one directly read from Fig. 18 and the other computed using (30). (Recall that the $SNR_{\rm sum}(0)$ is for the interference-free case, so $SNR_{\rm sum}(0)$ in (30) can be read from Fig. 18.) Define

$$\gamma = \frac{SNR_{\rm sum}(\beta) \text{ read from Fig. 18}}{SNR_{\rm sum}(\beta) \text{ computed using (30)}}. \tag{31}$$

A large γ indicates a high power cost incurred by pilot contamination, which is caused by the correlation among the pilots of different users. Such correlation results in wrongly steered interference signals and increases the effective interference power suffered by each user. The γ values for different schemes at BER $= 10^{-5}$ are compared in Fig. 19. We can see that I-MRC-DACE has much smaller γ than other alternatives, indicating its insensitivity to pilot contamination. This is because DACE increases the effective pilot length and alleviates the correlation problem mentioned above. More detailed discussions on this issue can be found in [22].

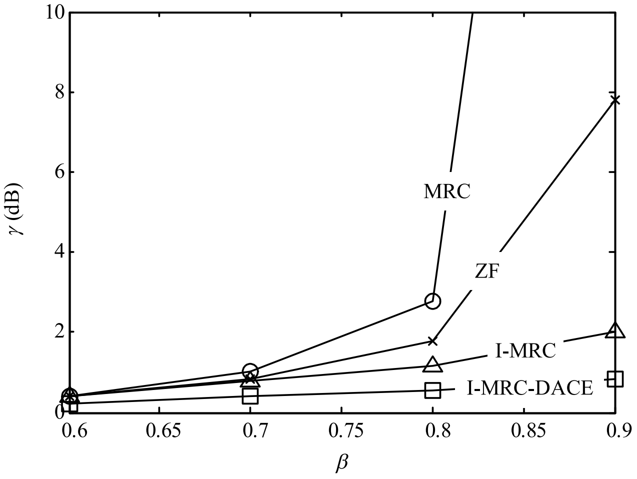


FIGURE 19. Comparison of performance loss due to pilot contamination for different schemes. The γ values are the difference between $SNR_{\rm sum}(\beta)$ estimated using (30) and $SNR_{\rm sum}(\beta)$ reading from Fig. 18 at BER $= 10^{-5}$. Other parameters and settings are the same as those in Fig. 18.

Incidentally, from (29), $SINR_{\rm sum}$ is upper bounded by

$$SINR_{\rm sum} \le \frac{1}{\beta}. \tag{32}$$

For a typical value of $\beta = 0.6$, we have $SINR_{\rm sum} \approx 2.2$ dB. Allowing a certain range for β, we may consider 0–10 dB as a typical range for $SINR_{\rm sum}$.

VI. CONCLUSIONS

A. MAIN FINDINGS

We now summarize the main findings of this paper as follows.

- Multi-user gain is the most attractive benefit from massive MIMO, with potential rate increase in the order of tens or even hundreds (for very large $N_{\rm BS}$) of times.
- Under accurate CSI, multi-user gain can be achieved by both OMA and NOMA. The difference between them is small in most practical situations. OMA can be preferred in this case since it allows low-cost SUD receiver. With resource allocation, OMA (such as ZF) can explore multi-user gain as well as improve fairness.
- Without accurate CSI, neither OMA nor NOMA can perform well.
- NOMA has an advantage over OMA in CSI acquisition. It allows gradual CSI refinement through iterative processing. This greatly reduces the power and time overheads related to pilot and also alleviates the pilot contamination problem.

We resort to NOMA only because we do not have reliable CSI to establish spatial orthogonality initially. A simple solution is IDMA for the up-link that facilitates I-MRC-DACE at the BS. Under TDD, the acquired CSI can be used to support ZF or MRT for the down-link. Overall, we can conclude that CSI acquisition at the BS is the most challenging issue for massive MIMO.

B. FUTURE WORKS

1) COMPREHENSIVE COMPARISON BETWEEN OMA AND NOMA

More studies are still required to carefully weigh the relative advantages of OMA and NOMA. In particular, for fair comparison, optimization over rate, power, frequency, time and space should also be considered for OMA. TDMA-ZF mentioned earlier, for example, can be used for this purpose. It is also necessary to consider multi-cell environments, where SINR is limited by cross-cell interference. For convenience, we used in this paper a single parameter β to characterize a cellular system. This provides useful insights into the problem, but practical situations are much more complicated. More studies are required to assess the impact of non-ideal factors such as CSI uncertainty [127] and cross-cell interference fluctuation.

2) POWER ALLOCATION FOR IDMA

A simple linear programming algorithm is developed in [66], [67], [125], and [126] for power allocation among different users in SISO IDMA systems. The problem is more complicated in MIMO IDMA systems. We used a heuristic search technique in our simulation work. More efficient algorithms are still required to cope with situations in massive MIMO systems with a large number of users.

3) COOPERATIVE AND DISTRIBUTIVE SYSTEMS

NOMA over cooperative SISO cellular systems is recently studied in [128], which can be extended to cooperative and distributed MIMO systems [129]–[133]. Sharing CSI among distributed transmitters is a main difficulty here. Decentralized control without global CSI can be a low cost alternative [132]. The role of CSI in cooperative or distributive systems is, again, an important research topic.

4) MILLIMETER WAVE SYSTEMS

The discussions in this paper are based on the conventional sub-millimeter model. Millimeter wave systems have been widely studied as a special case of massive MIMO. Angle and near-far diversities, as introduced in Section II-B, can also be defined in millimeter systems. Therefore we expect that the findings in this paper can be extended. In particular, I-MRC-DACE can also be used to enhance performance in millimeter wave systems.

