IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 19, 2019, accepted September 3, 2019, date of publication September 11, 2019,
date of current version September 26, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2940458

Multiuser Massive MIMO Systems With Time-Offset Pilots and Successive Interference Cancellation

THE KHAI NGUYEN[1], HA H. NGUYEN[1], (Senior Member, IEEE), AND TIEN HOA NGUYEN[2]
[1]Department of Electrical and Computer Engineering, University of Saskatchewan, Saskatoon, SK S7N 5A9, Canada
[2]School of Electronics and Telecommunications, Hanoi University of Science and Technology (HUST), Hanoi 100000, Vietnam

Corresponding author: Ha H. Nguyen (ha.nguyen@usask.ca)

This work was supported in part by the Natural Sciences and Engineering Research Council (NSERC) Discovery Grant (DG) under Grant RGPIN-2017-05899, and in part by the Vietnam's Ministry of Education and Training (MOET) Project under Grant B2019-BKA-10.

ABSTRACT This paper proposes time-offset pilots for a single-cell multiuser massive multiple-input multiple-output (MIMO) system and studies its performance under the minimum mean-squared error channel estimator and successive interference cancellation. With the proposed time-offset pilots, users are divided into two groups and the uplink pilots from one group are transmitted *simultaneously* with the uplink data of the other group, which allows the system to accommodate more users for a given number of pilots. Successive interference cancellation is developed to ease the effect of pilot contamination and enhance data detection. Closed-form expressions for lower bounds of the uplink spectral efficiencies in both the training and data phases are derived when the maximum-ratio combining receiver is used at the base station. The power control problem is formulated with the objective of either maximizing the quality of service that can be equally provided to all users, or minimizing the total transmit power. Since the original power control problems are NP-hard, we also propose algorithms based on the bisection method to solve the problems separately in training and data phases. Analysis and numerical results show that the effect of pilot contamination can be mitigated by successive interference cancellation and proper power control.

INDEX TERMS Massive MIMO, single-cell networks, pilot design, energy efficiency, spectrum efficiency, power control, optimization.

I. INTRODUCTION

Over the last decade, massive multiple-input multiple-output (MIMO) systems have gained a strong interest as a promising key technology for enabling the next and future generations of wireless communications. With hundreds of antennas equipped at each base station (BS), a massive MIMO system allows multiple users to simultaneously operate in the same time-frequency blocks, while co-channel interference can be effectively mitigated as a result of channel hardening and favorable propagation effects [1]–[4]. Furthermore, by utilizing proper power control algorithms, massive MIMO systems have the ability to achieve very high spectral efficiency (SE) and energy efficiency (EE) [5]–[7].

However, performance of a massive MIMO system is limited by the quality of channel estimation [8]–[10]. As discussed in these papers, in every coherence interval where the wireless channels between BSs and users are approximately constant, the number of symbols spent for channel estimation directly determines the maximum number of pairwise-orthogonal pilot sequences that can be generated for channel estimation. Conventionally and preferably, pilot sequences are designed to be mutually orthogonal and distinct pilot sequences are assigned to different users to avoid pairwise correlation between them. Unfortunately, the number of orthogonal pilot sequences could be limited by the small length of the coherence interval, especially when the propagation environment changes quickly. Therefore, if the number of users served by one BS keeps increasing, pilot sequences must be reused, resulting in the so-called pilot contamination [11]–[13]. As a consequence, with simple linear receivers such as maximum-ratio combining and zero-forcing, the network's SE becomes saturated, even when the number of antennas goes to infinity [14]–[17].

The associate editor coordinating the review of this manuscript and approving it for publication was Xianfu Lei.

A. RELATED WORKS

There have been many research works addressing the pilot contamination problem in multi-cell massive MIMO systems [1], [18]–[21]. For multi-cell massive MIMO systems, the basic approach to reduce the effect of pilot contamination is reusing pilots [16], [19], [22], [23]. With this approach, an arbitrary pilot sequence can be assigned one time only within a cluster of ϑ cells. This has been investigated in [16] and it was shown that using a higher pilot reuse factor helps to lessen pilot contamination. It should be pointed out that, a larger value of ϑ implies that the cell size, as well as the number of users who can be served within each cell, are reduced. Another method to reduce the effect of pilot contamination in a multi-cell massive MIMO system is to use different pilot sets [18]. Specifically, from a basic mutually-orthogonal pilot set, the authors in [18] construct the so-called dictionary of linear combinations of the original pilot set to exploit the degree of freedom, which is demonstrated to lower the interference level during the training phase. With this method, non-orthogonal pilots are used even within a cell.

All the works discussed above are for multi-cell massive MIMO systems where users are geographically separated into a cellular topology and hence, pilots can be reused across cells with large distance separations [9]. On the other hand, the joint pilot and payload power control problem in a single-cell massive MIMO system is investigated in [24]. In this work, the authors show that the optimal number of pilots should be set equal to the number of users in the system because using orthogonal pilots maximizes the signal-to-interference-plus-noise ratio (SINR). However, when the number of users increases and/or the coherence interval is short (as seen in fast-varying channels), the total throughput inversely decreases with the number of pilots. Another work examining the pilot contamination problem can be found in [6]. In this paper, a cell-free massive MIMO system with multiple access points (APs) is considered. As explained in [6], during the training phase, a set of orthogonal pilots can be assigned to a larger number of users by using a greedy algorithm. This assignment was shown to provide an improvement of approximately 10% in spectral efficiency as compared to a random pilot assignment. However, via the large-antenna analysis, it is shown in [6], [19] that if a pilot sequence is assigned to more than one user, the SINR is still upper-bounded because not only the desire signal power, but also the correlated interference power caused by pilot contamination increases proportionally with the number of antennas.

In all the works discussed above, uplink (UL) pilots are transmitted at the same time for all users. This method is known as aligned pilots in [1] or synchronous pilots in [25]. Another method to deal with pilot contamination is using time-offset (or asynchronous) pilots [1], [21], [25]–[27]. In particular, the authors in [1], [21] propose to schedule UL pilots so that the pilot signaling of one cell can be carried out while other cells are transmitting downlink (DL) data. Using the large-antenna analysis, these papers show that with such a pilot design, when the number of antennas goes to infinity, the SINRs in both UL and DL increase proportionally with the number of antennas. In addition, the authors also point out that having users in one cell transmitting UL pilots while users in other cells are transmitting UL data is not optimal because the performance is saturated when the number of antennas goes up to infinity.

To address the disadvantage of transmitting UL pilots simultaneously with UL data, the authors in [25], [26] propose a semi-blind pilot decontamination scheme. In such a scheme, under the assumption of *time-invariant* channel, least-square estimation of the channel is obtained by UL pilot sequences and with the aid of UL data extraction. This method is shown to significantly improve the quality of channel estimation when the length of data increases. However, such an improvement is difficult to achieve in the case of *fast-varying* channels as demonstrated in [10]. In particular, the authors in [10] show that, in practice, in order to allow data transmission plus channel estimation, the number of users needs to be well below the coherence length. The authors then propose a blind pilot decontamination method in which the pilot data is not required to find a subspace projection, which is used to improve channel estimation. Other research works on combating pilot contamination with time-offset pilots for multi-cell massive MIMO can be found in [28], [29] which introduce new coherence block structures with extra intervals for BS channel estimation [28] or null transmission [29]. However, if the coherence length is short, spending more symbols for channel estimation may result in an insufficient time interval for data transmission [10].

Another emerging technique to accommodate more users without requiring extra pilots is beam-domain user grouping for massive MIMO [30]–[32]. In these papers, the authors introduce a beam-domain grouping method that assigns users into different groups based on the direction of arrival (DOA) and then reuse pilots in different groups. The channel vector's elements are assumed to be correlated with an array response vector, which allows a beam-domain presentation of the actual channel. With such a method, it is shown that the training resources can be reduced, whereas inter-group interference and self interference at the BS can be effectively mitigated thanks to the properties of the beam-domain channel.

B. CONTRIBUTION

In this paper, we investigate a new approach with time-offset pilots in a single-cell massive MIMO system. For the system considered in this paper, all users are divided into two groups. During the training phase, one group transmits orthogonal pilot signals, while the other group sends data signals. The BS gathers all pilot signals and performs the minimum mean-square error (MMSE) channel estimation. With this method, channel estimation is not contaminated by correlation between pilots, but by the data transmitted

by the other group, whose power is typically much lower than the pilot power. In addition, with a fixed number of pairwise orthogonal pilot sequences, this approach allows to double the number of users compared to the orthogonal pilot approach. Different from previous works, in which the pilot power is usually set at the maximal level to maximize the channel estimation quality [6], [33], or assigned based on a long-term average power constraint [24], our work takes into account both pilot power and data power to optimally allocate users' UL power to satisfy a predetermined cost function. Moreover, we also develop a successive interference cancellation method that does not require the perfect channel state information. The method is shown to be able to significantly suppress the interference caused by pilot contamination. Naturally, this advantage comes at the expense of higher implementation complexity. The main contributions of the paper are as follows:

- We derive a closed-form expression of the UL ergodic spectral efficiency for the proposed time-offset pilot method under Rayleigh fading channels and when the maximum ratio combining (MRC) is used at the BS. Many interesting observations concerning the effects of array gain, interference, and additive noise are revealed.
- We develop a successive interference cancellation method for the detection of UL data at the BS to mitigate the impact of pilot contamination in UL transmission. Under the assumption of ideal error-free detection, it is shown that the UL SE is no longer bounded when the number of antennas increases.
- We formulate and solve the power control problem for two different cost functions: the first problem focuses on maximizing the minimum quality of service (QoS) or max-min QoS, whereas the second problem is on total power minimization. Because of the NP-hardness of the original problems, we propose algorithms based on the bisection method to decompose these NP-hard problems into two subproblems which can be solved in polynomial time.
- A group assignment method is also proposed to mitigate the interference that cannot be removed by the MRC. The proposed group assignment helps to further improve the UL ergodic spectral efficiency.

The remainder of this paper is organized as follows. Section II presents the model of a single-cell multiuser massive MIMO system with time-offset pilots and channel estimation. Section III analyzes UL spectral efficiencies in both training and data phases. Section IV studies power control problems. Section V provides simulation results and discussion. Section VI concludes the paper.

Notations: Vectors are formatted in bold lower-case, matrices are in bold upper-case. The transpose and conjugate transpose are denoted with superscripts T and H, respectively. The $K \times K$ identity matrix is $\boldsymbol{I}_K$. The operator $\mathbb{E}\{\cdot\}$ denotes the expectation of a random variable. The notation $\|\cdot\|$ stands for the Euclidean norm and tr$(\cdot)$ represents the trace of a matrix. The notation $\boldsymbol{n} \sim \mathcal{CN}(0, \boldsymbol{C})$ means that $\boldsymbol{n}$ is a zero-mean complex Gaussian vector with covariance matrix $\boldsymbol{C}$.

II. TIME-OFFSET PILOTS AND CHANNEL ESTIMATION

Consider a single-cell multi-user massive MIMO system in which one M-antenna BS serves N users, who are randomly distributed over the cell. The channels between the users and the BS are assumed to be frequency flat and approximately constant within a coherence interval of length τ_c symbols. The UL and DL transmissions in the system operate in time-division duplex (TDD) mode. As a result, conventional pilot designs can take advantage of channel reciprocity to estimate both UL and DL channels within a coherence interval. In massive MIMO systems, pilot sequences are usually transmitted synchronously by all users at the same time. This is problematic if the coherence interval is short, since to maintain orthogonal pilots, a smaller number of symbol periods can be used for data transmission. Motivated by the work in [6], we consider time-offset pilot design as illustrated in Fig. 1. Here, N users in the system are separated into $G = 2$ groups, each having $K = N/2$ users and taking turn to transmit UL pilots in different time slots. To improve the SE, the transmission of UL pilots by one group happens concurrently with UL data transmission from the other group. An important point to note is that pilot transmission must be carried out at the beginning of every coherence interval.

Dropping the block index for simplicity and without loss of generality, the $M \times 1$ received signal vector at the BS in one symbol time can be generally written as:

$$\boldsymbol{y} = \sum_{k=1}^{K} \left(\boldsymbol{h}_{1,k} \sqrt{p_{1,k}} x_{1,k} + \boldsymbol{h}_{2,k} \sqrt{p_{2,k}} x_{2,k} \right) + \boldsymbol{n}, \qquad (1)$$

where $x_{g,k}$ $(g = 1, 2)$ is the transmit signal of the kth user in the gth group that is normalized to have unit power, i.e., $\mathbb{E}\left\{|x_{g,k}|^2\right\} = 1$, whereas the actual transmit power is specified by $p_{g,k}$. Note that $x_{g,k}$ represents either the data or the pilot symbol during the training phase (see the illustration in Fig. 1). The term $\boldsymbol{n} \sim \mathcal{CN}\left(0, \sigma^2 \boldsymbol{I}_M\right)$ models additive white Gaussian noise (AWGN) at the BS. The channels are assumed to be uncorrelated Rayleigh fading. That is, the channel vector $\boldsymbol{h}_{g,k}$ from the kth user of the gth group to the BS is modeled as having a circularly-symmetric complex Gaussian distribution, $\boldsymbol{h}_{g,k} \sim \mathcal{CN}\left(0, \beta_{g,k} \boldsymbol{I}_M\right)$, where $\beta_{g,k}$ represents large-scale fading.

The BS does not know the exact channel coefficients but the channel statistics. To estimate the channels for each user group, a set of K length-τ_p pilot sequences is used. These pilots are collectively represented by a $\tau_p \times K$ pilot matrix $\boldsymbol{\Phi} = \left[\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_K\right]$ which satisfies $\boldsymbol{\Phi}^H \boldsymbol{\Phi} = \tau_p \boldsymbol{I}_K$. Usually, the pilot length is set at the minimum value $\tau_p = K$ in order to achieve the orthogonality between pilot sequences.

Without loss of generality, suppose that the first group transmits pilots first at the beginning of the training phase, while the other group transmits data. Then the signal matrix $\boldsymbol{Y} \in \mathbb{C}^{M \times \tau_p}$ received at the BS over τ_p time slots (symbols)

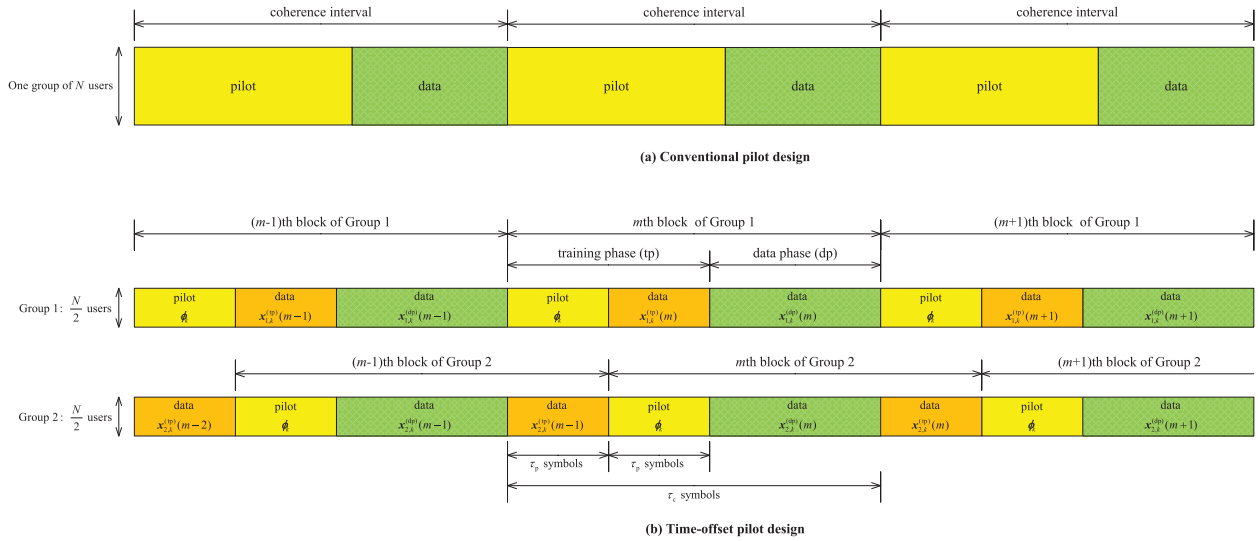


FIGURE 1. (a) Conventional pilot design, and (b) Time-offset pilot design.