5) MULTIPLE MT ANTENNAS

When $N_{\text{MT}} > 1$, CSIT at MTs is required for up-link precoder design. It is shown in [134] that CSIT accuracy at MTs is not crucial in this case if K is much smaller than N_{BS} (though accurate CSIR at the BS is still crucial). A statistical water-filling (SWF) technique is studied in [134] that greatly eases the burden of CSI acquisition at MTs. Significant multiuser gain is still achievable in this way. How to obtain the required statistical information for SWF at low cost is a topic of practical importance.

6) RANDOM ACCESS

Most NOMA schemes require centralized power control to facilitate SIC. Decentralized power control (DPC) is an alternative for random access. This issue was discussed in [135] by combining DPC and SIC (DPC-SIC) in SISO. The new technique can offer throughput improvement over conventional random access techniques such as ALOHA. It provides an attractive option in, e.g., machine-to-machine applications characterized by a large number of sporadic short packets [135]–[139]. The extension of DPC-SIC to MIMO is an interesting issue. Preliminary results can be found in [136].

C. SOFTWARE

We will make some of the software used for the simulation results in this paper available in the following site. http://www.ee.cityu.edu.hk/~liping/Research/Simulationpackage/

APPENDIX

A. DERIVATION OF (26)

Recall from (3) that $H_{n,k} = H_k^{\text{slow}} H_{n,k}^{\text{fast}}$. We will assume that $|H_k^{\text{slow}}|^2$, i.e. the long term average channel gain is known at the receiver, which is reasonable in practice. We will also assume that $H_{n,k}^{\text{fast}}$ is Gaussian distributed with zero mean and unit variance (i.e., Rayleigh fading). Thus $\mathrm{E}(|H_{n,k}|^2) = |H_k^{\text{slow}}|^2$. We have two sources of information to estimate $H_{n,k}$:

- the pilot sequence $\boldsymbol{p}_k$ and observation $\boldsymbol{z}_{n,k}^{\text{pilot}}$, and
- the data observation in (25).

The standard LMMSE estimation of $H_{n,k}$ based on $\boldsymbol{z}_{n,k}^{\text{pilot}}$ is given by [140]

$$\hat{H}_{n,k}^{z} = \mathrm{E}(H_{n,k}|\boldsymbol{z}_{n,k}^{\text{pilot}}) = \frac{1}{\frac{1}{|H_k^{\text{slow}}|^2} + \frac{\boldsymbol{p}_k^H \boldsymbol{p}_k}{N_0}} \cdot \frac{\boldsymbol{p}_k^H}{N_0} \cdot \boldsymbol{z}_{n,k}^{\text{pilot}}, \tag{33a}$$

with

$$MSE_{n,k}^{z} = \mathrm{Var}(H_{n,k}|\boldsymbol{z}_{n,k}^{\text{pilot}}) = \frac{1}{\frac{1}{|H_k^{\text{slow}}|^2} + \frac{\boldsymbol{p}_k^H \boldsymbol{p}_k}{N_0}}. \tag{33b}$$

After updating the mean and variance of $H_{n,k}$ using $\hat{H}_{n,k}^{z} = \mathrm{E}(H_{n,k}|\boldsymbol{z}_{n,k}^{\text{pilot}})$ and $MSE_{n,k}^{z} = \mathrm{Var}(H_{n,k}|\boldsymbol{z}_{n,k}^{\text{pilot}})$, we estimate $H_{n,k}$

again using $\bar{\boldsymbol{x}}_k$. Using [140, eq. (11.33)], we have

$$\hat{H}_{n,k} = \hat{H}^z_{n,k} + \frac{1}{\frac{1}{MSE^z_{n,k}} + \frac{\bar{\boldsymbol{x}}_k^H \bar{\boldsymbol{x}}_k}{V^\xi_{n,k}}} \cdot \frac{\bar{\boldsymbol{x}}_k^H}{V^\xi_{n,k}} \cdot \left(\boldsymbol{z}_{n,k} - \bar{\boldsymbol{x}}_k \hat{H}^z_{n,k}\right). \tag{34}$$

Substituting (33) into (34) gives (26).

B. VARIANCE COMPUTATION

On the right hand side of (26), all the variables are available except $V^\xi_{n,k} = \mathrm{E}(|\xi_{n,k}(j)|^2)$. We now briefly explain how to approximately compute $V^\xi_{n,k}$ in real time. We treat $\boldsymbol{\xi}_{n,k}$ as a zero-mean additive noise vector. From (25b), for each j, we have

$$V^\xi_{n,k} = \mathrm{E}\left(\left|H_{n,k}\right|^2\right) \cdot \mathrm{E}\left(\left|x_k(j) - \bar{x}_k(j)\right|^2\right) + \sum_{k' \neq k}^{K} \mathrm{E}\left(\left|H_{n,k'}x_{k'}(j) - \bar{H}_{n,k'}\bar{x}_{k'}(j)\right|^2\right) + N_0. \tag{35}$$

Here $\bar{x}_{k'}(j)$ is the feedback from the DEC and $\bar{H}_{n,k'}$ is obtained from the previous iteration. As mentioned above, $\mathrm{E}(|H_{n,k}|^2) = |H^{\text{slow}}_k|^2$ is assumed known. $\mathrm{E}(|x_k(j) - \bar{x}_k(j)|^2)$ is the variance of $x_k(j)$ feedback from the DEC and can be computed similarly as (19b) in Subsection IV-D. The term in the summation in (35) can be rewritten as

$$\mathrm{E}\left(\left|H_{n,k'}x_{k'}(j) - \bar{H}_{n,k'}\bar{x}_{k'}(j)\right|^2\right) = \mathrm{E}\left(\left|\Delta H_{n,k'}\bar{x}_{k'}(j) + H_{n,k'}\Delta x_{k'}\right|^2\right), \tag{36}$$

where $\Delta H_{n,k'} = H_{n,k'} - \bar{H}_{n,k'}$ and $\Delta x_{k'} = x_{k'}(j) - \bar{x}_{k'}(j)$.

Consider a standard linear model $y = hx + \eta$. Define MMSE estimation $\bar{x} \equiv \mathrm{E}(x|y)$ and error $\Delta x \equiv x - \bar{x}$. It is known that Δx is uncorrelated with both h and $\bar{x}$ [140]. Base on this, we assume that $\Delta H_{n,k'}$, $\Delta x_{k'}$, $\bar{x}_{k'}(j)$ and $H_{n,k'}$ are approximately uncorrelated with each other. We further assume that they are approximately Gaussian. Then they are independent of each other as well. We then have

$$\mathrm{E}\left(\left|H_{n,k'}x_{k'}(j) - \bar{H}_{n,k'}\bar{x}_{k'}(j)\right|^2\right) = \mathrm{E}\left(\left|\Delta H_{n,k'}\right|^2\right)\mathrm{E}\left(\left|\bar{x}_{k'}(j)\right|^2\right) + \mathrm{E}\left(\left|H_{n,k'}\right|^2\right)\mathrm{E}\left(\left|\Delta x_{k'}\right|^2\right). \tag{37}$$

The terms on the right hand side of (37) can be generated as follows. Assume that $V^\xi_{n,k'}$ is obtained from the previous iteration. Then $\mathrm{E}(|\Delta H_{n,k'}|^2)$ can be obtained as the MSE for estimating $H_{n,k'}$ [140],

$$\mathrm{E}\left(\left|\Delta H_{n,k'}\right|^2\right) = \mathrm{MSE} = \frac{1}{\frac{1}{|H^{\text{slow}}_k|^2} + \frac{\boldsymbol{p}_k^H \boldsymbol{p}_k}{N_0} + \frac{\bar{\boldsymbol{x}}_{k'}^H \bar{\boldsymbol{x}}_{k'}}{V^\xi_{n,k'}}}. \tag{38}$$

For $\mathrm{E}(|\bar{x}_{k'}(j)|^2)$, we use approximation

$$\mathrm{E}\left(\left|\bar{x}_{k'}(j)\right|^2\right) \approx \frac{1}{J}\sum_{j=1}^{J}\left|\bar{x}_{k'}(j)\right|^2. \tag{39}$$

Finally, from (19b), we have $\mathrm{E}\left(|\Delta x_{k'}|^2\right) = \mathrm{Var}(\mathrm{Re}(x_{k'}(j))) + \mathrm{Var}(\mathrm{Im}(x_{k'}(j))) \approx 2v^x_k$ for QPSK modulation. Note that $V^\xi_{n,k}$ serves as a weighting factor for the data observation in (25). From simulation, we observed that DACE is not sensitive to the accuracy of $V^\xi_{n,k}$.

ACKNOWLEDGEMENT

The last author would also like to thank Keying Wu, Jinhong Yuan, Yi Hong, Emmanuel Viterbo, Jiangzhou Wang, Cheng-Xiang Wang, Kit K. Wong, Peter Hoeher, Lin Dai and Moshe Zukerman for enlightening discussions during the writing up of this paper.

REFERENCES

[1] E. Telatar, "Capacity of multi-antenna Gaussian channels," *Eur. Trans. Telecommun.*, vol. 10, no. 6, pp. 585–595, Nov./Dec. 1999.

[2] D. N. C. Tse and P. Viswanath, *Fundamentals Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

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

[4] P. Viswanath, D. N. C. Tse, and V. Anantharam, "Asymptotically optimal water-filling in vector multiple-access channels," *IEEE Trans. Inf. Theory*, vol. 47, no. 1, pp. 241–267, Jan. 2001.

[5] 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, no. 10, pp. 2658–2668, Oct. 2003.

[6] W. Yu, W. Rhee, S. Boyd, and J. M. Cioffi, "Iterative water-filling for Gaussian vector multiple-access channels," *IEEE Trans. Inf. Theory*, vol. 50, no. 1, pp. 145–152, Jan. 2004.

[7] H. Weingarten, Y. Steinberg, and S. Shamai (Shitz), "The capacity region of the Gaussian multiple-input multiple-output broadcast channel," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 3936–3964, Sep. 2006.

[8] M. Kobayashi and G. Caire, "An iterative water-filling algorithm for maximum weighted sum-rate of Gaussian MIMO-BC," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1640–1646, Aug. 2006.

[9] J. Mietzner, R. Schober, L. Lampe, W. H. Gerstacker, and P. A. Hoeher, "Multiple-antenna techniques for wireless communications—A comprehensive literature survey," *IEEE Commun. Surveys Tuts.*, vol. 11, no. 2, pp. 87–105, 2nd Quart., 2009.

[10] G. Caire, N. Jindal, M. Kobayashi, and N. Ravindran, "Multiuser MIMO achievable rates with downlink training and channel state feedback," *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 2845–2866, Jan. 2010.

[11] D. Gesbert, S. Hanly, H. Huang, S. S. Shitz, O. Simeone, and W. Yu, "Multi-cell MIMO cooperative networks: A new look at interference," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 9, pp. 1380–1408, Dec. 2010.

[12] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[13] J. Jose, A. Ashikhmin, T. L. Marzetta, and S. Vishwanath, "Pilot contamination and precoding in multi-cell TDD systems," *IEEE Trans. Wireless Commun.*, vol. 10, no. 8, pp. 2640–2651, Aug. 2011.

[14] F. Rusek *et al.*, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, Jan. 2013.

[15] J. Hoydis, S. ten Brink, and M. Debbah, "Massive MIMO in the UL/DL of cellular networks: How many antennas do we need?" *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 160–171, Feb. 2013.

[16] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.

[17] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive MIMO: Benefits and challenges," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 742–758, Oct. 2014.