is given as:

$$\boldsymbol{Y} = \sum_{k=1}^{K} \left(\boldsymbol{h}_{1,k} \sqrt{\rho_{1,k}^{(\mathrm{p})}} \boldsymbol{\phi}_k^H + \boldsymbol{h}_{2,k} \sqrt{\rho_{2,k}^{(\mathrm{d})}} \boldsymbol{x}_{2,k}^{(\mathrm{tp})} \right) + \boldsymbol{N}, \quad (2)$$

where $\rho_{1,k}^{(\mathrm{p})}$ is the pilot power, $\rho_{2,k}^{(\mathrm{d})}$ is the power assigned to the normalized data signal vector $\boldsymbol{x}_{2,k}^{(\mathrm{tp})} \in \mathbb{C}^{1\times\tau_p}$ of the kth user in the second group during τ_p time slots of the training phase, which satisfies $\boldsymbol{x}_{2,k}^{(\mathrm{tp})} \sim \mathcal{CN}\left(0, \boldsymbol{I}_{\tau_p}\right)$.

To estimate the channel from the qth user in the first group, the BS multiplies the received signal with the corresponding pilot of the qth user. This results in:

$$\boldsymbol{r}_{1,q} = \boldsymbol{Y} \frac{\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|} = \boldsymbol{h}_{1,q} \sqrt{\rho_{1,q}^{(\mathrm{p})}} \tau_p + \sum_{k=1}^{K} \boldsymbol{h}_{2,k} \sqrt{\rho_{2,k}^{(\mathrm{d})}} \boldsymbol{x}_{2,k}^{(\mathrm{tp})} \frac{\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|} + \boldsymbol{N} \frac{\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|}. \quad (3)$$

Then, the estimate of $\boldsymbol{h}_{1,q}$ can be obtained by using the MMSE estimator [34] as:

$$\hat{\boldsymbol{h}}_{1,q} = \frac{\mathrm{cov}\left\{\boldsymbol{h}_{1,q}, \boldsymbol{r}_{1,q}\right\}}{\mathrm{var}\left\{\boldsymbol{r}_{1,q}\right\}} \boldsymbol{r}_{1,q} = \mu_{1,q} \boldsymbol{r}_{1,q}, \quad (4)$$

where

$$\mu_{1,q} = \frac{\sqrt{\rho_{1,q}^{(\mathrm{p})} \tau_p} \beta_{1,q}}{\rho_{1,q}^{(\mathrm{p})} \tau_p \beta_{1,q} + \sum_{k=1}^{K} \rho_{2,k}^{(\mathrm{d})} \beta_{2,k} + \sigma^2}.$$

As a result, the estimated channel is a random vector with distribution $\hat{\boldsymbol{h}}_{1,q} \sim \mathcal{CN}\left(0, \gamma_{1,q} \boldsymbol{I}_M\right)$, where

$$\gamma_{1,q} = \frac{\rho_{1,q}^{(\mathrm{p})} \tau_p \beta_{1,q}^2}{\rho_{1,q}^{(\mathrm{p})} \tau_p \beta_{1,q} + \sum_{k=1}^{K} \rho_{2,k}^{(\mathrm{d})} \beta_{2,k} + \sigma^2}. \quad (5)$$

Furthermore, the estimation error $\boldsymbol{e}_{1,q} = \boldsymbol{h}_{1,q} - \hat{\boldsymbol{h}}_{1,q}$ is independent of the estimated channel and distributed as $\boldsymbol{e}_{1,q} \sim \mathcal{CN}\left(0, (\beta_{1,q} - \gamma_{1,q}) \boldsymbol{I}_M\right)$.

After obtaining the channel estimation, the BS applies a linear processing vector for the detection of the UL data belonging to the same user. By employing the maximum ratio combining (MRC), the combining vector is given as:

$$\boldsymbol{v}_{1,q} = \hat{\boldsymbol{h}}_{1,q}. \quad (6)$$

For the second group, the same channel estimation process applies, but with the roles of the two groups reversed.

III. UPLINK DATA TRANSMISSION

A. ANALYSIS IN THE TRAINING PHASE

To examine data detection in the training phase, focus on the time slots over which the first group transmits UL data while the second group transmits UL pilots for channel estimation. The signal received at the BS over one symbol can be rewritten from (1) as:

$$\boldsymbol{y}^{(\mathrm{tp})} = \sum_{k=1}^{K} \boldsymbol{h}_{1,k} \sqrt{\rho_{1,k}^{(\mathrm{d})}} x_{1,k}^{(\mathrm{tp})} + \sum_{k=1}^{K} \boldsymbol{h}_{2,k} \sqrt{\rho_{2,k}^{(\mathrm{p})}} \phi_k + \boldsymbol{n}, \quad (7)$$

where ϕ_k simply denotes one entry of the pilot vector $\boldsymbol{\phi}_k$. To detect data of the qth user of the first group, the BS multiplies the above received signal with the corresponding combining vector $\boldsymbol{v}_{1,q}$ as specified in (6). This yields

$$\boldsymbol{v}_{1,q}^H \boldsymbol{y}^{(\mathrm{tp})} = \sum_{k=1}^{K} \boldsymbol{v}_{1,q}^H \boldsymbol{h}_{1,k} \sqrt{\rho_{1,k}^{(\mathrm{d})}} x_{1,k}^{(\mathrm{tp})} + \sum_{k=1}^{K} \boldsymbol{v}_{1,q}^H \boldsymbol{h}_{2,k} \sqrt{\rho_{2,k}^{(\mathrm{p})}} \phi_k + \boldsymbol{v}_{1,q}^H \boldsymbol{n}. \quad (8)$$

To see how the desired data is affected by different components, decompose the signal in (8) as:

$$\boldsymbol{v}_{1,q}^H \boldsymbol{y}^{(\mathrm{tp})} = \underbrace{\boldsymbol{v}_{1,q}^H \boldsymbol{h}_{1,q} \sqrt{\rho_{1,q}^{(\mathrm{d})}} x_{1,q}^{(\mathrm{tp})}}_{\mathrm{DS}_{1,q}^{(\mathrm{tp})}\text{ – Desired signal}} + \underbrace{\sum_{k=1,k\neq t}^{K} \boldsymbol{v}_{1,q}^H \boldsymbol{h}_{1,k} \sqrt{\rho_{1,k}^{(\mathrm{d})}} x_{1,k}^{(\mathrm{tp})}}_{\mathrm{IwG}_{1,q}^{(\mathrm{tp})}\text{ – Interference within group}}$$
$$+ \underbrace{\sum_{k=1}^{K} \boldsymbol{v}_{1,q}^H \boldsymbol{h}_{2,k} \sqrt{\rho_{2,k}^{(\mathrm{p})}} \phi_k}_{\mathrm{IP}_{1,q}^{(\mathrm{tp})}\text{ – Interference from pilot}} + \underbrace{\boldsymbol{v}_{1,q}^H \boldsymbol{n}}_{\mathrm{N}_{1,q}^{(\mathrm{tp})}\text{ – Noise}} . \quad (9)$$

The first component in (9) is the desired signal for the detection of data $x_{1,q}^{(\mathrm{tp})}$. The second term accounts for interference from users in the same group. The terms $\mathrm{IP}_{1,q}^{(\mathrm{tp})}$ quantifies the interference from pilot transmissions conducted by users in the second group. The last component in (9) is filtered additive Gaussian noise.