[18] P. Hoeher and N. Doose, "A massive MIMO terminal concept based on small-size multi-mode antennas," *Trans. Emerging Telecommun. Technol.*, vol. 28, no. 2, Feb. 2017.

[19] P. Hoeher and F. Tufvesson, "Channel estimation with superimposed pilot sequence," in *Proc. IEEE GLOBECOM*, vol. 4. Dec. 1999, pp. 2162–2166.

[20] J. Vieira, F. Rusek, and F. Tufvesson, "Reciprocity calibration methods for massive MIMO based on antenna coupling," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2014, pp. 3708–3712.

[21] A. Ashikhmin and T. Marzetta, "Pilot contamination precoding in multi-cell large scale antenna systems," in *Proc. IEEE ISIT*, Cambridge, MA, USA, Jul. 2012, pp. 1137–1141.

[22] J. Ma and L. Ping, "Data-aided channel estimation in large antenna systems," *IEEE Trans. Signal Process.*, vol. 62, no. 12, pp. 3111–3124, Jun. 2014.

[23] P. Wang, J. Xiao, and L. Ping, "Comparison of orthogonal and non-orthogonal approaches to future wireless cellular systems," *IEEE Veh. Technol. Mag.*, vol. 1, no. 3, pp. 4–11, Sep. 2006.

[24] P. Wang and L. Ping, "On maximum eigenmode beamforming and multi-user gain," *IEEE Trans. Inf. Theory*, vol. 57, no. 7, pp. 4170–4180, Jul. 2011.

[25] N. Otao, Y. Kishiyama, and K. Higuchi, "Performance of non-orthogonal access with SIC in cellular downlink using proportional fairbased resource allocation," in *Proc. IEEE ISWCS*, Paris, France, Aug. 2012, pp. 476–480.

[26] J. Umehara, Y. Kishiyama, and K. Higuchi, "Enhancing user fairness in non-orthogonal access with successive interference cancellation for cellular downlink," in *Proc. IEEE ICCS*, Munich, Germany, Nov. 2012, pp. 324–328.

[27] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *Proc. IEEE VTC-Spring*, Dresden, Germany, Jun. 2013, pp. 1–5.

[28] B. Kim *et al.*, "Non-orthogonal multiple access in a downlink multiuser beamforming system," in *Proc. IEEE Military Commun. Conf. (MILCOM)*, Nov. 2013, pp. 1278–1283.

[29] K. Higuchi and A. Benjebbour, "Non-orthogonal multiple access (NOMA) with successive interference cancellation for future radio access," *IEICE Trans. Commun.*, vol. E98–B, no. 3, pp. 403–414, Mar. 2015.

[30] Z. Ding, F. Adachi, and H. V. Poor, "The application of MIMO to non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 537–552, Jan. 2016.

[31] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G nonorthogonal multiple access," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6010–6023, Aug. 2016.

[32] Z. Wei, D. W. K. Ng, and J. Yuan. (2016). "Power-efficient resource allocation for MC-NOMA with statistical channel state information." [Online]. Available: https://arxiv.org/abs/1607.01116

[33] S. Verdu, *Multiuser Detection*. Cambridge, U.K.: Cambridge Univ. Press, 1998.

[34] R. Lupas and S. Verdu, "Linear multiuser detectors for synchronous code-division multiple-access channels," *IEEE Trans. Inf. Theory*, vol. 35, no. 1, pp. 123–136, Jan. 1989.

[35] P. Hoeher, "On channel coding and multiuser detection for DS-CDMA," in *Proc. IEEE Int. Conf. Univ. Pers. Commun.*, Ottawa, ON, Canada, Oct. 1993, pp. 641–646.

[36] M. Moher, "An iterative multiuser decoder for near-capacity communications," *IEEE Trans. Commun.*, vol. 46, no. 7, pp. 870–880, Jul. 1998.

[37] M. C. Reed, C. B. Schlegel, P. D. Alexander, and J. A. Asenstorfer, "Iterative multi-user detection for CDMA with FEC: Near-single-user performance," *IEEE Trans. Commun.*, vol. 46, no. 12, pp. 1693–1699, Dec. 1998.

[38] X. Wang and H. V. Poor, "Iterative (turbo) soft interference cancellation and decoding for coded CDMA," *IEEE Trans. Commun.*, vol. 47, no. 7, pp. 1046–1061, Jul. 1999.

[39] F. N. Brannstrom, T. M. Aulin, and L. K. Rasmussen, "Iterative multi-user detection of trellis code multiple access using *a posteriori* probabilities," in *Proc. IEEE ICC*, vol. 1. Jun. 2001, pp. 11–15.

[40] T. Yoo and A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 3, pp. 528–541, Mar. 2006.

[41] K. S. Gilhousen, I. M. Jacobs, R. Padovani, A. J. Viterbi, L. A. Weaver, Jr., and C. E. Wheatley, III, "On the capacity of a cellular CDMA system," *IEEE Trans. Veh. Technol.*, vol. 40, no. 2, pp. 303–312, May 1991.

[42] F. Adachi, M. Sawahashi, and H. Suda, "Wideband DS-CDMA for next generation mobile communications systems," *IEEE Commun. Mag.*, vol. 36, no. 9, pp. 56–69, Sep. 1998.

[43] M. Peng, W. Wang, and H.-H. Chen, "TD-SCDMA evolution," *IEEE Veh. Technol. Mag.*, vol. 5, no. 2, pp. 28–41, Jun. 2010.

[44] L. Ping, L. Liu, K. Wu, and W. K. Leung, "Interleave-division multiple access," *IEEE Trans. Wireless Commun.*, vol. 5, no. 4, pp. 938–947, Apr. 2006.