Next, consider the case that the MRC is used at the BS, i.e., $\boldsymbol{v}_{1,q} = \hat{\boldsymbol{h}}_{1,q} = \mu_{1,q} \boldsymbol{Y} \frac{\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|}$. Given the distribution of the channel estimate $\hat{\boldsymbol{h}}_{1,q} \sim \mathcal{CN}\left(0, \gamma_{1,q} \boldsymbol{I}_M\right)$, the following analyzes the behavior of each term in (9) when $M \to \infty$.

First, the desired signal component is

$$\begin{aligned}
\mathrm{DS}_{1,q}^{(\mathrm{tp})} &= \left(\mu_{1,q} \boldsymbol{Y} \frac{\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|}\right)^H \boldsymbol{h}_{1,q} \sqrt{\rho_{1,q}^{(\mathrm{d})}} x_{1,q}^{(\mathrm{tp})} \\
&= \Bigg[\mu_{1,q} \boldsymbol{h}_{1,q}^H \boldsymbol{h}_{1,q} \sqrt{\rho_{1,q}^{(\mathrm{p})} \rho_{1,q}^{(\mathrm{d})}} \tau_p \\
&\quad + \mu_{1,q} \sum_{k=1}^{K} \boldsymbol{h}_{2,k}^H \boldsymbol{h}_{1,q} \sqrt{\rho_{2,k}^{(\mathrm{d})} \rho_{1,q}^{(\mathrm{d})}} \boldsymbol{x}_{2,k}^{(\mathrm{tp})} \frac{\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|} \\
&\quad + \mu_{1,q} \left(\boldsymbol{N} \frac{\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|}\right)^H \boldsymbol{h}_{1,q} \sqrt{\rho_{1,q}^{(\mathrm{d})}} \Bigg] x_{1,q}^{(\mathrm{tp})}. \quad (10)
\end{aligned}$$

Due to the fact that the channels from the BS to all users are mutually independent, by applying the law of large numbers, the second and third components in (10) go to zero when M goes to infinity. It follows that

$$\frac{1}{M} \mathrm{DS}_{1,q}^{(\mathrm{tp})} \xrightarrow[M\to\infty]{\text{a.s}} \mu_{1,q} \beta_{1,q} \sqrt{\rho_{1,q}^{(\mathrm{p})} \rho_{1,q}^{(\mathrm{d})}} \tau_p x_{1,q}^{(\mathrm{tp})}, \quad (11)$$

where the notation $\xrightarrow[M\to\infty]{\text{a.s}}$ means almost sure convergence as $M \to \infty$. On the other hand, due to fact that all the components of $\boldsymbol{v}_{1,q} = \hat{\boldsymbol{h}}_{1,q}$ are statically independent of $\boldsymbol{h}_{1,k}$ for all $k \neq q$, $\mathrm{IwG}_{1,q}^{(\mathrm{tp})}$ and $\mathrm{N}_{1,q}^{(\mathrm{tp})}$ vanish when $M \to \infty$. That is,

$$\frac{1}{M} \mathrm{IwG}_{1,q}^{(\mathrm{tp})} \xrightarrow[M\to\infty]{\text{a.s}} 0, \quad (12)$$

and

$$\frac{1}{M} \mathrm{N}_{1,q}^{(\mathrm{tp})} \xrightarrow[M\to\infty]{\text{a.s}} 0. \quad (13)$$

Next, the interference term $\mathrm{IP}_{1,q}^{(\mathrm{tp})}$ that originates from the second group which transmits UL pilots is decomposed as:

$$\begin{aligned}
&\boldsymbol{v}_{1,q}^H \boldsymbol{h}_{2,k} \sqrt{\rho_{2,k}^{(\mathrm{p})}} \phi_k \\
&= \left(\mu_{1,q} \boldsymbol{Y} \frac{\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|}\right)^H \boldsymbol{h}_{2,k} \sqrt{\rho_{2,k}^{(\mathrm{p})}} \phi_k \\
&= \mu_{1,q} \boldsymbol{h}_{1,q}^H \boldsymbol{h}_{2,k} \sqrt{\rho_{1,q}^{(\mathrm{p})} \rho_{2,k}^{(\mathrm{p})} \tau_p} \phi_k \\
&\quad + \mu_{1,q} \sum_{k=1}^{K} \boldsymbol{h}_{2,k}^H \boldsymbol{h}_{2,k} \sqrt{\rho_{2,k}^{(\mathrm{d})} \rho_{2,k}^{(\mathrm{p})}} \boldsymbol{x}_{2,k}^{(\mathrm{tp})} \frac{\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|} \phi_k \\
&\quad + \mu_{1,q} \left(\boldsymbol{N} \frac{\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|}\right)^H \boldsymbol{h}_{2,k} \sqrt{\rho_{2,k}^{(\mathrm{d})}} \phi_k. \quad (14)
\end{aligned}$$

The first and third terms of (14) also vanish when $M \to \infty$. Therefore, the remaining term is:

$$\frac{1}{M} \mathrm{IP}_{1,q}^{(\mathrm{tp})} \xrightarrow[M\to\infty]{\text{a.s}} \sum_{k=1}^{K} \mu_{1,q} \beta_{2,k} \sqrt{\rho_{2,k}^{(\mathrm{d})} \rho_{2,k}^{(\mathrm{p})}} \boldsymbol{x}_{2,k}^{(\mathrm{tp})} \frac{\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|} \phi_k. \quad (15)$$

In summary, the above analysis shows that, when the number of antennas at the BS goes to infinity, the received signal for the qth user of the first group consists of the desired signal component as in (11) and the interference caused by users of the other group as a result of pilot contamination during its training phase (15).

The presence of $\mathrm{IP}_{1,q}^{(\mathrm{tp})}$ in (15) is due to the correlation between the channel estimation errors of the pilot-transmitting group and the received signals of the data-transmitting group. The impact of this interference can be reduced by applying the following interference cancellation method. At first, it can be seen from (14) that the part in $\mathrm{IP}_{1,q}^{(\mathrm{tp})}$ that remains when $M \to \infty$ is:

$$\Upsilon_{1,q}^{(\mathrm{IP})} = \mu_{1,q} \sum_{k=1}^{K} \boldsymbol{h}_{2,k}^H \boldsymbol{h}_{2,k} \sqrt{\rho_{2,k}^{(\mathrm{d})} \rho_{2,k}^{(\mathrm{p})}} \boldsymbol{x}_{2,k}^{(\mathrm{tp})} \frac{\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|} \phi_k. \quad (16)$$

Since the UL transmit power and UL pilot sequences are known and the UL signal $\boldsymbol{x}_{2,k}^{(\mathrm{tp})}$ was already detected first, the term $\Upsilon_{1,q}^{(\mathrm{IP})}$ can be estimated by replacing $\boldsymbol{h}_{2,k}^H \boldsymbol{h}_{2,k}$ with its statistical average. That is,

$$\hat{\Upsilon}_{1,q}^{(\mathrm{IP})} = \mu_{1,q} \sum_{k=1}^{K} \mathbb{E}\left\{\|\boldsymbol{h}_{2,k}\|^2\right\} \sqrt{\rho_{2,k}^{(\mathrm{d})} \rho_{2,k}^{(\mathrm{p})}} \boldsymbol{x}_{2,k}^{(\mathrm{tp})} \frac{\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|} \phi_k. \quad (17)$$