[45] H. Schoeneich and P. Hoeher, "Iterative pilot-layer aided channel estimation with emphasis on interleave-division multiple access systems," *EURASIP J. Appl. Signal Process.*, vol. 2006, pp. 1–15, Jan. 2006.

[46] Y. Hong and L. K. Rasmussen, "Iterative switched decoding for interleave-division multiple-access systems," *IEEE Trans. Veh. Technol.*, vol. 57, no. 3, pp. 1939–1944, May 2008.

[47] B. Senanayake, M. C. Reed, and Z. Shi, "Timing acquisition for multi-user IDMA," in *Proc. IEEE ICC*, Beijing, China, May 2008, pp. 5077–5081.

[48] R. Zhang and L. Hanzo, "Three design aspects of multicarrier interleave division multiple access," *IEEE Trans. Veh. Technol.*, vol. 57, no. 6, pp. 3607–3617, Nov. 2008.

[49] X. Xiong, J. Hu, F. Yang, and X. Ling, "Effect of channel estimation error on the performance of interleave-division multiple access systems," in *Proc. IEEE 11th ICACT*, Feb. 2009, pp. 1538–1542.

[50] T. Yang, J. Yuan, and Z. Shi, "Rate optimization for IDMA systems with iterative joint multi-user decoding," *IEEE Trans. Wireless Commun.*, vol. 8, no. 3, pp. 1148–1153, Mar. 2009.

[51] B. Cristea, D. Roviras, and B. Escrig, "Turbo receivers for interleave-division multiple-access systems," *IEEE Trans. Commun.*, vol. 57, no. 7, pp. 2090–2097, Jul. 2009.

[52] H. Chung, Y.-C. Tsai, and M.-C. Lin, "IDMA using non-gray labelled modulation," *IEEE Trans. Commun.*, vol. 59, no. 9, pp. 2492–2501, Sep. 2011.

[53] K. Kusume, G. Bauch, and W. Utschick, "IDMA vs. CDMA: Analysis and comparison of two multiple access schemes," *IEEE Trans. Wireless Commun.*, vol. 11, no. 1, pp. 78–87, Jan. 2012.

[54] P. Hammarberg and F. Rusek, "Channel estimation algorithms for OFDM-IDMA: Complexity and performance," *IEEE Trans. Wireless Commun.*, vol. 11, no. 5, pp. 1722–1732, May 2012.

[55] I. M. Mahafeno, C. Langlais, and C. Jego, "OFDM-IDMA versus IDMA with ISI cancellation for quasistatic Rayleigh fading multipath channels," in *Proc. 4th Int. Symp. Turbo Codes Rel. Topics*, Munich, Germany, Apr. 2006, pp. 1–6.

[56] C. Novak, G. Matz, and F. Hlawatsch, "IDMA for the multiuser MIMO-OFDM uplink: A factor graph framework for joint data detection and channel estimation," *IEEE Trans. Signal Process.*, vol. 61, no. 16, pp. 4051–4066, Aug. 2013.

[57] K. Wu, K. Anwar, and T. Matsumoto, "BICM-ID-based IDMA using extended mapping," *IEICE Trans. Commun.*, vols. E97–B, no. 7, pp. 1483–1492, Jul. 2014.

[58] L. Liu, Y. Li, Y. Su, and Y. Sun, "Quantize-and-forward strategy for interleave-division multiple-access relay channel," *IEEE Trans. Veh. Technol.*, vol. 65, no. 3, pp. 1808–1814, Mar. 2016.

[59] L. Bing, T. Aulin, B. Bai, and H. Zhang, "Design and performance analysis of multiuser CPM with single user detection complexity," *IEEE Trans. Wireless Commun.*, vol. 15, no. 6, pp. 4032–4044, Jun. 2016.

[60] G. Song and J. Cheng, "Distance enumerator analysis for interleave-division multi-user codes," *IEEE Trans. Inf. Theory*, vol. 62, no. 7, pp. 4039–4053, Jul. 2016.

[61] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon limit error-correcting coding and decoding: Turbo-codes. 1," in *Proc. IEEE Int. Conf. Commun. (ICC)*, vol. 2. Geneva, Switzerland, May 1993, pp. 1064–1070.

[62] D. J. C. MacKay and R. M. Neal, "Near Shannon limit performance of low density parity check codes," *Electron. Lett.*, vol. 32, no. 18, p. 1645, Aug. 1996.

[63] S. Lin and J. D. J. Costello, *Error Control Coding: Fundamentals and Applications*, 2nd ed. Upper Saddle River, NJ, USA: Prentice-Hall, 2004.

[64] T. Richardson and R. Urbanke, *Modern Coding Theory*. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[65] L. Ping, L. Liu, and W. K. Leung, "A simple approach to near-optimal multiuser detection: Interleaver-division multiple-access," in *Proc. IEEE WCNC*, New Orleans, LA, USA, Mar. 2003, pp. 391–396.

[66] L. Ping and L. Liu, "Analysis and design of IDMA systems based on SNR evolution and power allocation," in *Proc. IEEE VTC-Fall*, Los Angeles, CA, USA, Sep. 2004, pp. 1068–1072.

[67] L. Liu, J. Tong, and L. Ping, "Analysis and optimization of CDMA systems with chip-level interleavers," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 1, pp. 141–150, Jan. 2006.

[68] L. Hanzo, M. El-Hajjar, and O. Alamri, "Near-capacity wireless transceivers and cooperative communications in the MIMO era: Evolution of standards, waveform design, and future perspectives," *Proc. IEEE*, vol. 99, no. 8, pp. 1343–1385, Aug. 2011.

[69] G. Wunder *et al.*, "5GNOW: Non-orthogonal, asynchronous waveforms for future mobile applications," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 97–105, Feb. 2014.

[70] A. Osseiran *et al.*, "Scenarios for 5G mobile and wireless communications: The vision of the METIS project," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 26–35, May 2014.

[71] V. Jungnickel *et al.*, "The role of small cells, coordinated multipoint, and massive MIMO in 5G," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 44–51, May 2014.

[72] J. G. Andrews *et al.*, "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[73] *Sparse Code Multiple Access (SCMA) for 5G Radio Transmission*, document R1-162155, Huawei, HiSilicon, 3GPP TSG RAN WG1 Meeting, Busan, Korea, Apr. 2016.

[74] *Non-Orthogonal Multiple Access Candidate for NR*, document R1-163992, Samsung, 3GPP TSG RAN WG1 Meeting, Nanjing, China, May 2016.

[75] *Uplink Contentionbased Access in 5G New Radio*, document R1-165022, Nokia, Alcatel-Lucent Shanghai Bell, 3GPP TSG RAN WG1 Meeting, Nanjing, China, May 2016.

[76] P. Li, L. Liu, K. Wu, and W. Leung, "Interleave division multiple access (IDMA) communication systems," in *Proc. 3rd Int. Symp. Turbo Codes Rel. Topics*, Brest, France, Sep. 2003, pp. 173–180.

[77] M. Zhao, Z. Shi, and M. C. Reed, "Iterative turbo channel estimation for OFDM system over rapid dispersive fading channel," *IEEE Trans. Wireless Commun.*, vol. 7, no. 8, pp. 3174–3184, Aug. 2008.

[78] C.-K. Wen, C.-J. Wang, S. Jin, K.-K. Wong, and P. Ting, "Bayes-optimal joint channel-and-data estimation for massive MIMO with low-precision ADCs," *IEEE Trans. Signal Process.*, vol. 64, no. 10, pp. 2541–2556, May 2016.

[79] W. Yu, "Uplink-downlink duality via minimax duality," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 361–374, Feb. 2006.

[80] S. Hur, T. Kim, D. J. Love, J. V. Krogmeier, T. A. Thomas, and A. Ghosh, "Millimeter wave beamforming for wireless backhaul and access in small cell networks," *IEEE Trans. Commun.*, vol. 61, no. 10, pp. 4391–4403, Oct. 2013.

[81] O. El Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, Jr., "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, Mar. 2014.

[82] P. Wang, Y. Li, X. Yuan, L. Song, and B. Vucetic, "Tens of gigabits wireless communications over E-band LoS MIMO channels with uniform linear antenna arrays," *IEEE Trans. Wireless Commun.*, vol. 13, no. 7, pp. 3791–3805, Jul. 2014.

[83] P. Wang, Y. Li, L. Song, and B. Vucetic, "Multi-gigabit millimeter wave wireless communications for 5G: From fixed access to cellular networks," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 168–178, Jan. 2015.

[84] H. Ghauch, T. Kim, M. Bengtsson, and M. Skoglund, "Subspace estimation and decomposition for large millimeter-wave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 528–542, Apr. 2016.

[85] R. W. Heath, N. Gonzalez-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 436–453, Apr. 2016.

[86] X. Wu *et al.*, "60 GHz millimeter-wave channel measurements and modeling for indoor office environments," *IEEE Trans. Antennas Propag.*, vol. 65, no. 4, pp. 1912–1924, Apr. 2017.

[87] B. Hochwald, T. Marzetta, and V. Tarokh, "Multi-antenna channel-hardening and its implications for rate feedback and scheduling," *IEEE Trans. Inf. Theory*, vol. 50, no. 9, pp. 1893–1909, Sep. 2004.

[88] M. Mohseni, R. Zhang, and J. M. Cioffi, "Optimized transmission of fading multiple-access and broadcast channels with multiple antennas," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1627–1639, Aug. 2006.

[89] C. Yang, W. Wang, Y. Qian, and X. Zhang, "A weighted proportional fair scheduling to maximize best-effort service utility in multicell network," in *Proc. IEEE 19th Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, Sep. 2008, pp. 1–5.

[90] J. Lee and N. Jindal, "Symmetric capacity of MIMO downlink channels," in *Proc. IEEE ISIT*, Washington, DC, USA, Jul. 2006, pp. 1031–1035.

[91] C. Hellings, M. M. Joham, M. Riemensberger, and W. Utschick, "Minimal transmit power in parallel vector broadcast channels with linear precoding," *IEEE Trans. Signal Process.*, vol. 60, no. 4, pp. 1890–1898, Apr. 2012.

[92] 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, no. 8, pp. 1912–1921, Aug. 2003.

[93] H. Huh, S.-H. Moon, Y.-T. Kim, I. Lee, and G. Caire, "Multi-cell MIMO downlink with cell cooperation and fair scheduling: A large-system limit analysis," *IEEE Trans. Inf. Theory*, vol. 57, no. 12, pp. 7771–7786, Dec. 2011.

[94] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 2006.

[95] N. Jindal and A. Goldsmith, "Dirty paper coding versus TDMA for MIMO broadcast channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 5, pp. 1783–1794, May 2005.

[96] A. Goldsmith, S. A. Jafar, N. Jindal, and S. Vishwanath, "Capacity limits of MIMO channels," *IEEE J. Sel. Areas Commun.*, vol. 51, no. 6, pp. 684–702, Jun. 2003.

[97] T. Yoo and A. Goldsmith, "Capacity and power allocation for fading MIMO channels with channel estimation error," *IEEE Trans. Inf. Theory*, vol. 52, no. 5, pp. 2203–2214, May 2006.

[98] M. Joham, W. Utschick, and J. A. Nossek, "Linear transmit processing in MIMO communications systems," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 2700–2712, Aug. 2005.

[99] N. Wiberg, H.-A. Loeliger, and R. Kotter, "Codes and iterative decoding on general graphs," *Eur. Trans. Telecomm.*, vol. 6, pp. 513–525, Sep./Oct. 1995.