The above estimated value can then be subtracted from the received signal of the qth user in the first group (see (9)), which should reduce the interference caused by pilot contamination. The result after performing interference cancellation in (9) is:

$$\begin{aligned}
s_{1,q} &= \boldsymbol{v}_{1,q}^H \boldsymbol{y}^{(\mathrm{tp})} - \hat{\Upsilon}_{1,q}^{(\mathrm{IP})} \\
&= \boldsymbol{v}_{1,q}^H \boldsymbol{h}_{1,q} \sqrt{\rho_{1,q}^{(\mathrm{d})}} x_{1,q}^{(\mathrm{tp})} + \sum_{k=1,k\neq t}^{K} \boldsymbol{v}_{1,q}^H \boldsymbol{h}_{1,k} \sqrt{\rho_{1,k}^{(\mathrm{d})}} x_{1,k}^{(\mathrm{tp})}
\end{aligned}$$

$$+\underbrace{\sum_{k=1}^{K}\left(\boldsymbol{v}_{1,q}^{H}\boldsymbol{h}_{2,k}-\mu_{1,q}\mathbb{E}\left\{\left\|\boldsymbol{h}_{2,k}\right\|^{2}\right\}\frac{\sqrt{\rho_{2,k}^{(\mathrm{d})}}\boldsymbol{x}_{2,k}^{(\mathrm{tp})}\boldsymbol{\phi}_{q}}{\|\boldsymbol{\phi}_{q}\|}\right)}_{\mathrm{IP}_{1,q}^{(\mathrm{tp})}-\hat{\Upsilon}_{1,q}^{(\mathrm{IP})}}$$

$$\times\sqrt{\rho_{2,k}^{(\mathrm{p})}}\phi_{k}+\boldsymbol{v}_{1,q}^{H}\boldsymbol{n}. \tag{18}$$

It can be seen from (17) that:

$$\frac{1}{M}\hat{\Upsilon}_{1,q}^{(\mathrm{IP})}\xrightarrow[M\to\infty]{\text{a.s}}\sum_{k=1}^{K}\mu_{1,q}\beta_{2,k}\sqrt{\rho_{2,k}^{(\mathrm{d})}\rho_{2,k}^{(\mathrm{p})}}\boldsymbol{x}_{2,k}^{(\mathrm{tp})}\frac{\boldsymbol{\phi}_{q}}{\|\boldsymbol{\phi}_{q}\|}\phi_{k}, \tag{19}$$

which means that the term $\left(\mathrm{IP}_{1,q}^{(\mathrm{tp})}-\hat{\Upsilon}_{1,q}^{(\mathrm{IP})}\right)$ converges to zero when $M\to\infty$, hence pilot contamination can be removed. Thus, as M goes to infinity, only the desired signal component $\mathrm{DS}_{1,q}^{(\mathrm{tp})}$ remains in (18).

Next, Theorem 1 gives a closed-form expression for a lower bound of the UL spectral efficiency for the qth user of the first group when MRC is used at the BS. Note that this result is valid for finite M.

Theorem 1: The UL spectral efficiency of the qth user in the first group in the training phase with MRC at the BS and successive interference cancellation is given as:

$$R_{1,q}^{(\mathrm{tp})}\geq\log_{2}(1+\mathrm{SINR}_{1,q}^{(\mathrm{tp,\,MRC})}), \tag{20}$$

where $\mathrm{SINR}_{1,q}^{(\mathrm{tp,\,MRC})}$ is given as in (21), as shown at the bottom of this page.

Proof: See the Appendix. It should be pointed out that, the same result applies to users in the second group during its training phase.

From (21), one can see that the array gain is proportional to the number of antennas, while the power of interference in the denominator is independent of the number of antennas. In particular, the denominator consists of two components: (i) uncorrelated interference, whose power equals to the signal power of all users received at the BS and noise power, and (ii) correlated interference caused by users in the second group as a result of pilot contamination.

B. ANALYSIS IN THE DATA PHASE

In the data phase, both groups transmit their UL data. The received signal in the data phase is given as in (1) by substituting $x_{g,k}=x_{g,k}^{(\mathrm{dp})}$ (for $g=1,2$). Similar to the training phase, after applying a combining vector $\boldsymbol{v}_{1,q}$, the received signal of the qth user from the first group is decomposed as:

$$\begin{aligned}&\boldsymbol{v}_{1,q}^{H}\boldsymbol{y}^{(\mathrm{dp})}\\&=\sum_{k=1}^{K}\left(\boldsymbol{v}_{1,q}^{H}\boldsymbol{h}_{1,k}\sqrt{p_{1,k}}x_{1,k}^{(\mathrm{dp})}+\boldsymbol{v}_{1,q}^{H}\boldsymbol{h}_{2,k}\sqrt{p_{2,k}}x_{2,k}^{(\mathrm{dp})}\right)+\boldsymbol{v}_{1,q}^{H}\boldsymbol{n}\\&=\underbrace{\boldsymbol{v}_{1,q}^{H}\boldsymbol{h}_{1,q}\sqrt{p_{1,q}}x_{1,q}^{(\mathrm{dp})}}_{\mathrm{DS}_{1,q}^{(\mathrm{dp})}\text{ – Desired signal}}+\underbrace{\sum_{k=1,k\neq q}^{K}\boldsymbol{v}_{1,q}^{H}\boldsymbol{h}_{1,k}\sqrt{p_{1,k}}x_{1,k}^{(\mathrm{dp})}}_{\mathrm{IwG}_{1,q}^{(\mathrm{dp})}\text{ – Interference within group}}\\&+\underbrace{\sum_{k=1}^{K}\boldsymbol{v}_{1,q}^{H}\boldsymbol{h}_{2,k}\sqrt{p_{2,k}}x_{2,k}^{(\mathrm{dp})}}_{\mathrm{IoG}_{1,q}^{(\mathrm{dp})}\text{ – Interference from other group}}+\underbrace{\boldsymbol{v}_{1,q}^{H}\boldsymbol{n}}_{\mathrm{N}_{1,q}^{(\mathrm{dp})}\text{ – Noise}}.\end{aligned} \tag{22}$$

Unlike the training phase, there is no interference caused by pilot transmission of the other group. Instead, there is interference, denoted as $\mathrm{IoG}_{1,q}^{(\mathrm{dp})}$, caused by concurrent data transmission from the other group. Following the same analysis as in the training phase, the terms $\mathrm{IwG}_{1,q}^{(\mathrm{dp})}$ and $\mathrm{N}_{1,q}^{(\mathrm{dp})}$ vanish when M goes to infinity. The only terms remained in (22) are the desired signal component,

$$\frac{1}{M}\mathrm{DS}_{1,q}^{(\mathrm{dp})}\xrightarrow[M\to\infty]{\text{a.s}}\mu_{1,q}\beta_{1,q}\sqrt{\rho_{1,q}^{(\mathrm{p})}p_{1,q}}\tau_{p}x_{1,q}^{(\mathrm{dp})}, \tag{23}$$

and interference from users in the other group:

$$\begin{aligned}&\frac{1}{M}\mathrm{IoG}_{1,q}^{(\mathrm{dp})}\\&\xrightarrow[M\to\infty]{a.s}\sum_{k=1}^{K}\mu_{1,q}\beta_{2,k}\sqrt{\rho_{2,k}^{(\mathrm{d})}p_{2,k}}\boldsymbol{x}_{2,k}^{(\mathrm{tp})}\frac{\boldsymbol{\phi}_{q}}{\|\boldsymbol{\phi}_{q}\|}x_{2,k}^{(\mathrm{dp})}.\end{aligned} \tag{24}$$

Different from the training phase, where the correlated interference from the pilot-transmitting group can be subtracted from the received signal of the data-transmitting group, the data signals $x_{1,k}^{(\mathrm{dp})}$ and $x_{2,k}^{(\mathrm{dp})}$ in (24) are unknown, and therefore the interference cancellation method cannot be applied in the same way as in the training phase. However, assuming that the signal from the second group, $x_{2,k}^{(\mathrm{dp})}$, is detected first by treating $x_{1,k}^{(\mathrm{dp})}$ as noise, then one can subtract an estimated value of $\mathrm{IoG}_{1,q}^{(\mathrm{dp})}$ from the received signal of the first group. This successive cancellation has been investigated in [35] and [36] and shown to significantly improve the minimal SINR value of users.