[100] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 498–519, Feb. 2001.

[101] J. Thorpe, "Low-density parity-check (LDPC) codes constructed from protographs," INP, Pasadena, CA, USA, Tech. Rep. 42-154, Aug. 2003, pp. 1–7.

[102] M. Noemm, T. Wo, and P. Hoeher, "Multilayer app detection for IDM," *Electron. Lett.*, vol. 46, no. 1, pp. 96–97, Jan. 2010.

[103] J. Tong, L. Ping, Z. Zhang, and V. K. Bhargava, "Iterative soft compensation for OFDM systems with clipping and superposition coded modulation," *IEEE Trans. Commun.*, vol. 58, no. 10, pp. 2861–2870, Oct. 2010.

[104] C. Douillard *et al.*, "Iterative correction of intersymbol interference: Turbo-equalization," *Eur. Trans. Telecommun.*, vol. 6, no. 5, pp. 507–511, Sep. 1995.

[105] G. Caire, G. Taricco, and E. Biglieri, "Bit-interleaved coded modulation," *IEEE Trans. Inf. Theory*, vol. 44, no. 3, pp. 927–946, May 1998.

[106] L. Ping, J. Tong, X. Yuan, and Q. Guo, "Superposition coded modulation and iterative linear MMSE detection," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 6, pp. 995–1004, Aug. 2009.

[107] I. Pupeza, A. Kavcic, and L. Ping, "Efficient generation of interleavers for IDMA," in *Proc. IEEE ICC*, Istanbul, Turkey, Jun. 2006, pp. 1508–1513.

[108] D. Hao and P. A. Hoeher, "Helical interleaver set design for interleave-division multiplexing and related techniques," *IEEE Commun. Lett.*, vol. 12, no. 11, pp. 843–845, Nov. 2008.

[109] R. Hoshyar, F. Wathan, and R. Tafazolli, "Novel low-density signature for synchronous CDMA systems over AWGN channel," *IEEE Trans. Signal Process.*, vol. 56, no. 4, pp. 1616–1626, Apr. 2008.

[110] J. van de Beek and B. Popovi, "Multiple access with low-density signatures," in *Proc. IEEE ICC*, Dresden, Germany, Jun. 2009, pp. 1–6.

[111] P. Hoeher, H. Schoeneich, and J. C. Fricke, "Multi-layer interleave-division multiple access: Theory and practice," *Eur. Trans. Telecommun.*, vol. 19, no. 5, pp. 523–536, Aug. 2008.

[112] H. Nikopour and H. Baligh, "Sparse code multiple access," in *Proc. IEEE 24th Int. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, Sep. 2013, pp. 332–336.

[113] X. Ma and L. Ping, "Coded modulation using superimposed binary codes," *IEEE Trans. Inf. Theory*, vol. 50, no. 12, pp. 3331–3343, Dec. 2004.

[114] T. Yang and J. Yuan, "Performance of iterative decoding for superposition modulation-based cooperative transmission," *IEEE Trans. Wireless Commun.*, vol. 9, no. 1, pp. 51–59, Jan. 2010.

[115] P. A. Hoeher and T. Wo, "Superposition modulation: Myths and facts," *IEEE Commun. Mag.*, vol. 49, no. 12, pp. 110–116, Dec. 2011.

[116] M. Taherzadeh, H. Nikopour, A. Bayesteh, and H. Baligh, "SCMA codebook design," in *Proc. IEEE 80th Veh. Technol. Conf. (VTC Fall)*, Sep. 2014, pp. 1–5.

[117] S. Kudekar, T. J. Richardson, and R. L. Urbanke, "Threshold saturation via spatial coupling: Why convolutional LDPC ensembles perform so well over the BEC," *IEEE Trans. Inf. Theory*, vol. 57, no. 2, pp. 803–834, Feb. 2011.

[118] K. Takeuchi, T. Tanaka, and T. Kawabata, "Improvement of BP-based CDMA multiuser detection by spatial coupling," in *Proc. IEEE ISIT*, Saint-Petersburg, Russian, Aug. 2011, pp. 1489–1493.

[119] Z. Zhang, C. Xu, and L. Ping, "Spatially coupled LDPC coding and linear precoding for MIMO systems," *IEICE Trans. Commun.*, vol. E95-B, no. 12, pp. 3663–3670, Dec. 2012.

[120] C. Schlegel and D. Truhachev, "Multiple access demodulation in the lifted signal graph with spatial coupling," *IEEE Trans. Inf. Theory*, vol. 59, no. 4, pp. 2459–2470, Apr. 2013.

[121] D. J. Costello, L. Dolecek, T. Fuja, J. Kliewer, D. Mitchell, and R. Smarandache, "Spatially coupled sparse codes on graphs: Theory and practice," *IEEE Commun. Mag.*, vol. 52, no. 7, pp. 168–176, Jul. 2014.

[122] S. Kumar, A. J. Young, N. Macris, and H. D. Pfister, "Threshold saturation for spatially coupled LDPC and LDGM codes on BMS channels," *IEEE Trans. Inf. Theory*, vol. 60, no. 12, pp. 7389–7415, Dec. 2014.

[123] C. Liang, J. Ma, and L. Ping, "Towards Gaussian capacity, universality and short block length," in *Proc. 9th Int. Symp. Turbo Codes*, Brest, France, Sep. 2016, pp. 412–416.

[124] C. Liang, J. Ma, and L. Ping, "Compressed FEC codes with spatial-coupling," *IEEE Commun. Lett.*, vol. 21, no. 5, pp. 987–990, May 2017.

[125] P. Wang, L. Ping, and L. Liu, "Power allocation for multiple access systems with practical coding and iterative multi-user detection," in *Proc. IEEE ICC*, Istanbul, Turkey, Dec. 2007, pp. 3179–3183.

[126] P. Wang and L. Ping, "On multi-user gain in MIMO systems with rate constraints," in *Proc. IEEE GLOBECOM*, Washington, DC, USA, Dec. 2007, pp. 3179–3183.

[127] E. A. Gharavol, Y.-C. Liang, and K. Mouthaan, "Robust downlink beamforming in multiuser MISO cognitive radio networks with imperfect channel-state information," *IEEE Trans. Veh. Technol.*, vol. 59, no. 6, pp. 2852–2860, Jun. 2010.

[128] W. Shin, M. Vaezi, B. Lee, D. J. Love, J. Lee, and H. V. Poor. (2016). "Non-orthogonal multiple access in multi-cell networks: Theory, performance, and practical challenges." [Online]. Available: https://arxiv.org/abs/1611.01607

[129] J. Wang, F. Adachi, and X. Xia, "Coordinated and distributed MIMO," *IEEE Wireless Commun.*, vol. 17, no. 3, pp. 24–25, Jun. 2010.

[130] W. W. L. Ho, T. Q. S. Quek, S. Sun, and R. W. Heath, "Decentralized precoding for multicell MIMO downlink," *IEEE Trans. Wireless Commun.*, vol. 10, no. 6, pp. 1798–1809, Jun. 2011.

[131] P. Wang, H. Wang, L. Ping, and X. Lin, "On the capacity of MIMO cellular systems with base station cooperation," *IEEE Trans. Wireless Commun.*, vol. 10, no. 11, pp. 3720–3731, Nov. 2011.

[132] J. Zhang, X. Yuan, and L. Ping, "Hermitian precoding for distributed MIMO systems with individual channel state information," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 241–250, Feb. 2013.

[133] J. Wang and L. Dai, "Downlink rate analysis for virtual-cell based large-scale distributed antenna systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 3, pp. 1998–2011, Mar. 2016.

[134] C. Xu, P. Wang, Z. Zhang, and L. Ping, "Transmitter design for uplink MIMO systems with antenna correlation," *IEEE Trans. Wireless Commun.*, vol. 14, no. 4, pp. 1772–1784, Apr. 2015.

[135] C. Xu, L. Ping, P. Wang, S. Chan, and X. Lin, "Decentralized power control for random access with successive interference cancellation," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 11, pp. 2387–2396, Nov. 2013.

[136] C. Xu, X. Wang, and L. Ping, "Random access with massive-antenna arrays," in *Proc. IEEE VTC-Spring*, Nanjing, China, May 2016, pp. 1–5.

[137] L. Lu, L. You, and S. C. Liew, "Network-coded multiple access," *IEEE Trans. Mobile Comput.*, vol. 13, no. 12, pp. 2853–2869, Dec. 2014.

[138] H. Lin, K. Ishibashi, W.-Y. Shin, and T. Fujii, "A simple random access scheme with multilevel power allocation," *IEEE Commun. Lett.*, vol. 19, no. 12, pp. 2118–2121, Dec. 2015.

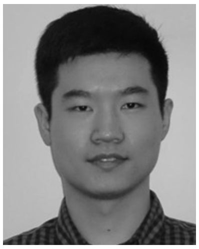
[139] M. Zou, S. Chan, H. Vu, and L. Ping, "Throughput improvement of 802.11 networks via randomization of transmission power levels," *IEEE Trans. Veh. Technol.*, vol. 65, no. 4, pp. 2703–2714, Apr. 2016.

[140] S. M. Kay, *Fundamentals Statistical Signal Processing: Estimation Theory*. Upper Saddle River, NJ, USA: Prentice-Hall, 1993.

CHONGBIN XU received the B.S. degree in information engineering from Xi'an Jiaotong University in 2005 and the Ph.D. degree in information and communication engineering from Tsinghua University in 2012. Since 2014, he has been with the Department of Communication Science and Engineering, Fudan University, China. His research interests are in the areas of signal processing and communication theory, including linear precoding, iterative detection, and random access techniques.

YANG HU received the B.S. degree from Tianjin University, China, in 2011, and the M.S. degree from Tsinghua University, China, in 2014. He is currently pursuing the Ph.D. degree at the City University of Hong Kong. His research interests include information theory, multiuser detection, and random access techniques.

CHULONG LIANG received the B.E. degree in communication engineering and the Ph.D. degree in communication and information systems from Sun Yat-sen University, Guangzhou, China, in 2010 and 2015, respectively. He is currently a Post-Doctoral Fellow with the City University of Hong Kong, Hong Kong. His current research interests include channel coding theory and its applications to communication systems.



JUNJIE MA received the B.E. degree from Xidian University, China, in 2010, and the Ph.D. degree from the City University of Hong Kong in 2015. He was a Research Fellow with the Department of Electronic Engineering, City University of Hong Kong, from 2015 to 2016. Since 2016, he has been a Post-Doctoral Researcher with the Department of Statistics, Columbia University. His current research interests include statistical signal processing, compressed sensing, and optimization methods.

LI PING (S'87–M'91–SM'06–F'10) received the Ph.D. degree from Glasgow University in 1990. He was a Lecturer with the Department of Electronic Engineering, Melbourne University, from 1990 to 1992, and a Member of Research Staff with Telecom Australia Research Laboratories from 1993 to 1995. He has been with the Department of Electronic Engineering, City University of Hong Kong, since 1996, where he is currently a Chair Professor. He served as a member of Board of Governors of the IEEE Information Theory Society from 2010 to 2012. He was the recipient of a British Telecom-Royal Society Fellowship in 1986, the IEE J. J. Thomson Premium in 1993, the Croucher Foundation Award in 2005, and a British Royal Academy of Engineering Distinguished Visiting Fellowship in 2010.