With the knowledge of $x_{2,k}^{(\mathrm{dp})}$, an estimation of $\mathrm{IoG}_{1,q}^{(\mathrm{dp})}$ can be formed as:

$$\begin{aligned}&\hat{\Upsilon}_{1,q}^{((\mathrm{IoG})}\\&=\mu_{1,q}\sum_{k=1}^{K}\mathbb{E}\left\{\left\|\boldsymbol{h}_{2,k}\right\|^{2}\right\}\sqrt{\rho_{2,k}^{(\mathrm{d})}p_{2,k}}\boldsymbol{x}_{2,k}^{(\mathrm{tp})}\frac{\boldsymbol{\phi}_{t}}{\|\boldsymbol{\phi}_{t}\|}x_{2,k}^{(\mathrm{dp})},\end{aligned} \tag{25}$$

$$\mathrm{SINR}_{1,q}^{(\mathrm{tp,\,MRC})}=\frac{M\rho_{1,q}^{(\mathrm{d})}\gamma_{1,q}}{\sum_{k=1}^{K}\left(\rho_{1,k}^{(\mathrm{d})}\beta_{1,k}+\rho_{2,k}^{(\mathrm{p})}\beta_{2,k}\right)+\sum_{k=1}^{K}\rho_{2,k}^{(\mathrm{p})}\frac{\rho_{2,k}^{(\mathrm{d})}}{\rho_{1,q}^{(\mathrm{p})}\tau_{p}}\gamma_{1,q}\left(\frac{\beta_{2,k}}{\beta_{1,q}}\right)^{2}+\sigma^{2}}. \tag{21}$$

$$\mathrm{SINR}_{1,q}^{(\mathrm{dp,MRC})} = \frac{M p_{1,q}\gamma_{1,q}}{\sum_{k=1}^{K}\left(p_{1,k}\beta_{1,k} + p_{2,k}\beta_{2,k}\right) + \sum_{k=1}^{K} p_{2,k}\frac{\rho_{2,k}^{(\mathrm{d})}}{\rho_{1,q}^{(\mathrm{p})}\tau_p}\gamma_{1,q}\left(\frac{\beta_{2,k}}{\beta_{1,q}}\right)^2 + \sigma^2} \tag{28}$$

$$\mathrm{SINR}_{2,q}^{(\mathrm{dp,MRC})} = \frac{M p_{2,q}\gamma_{2,q}}{\sum_{k=1}^{K}\left(p_{1,k}\beta_{1,k} + p_{2,k}\beta_{2,k}\right) + \sum_{k=1}^{K} p_{1,k}\frac{\rho_{1,k}^{(\mathrm{d})}}{\rho_{2,q}^{(\mathrm{p})}\tau_p}\gamma_{2,q}\left(\frac{\beta_{1,k}}{\beta_{2,q}}\right)^2 (M+1) + \sigma^2} \tag{29}$$

By subtracting (25) from (22), the received signal corresponding to the qth user in the first group now becomes:

$$\begin{aligned}
&\boldsymbol{v}_{1,q}^H \boldsymbol{y}^{(\mathrm{dp})} - \hat{\Upsilon}_{1,q}^{(\mathrm{IoG})} \\
&= \boldsymbol{v}_{1,q}^H \boldsymbol{h}_{1,q}\sqrt{\rho_{1,q}^{(\mathrm{d})}}x_{1,q}^{(\mathrm{tp})} + \sum_{k=1,k\neq t}^{K} \boldsymbol{v}_{1,q}^H \boldsymbol{h}_{1,k}\sqrt{\rho_{1,k}^{(\mathrm{d})}}x_{1,k}^{(\mathrm{tp})} \\
&\quad + \sum_{k=1}^{K}\left(\boldsymbol{v}_{1,q}^H \boldsymbol{h}_{2,k} - \mu_{1,q}\mathbb{E}\left\{\left\|\boldsymbol{h}_{2,k}\right\|^2\right\}\frac{\sqrt{\rho_{2,k}^{(\mathrm{d})}}\boldsymbol{x}_{2,k}^{(\mathrm{tp})}\boldsymbol{\phi}_q}{\|\boldsymbol{\phi}_q\|}\right) \\
&\quad \times \sqrt{p_{2,k}}x_{2,k}^{(\mathrm{dp})} + \boldsymbol{v}_{1,q}^H \boldsymbol{n}.
\end{aligned} \tag{26}$$

Based on (26), a lower bound on the UL SE of the qth user of the first group when the MRC is employed at the BS during the data phase can be obtained as in Theorem 2.

Theorem 2: The UL spectral efficiency of the qth user in the gth group ($g = 1, 2$) when the MRC is employed at the BS in the data phase is given as:

$$R_{g,q}^{(\mathrm{dp})} \geq \log_2(1 + \mathrm{SINR}_{g,q}^{(\mathrm{dp,MRC})}), \tag{27}$$

where the SINRs of the qth users of the first and second groups can be calculated as in (28) and (29), as shown at the top of this page, respectively.

Proof: The proof follows the same steps as carried out for proving the UL spectral efficiency in the training phase.

From (29), it can be seen that the correlated interference originating from pilot contamination in the denominator is proportional to $(M + 1)$. As a consequence, for the second group, this component does not vanish when the number of antennas goes to infinity, unless the data power in the training phase is set to zero (equivalently, no data is transmitted in the training phase). Based on this observation, an adaptive power control method is proposed in the next section to optimize the UL data rate.

Before closing this section, it is worth pointing out that how to assign users into two different groups (i.e., group assignment method) can affect the spectral efficiency. In general, it is desired to optimally assign users into two groups such that the highest SE can be obtained. With time-offset pilots, group assignment impacts performance in both training phase and data phase, and not in the same way.

Unfortunately, optimizing group assignment is a combinatorial problem and, therefore, difficult to find the optimal solution. As discussed at the end of Section I-A, a beam-domain group assignment approach was proposed in [30]–[32] which assigns users to different groups based on DOA. However, this approach is not applicable for the system model considered in this paper in which the channel vector's elements are mutually independent and hence cannot exploit the advantage of the beam-domain channel presentation. In this paper, in order to remove as much correlated interference as possible, we instead consider a group assignment exploiting the large-scale fading conditions of users. In this method, users in the cell are divided into inner and outer regions based on their locations. Since users in the inner region have generally better channel conditions compared to users in the outer region, they are assigned to the second group, whose data is detected first as described in Section III-B. The other users belong to the first group.

Before closing this section, it should be pointed out that the proposed time-offset pilot approach can be extended to more than 2 groups. In such a design, users are also grouped based on large-scale fading by dividing the coverage area into ring regions with different radii. The users in the group experiencing better channels will have their signals detected first, followed by users in the group having the next best channels, and so on. As a result, an arbitrary group can remove the known UL signals of all groups which have been already detected before by using successive interference cancellation (SIC). With more than 2 groups, the training phase needs to be divided into more sub-intervals, each one for one group to transmit its UL pilots. This implies that a more complicated power control method is required. Given the more severe pilot contamination and the higher complexity of interference cancellation if the system is designed with more than 2 groups, considering only 2 groups in the proposed approach appears most attractive and practical.

IV. POWER CONTROL

This section studies power control problems under two cost functions: max-min QoS and total power minimization. The approach to solve these two problems is to decompose the original problem into two subproblems corresponding to two phases (training and data) in one coherence interval.

A. MAX-MIN QOS OPTIMIZATION

Consider the optimization problem in which the cost function is to maximize the QoS value that can be equally provided to all users in the system. In the considered system model with time-offset pilots, data transmission of each user happens in both training and data phases. The training phase lasts for τ_p time slots and has a SE of $R_{g,q}^{(\mathrm{tp})}$ ($g = 1, 2$). The data phase

has a SE of $R_{g,q}^{(\mathrm{dp})}$ and is over $(1-2\tau_p)$ time slots. As a result, the average UL SE in one coherence interval of τ_c time slots can be calculated as:

$$R_{g,q}^{(\mathrm{total})} = \frac{\tau_p}{\tau_c} R_{g,q}^{(\mathrm{tp})} + \left(1 - \frac{2\tau_p}{\tau_c}\right) R_{g,q}^{(\mathrm{dp})}. \tag{30}$$

With the above UL spectral efficiency, the max-min QoS optimization problem is formulated as:

$$\begin{aligned} \underset{\rho_{g,q}^{(\mathrm{p})}, \rho_{g,q}^{(\mathrm{d})}, p_{g,q}}{\text{maximize}} \; & \min_{q=1,\ldots,K} \left\{ R_{1,q}^{(\mathrm{total})}, R_{2,q}^{(\mathrm{total})} \right\} \\ \text{subject to } & 0 \le \rho_{g,q}^{(\mathrm{p})} \le p_{\max}, \quad \forall g, q, \\ & 0 \le \rho_{g,q}^{(\mathrm{d})} \le p_{\max}, \quad \forall g, q, \\ & 0 \le p_{g,q} \le p_{\max}, \quad \forall g, q, \end{aligned} \tag{31}$$

where the objective is to maximize the minimum QoS and the constraints are to limit the data and pilot powers under a predetermined maximum UL transmit power $p_{\max}$.

The above optimization problem has the same form as the max-sum SE optimization problem studied in [37], which is a signomial programming and proved to be NP-hard [38]. Therefore, an algorithm based on the bisection method is proposed here, which iteratively solves the max-min QoS problem in training phase and data phase, separately.

In Training Phase: The power allocation problem to maximize min-QoS with UL transmit power constraints in the training phase can be formulated as:

$$\begin{aligned} \underset{\rho_{g,q}^{(\mathrm{p})}, \rho_{g,q}^{(\mathrm{d})}}{\text{maximize}} \; & \min_{q=1,\ldots,K} \left\{ R_{1,q}, R_{2,q} \right\} \\ \text{subject to } & 0 \le \rho_{g,q}^{(\mathrm{p})} \le p_{\max}, \quad \forall g, q, \\ & 0 \le \rho_{g,q}^{(\mathrm{d})} \le p_{\max}, \quad \forall g, q. \end{aligned} \tag{32}$$

The epigraph form of the above problem is:

$$\begin{aligned} \underset{\rho_{g,q}^{(\mathrm{p})}, \rho_{g,q}^{(\mathrm{d})}, \lambda^{(\mathrm{tp})}}{\text{maximize}} \; & \lambda^{(\mathrm{tp})} \\ \text{subject to } & \mathrm{SINR}_{g,q}^{(\mathrm{tp})} \ge \lambda^{(\mathrm{tp})}, \quad \forall g, q, \\ & 0 \le \rho_{g,q}^{(\mathrm{p})} \le p_{\max}, \quad \forall g, q, \\ & 0 \le \rho_{g,q}^{(\mathrm{d})} \le p_{\max}, \quad \forall g, q. \end{aligned} \tag{33}$$

By dividing both the nominator and denominator of $\mathrm{SINR}_{g,q}^{(\mathrm{tp})}$ to $\gamma_{g,q}$, the first constraint of this problem can be converted into a valid constraint of GP where the left-hand side of the "greater-than" inequality is a monomial and the right-hand side is a posynomial. As a result, this GP can be solved in polynomial time by using GP solvers like MOSEK solver with CVX [38], [39].

In Data Phase: With the power allocation strategy obtained in the training phase, the value of $\gamma_{g,k}$ can be calculated as in (5). Similar to the training phase, the max-min QoS power control in the data phase can be formulated in an epigraph form as:

$$\begin{aligned} \underset{p_{g,q}, \lambda^{(\mathrm{dp})}}{\text{maximize}} \; & \lambda^{(\mathrm{dp})} \\ \text{subject to } & \mathrm{SINR}_{g,q}^{(\mathrm{dp})} \ge \lambda^{(\mathrm{dp})}, \quad \forall g, q, \\ & 0 \le p_{g,q} \le p_{\max}, \quad \forall n, q. \end{aligned} \tag{34}$$

The objective is to maximize $\lambda^{(\mathrm{dp})}$, which is the lower bound of all $\mathrm{SINR}_{g,q}^{(\mathrm{dp})}$ as expressed in the first constraint, whereas the transmit power is limited as in the second constraint. Similar to the training phase, the max-min QoS optimization problem in the data phase is also a GP and hence can be solved in polynomial time.

Max-Min QoS Power Allocation Using the Bisection Method: With the optimization problems formulated above for training and data phases, a joint adaptive max-min QoS power allocation using the bisection method can be performed as follows. In the first stage, the max-min QoS problem in the training phase (33) is solved to obtain the maximum value of the achievable QoS (say $R_{\mathrm{ini}}^{(\mathrm{tp})}$) and the corresponding SE $R_{\mathrm{ini}}^{(\mathrm{dp})}$. Intuitively, a higher rate in the training phase causes a lower rate in the data phase because of lower-quality channel estimation. Hence, to find the value of $R_{g,q}^{(\mathrm{tp})}$ that maximizes the total rate $R_{g,q}^{(\mathrm{total})}$, its lower and upper bounds $R_{\min} \le R_{g,q}^{(\mathrm{tp})} \le R_{\max}$ are chosen such that $R_{\max} = R_{\mathrm{ini}}^{(\mathrm{tp})}$ is the optimal solution for (33) and $R_{\min} = 0$. Applying bisection searching within this interval, in each iteration, the following problem is solved

$$\begin{aligned} \underset{\rho_{g,q}^{(\mathrm{p})}, \rho_{g,q}^{(\mathrm{d})}}{\text{minimize}} \; & \theta \\ \text{subject to } & \frac{\rho_{g,k}^{(\mathrm{p})}}{\rho_{g',q}^{(\mathrm{d})}} \le \theta, \quad \forall k, q, g \ne g', \\ & \mathrm{SINR}_{g,q}^{(\mathrm{tp})} \ge \lambda_{\mathrm{req}}^{(\mathrm{tp})}, \quad \forall g, q, \\ & 0 \le \rho_{g,q}^{(\mathrm{p})} \le p_{\max}, \quad \forall g, q, \\ & 0 \le \rho_{g,q}^{(\mathrm{d})} \le p_{\max}, \quad \forall g, q, \end{aligned} \tag{35}$$

where $\lambda_{\mathrm{req}}^{(\mathrm{tp})}$ is the value of the SINR corresponding to

$$R_{\mathrm{req}}^{(\mathrm{tp})} = \log_2(1 + \lambda_{\mathrm{req}}^{(\mathrm{tp})}). \tag{36}$$

The cost function and the first constraint in (35) aim to minimize the interference caused by pilot-transmitting group as in (21), while maintaining a required QoS as expressed in the second constraint. After obtaining the power allocation with respect to (35), the total achievable UL rate can be calculated by (30). This procedure is iterated until $R_{g,q}^{(\mathrm{total})}$ converges. The proposed power allocation method is summarized in Algorithm 1.

With the proposed power control algorithm, the max-min QoS with time-offset pilots is not upper bounded by a saturation level as in the case of using non-orthogonal pilots. The reason is that the SINR in the training phase grows proportionally with the number of antennas. When the coherence interval is short, this even leads to a larger amount of SE compared to the orthogonal pilot method. This is because the orthogonal pilot method has to spend more time slots for pilot signaling and there will be fewer time slots left for data transmission. On the other hand, when the coherence interval is large, the proposed algorithm can adaptively reduce data

Algorithm 1 Bisection-Based Algorithm for Max-Min QoS Power Control

Require: The maximum achievable rate in training phase $R_{\text{ini}}^{(\text{tp})}$ and its corresponding SE $R_{\text{ini}}^{(\text{dp})}$
$\delta = \infty$;
$R_{\max} = R_{\text{ini}}^{(\text{tp})}$
$R_{\min} = 0$
$R_{\text{prev}}^{(\text{total})} = R_{\text{ini}}^{(\text{total})}$ where $R_{\text{ini}}^{(\text{total})}$ is calculated as in (30)
while $\delta > \delta_{\text{threshold}}$ **do**
  $R_{\text{req}}^{(\text{tp})} = (R_{\min} + R_{\max})/2$
  Solve (35) with respect to $\lambda_{\text{req}} = R_{\text{req}}^{(\text{tp})}$
  Recalculate the corresponding $R_{\text{req}}^{(\text{tp})}$ and obtain the new $R_{\text{new}}^{(\text{total})}$ by applying (30).
  if $R_{\text{new}}^{(\text{total})} \leq R_{\text{prev}}^{(\text{total})}$ **then**
    $R_{\max} = R_{\text{req}}^{(\text{tp})}$
  else
    $R_{\min} = R_{\text{req}}^{(\text{tp})}$
    $R_{\text{prev}}^{(\text{total})} = R_{\text{new}}^{(\text{total})}$
  end if
  $\delta = R_{\max} - R_{\min}$;
end while
return $R_{\text{new}}^{(\text{total})}$

power in the training phase to ease the effect of pilot contamination to the data phase. It should also be pointed out that when the data power in the training phase is set to 0, there is no pilot contamination, and the system with time-offset pilots is equivalent to the system using orthogonal pilots with the pilot length of $2\tau_p$.

B. MINIMIZATION OF TOTAL POWER

This section studies the power control problem in which the objective is to minimize the total transmit power of the system while guaranteeing a predetermined QoS to be equally provided to all users. Because one coherence interval is separated into two phases, the average power is:

$$P_{g,k}^{(\text{total})} = \frac{\tau_p(\rho_{g,k}^{(\text{p})} + \rho_{g,k}^{(\text{d})}) + (\tau_c - 2\tau_p)p_{g,k}}{\tau_c}. \tag{37}$$

For a required QoS value of ξ that is equally provided to all users, the optimization problem is:

$$\begin{aligned} \underset{\rho_{g,k}^{(\text{p})}, \rho_{g,k}^{(\text{d})}, p_{g,k}}{\text{minimize}} \quad & \sum_{g=1}^{2}\sum_{k=1}^{K} P_{g,k}^{(\text{total})} \\ \text{subject to } & R_{g,k}^{(\text{total})} \geq \xi, \quad \forall g, k, \\ & 0 \leq \rho_{g,k}^{(\text{p})} \leq p_{\max}, \quad \forall g, k, \\ & 0 \leq \rho_{g,k}^{(\text{d})} \leq p_{\max}, \quad \forall g, k, \end{aligned} \tag{38}$$

where the first constraint is to ensure that the required QoS value of ξ is equally served to all users, whereas the next two constraints limit the transmit power by a maximum value of $p_{\max}$. The left-hand-side of the SE constraint is in the form of a fraction whose denominator and nominator are posynomials, while the right-hand-side is a constant. This means that the above optimization problem is a signomial programming, which is NP-hard [38]. Hence, like in the previous section, the original problem in (38) are separated into two subproblems for the training phase and data phase. By iteratively solving this two subproblems until convergence, a suboptimal solution for (38) is obtained. The two subproblems are formulated and discussed next.

Power Control in Training Phase: The power minimization problem in the training phase can be written as:

$$\begin{aligned} \underset{\rho_{g,k}^{(\text{p})}, \rho_{g,k}^{(\text{d})}}{\text{minimize}} \quad & \sum_{g=1}^{2}\sum_{k=1}^{K}(\rho_{g,k}^{(\text{p})} + \rho_{g,k}^{(\text{d})}) \\ \text{subject to } & \text{SINR}_{g,k}^{(\text{tp})} \geq \lambda_{\text{req}}^{(\text{tp})}, \quad \forall g, k, \\ & 0 \leq \rho_{g,k}^{(\text{p})} \leq p_{\max}, \quad \forall g, k, \\ & 0 \leq \rho_{g,k}^{(\text{d})} \leq p_{\max}, \quad \forall g, k, \end{aligned} \tag{39}$$

where $\lambda_{\text{req}}^{(\text{tp})}$ is the required SINR, which is equivalent to a predetermined value of QoS as defined in (36). This optimization problem is a GP, and hence can be solved in polynomial time.

Power Control in Data Phase: In the data phase, the power minimization problem is:

$$\begin{aligned} \underset{p_{g,k}}{\text{minimize}} \quad & \sum_{g=1}^{G}\sum_{k=1}^{K} p_{g,k} \\ \text{subject to } & \text{SINR}_{g,k}^{(\text{dp})} \geq \lambda_{\text{req}}^{(\text{dp})}, \quad \forall g, k, \\ & 0 \leq p_{g,k} \leq p_{\max}, \quad \forall g, k, \end{aligned} \tag{40}$$

where $\lambda_{\text{req}}^{(\text{dp})}$ is the required SINR, which is equivalent to a predetermined value of QoS $R_{\text{req}}^{(\text{dp})}$:

$$R_{\text{req}}^{(\text{dp})} = \log_2(1 + \lambda_{\text{req}}^{(\text{dp})}). \tag{41}$$

The above power minimization is convex, hence it is easily solved by existing convex optimization packages such as CVX.

Joint Power Minimization: To minimize the total transmit power during a coherence interval which includes both the training and data phases as in (38) is a problem with high complexity. Hence, an iterative method based on the bisection algorithm is performed as follows.

For a QoS requirement of ξ, in the first stage, we solve the max-min QoS problem in the training phase to obtain the maximum value of the achievable QoS in this phase, say $R_{\max}^{(\text{tp})}$. In the next step, we find the optimal UL rate contributed by the data phase, $R_{\text{req}}^{(\text{tp})}$, which minimizes the total UL transmit power. This can be done by bounding $R_{\min} \leq R_{\text{req}}^{(\text{tp})} \leq R_{\max}$ where the upper-bound and lower-bound are initially chosen as $R_{\min} = 0$ and $R_{\max} = R_{\max}^{(\text{tp})}$ and then updated in each iteration until the two bounds converge. With the allocated UL rate in the training phase, $R_{\text{req}}^{(\text{tp})}$, the required QoS in the data phase is:

$$R_{\text{req}}^{(\text{dp})} = \frac{\tau_c\xi - \tau_p R_{\text{req}}^{(\text{tp})}}{\tau_c - 2\tau_p}. \tag{42}$$

By using the power profile obtained in the training phase to estimate the channel coefficients, we can solve (40) with respect to the required data rate as in (42) and acquire the optimal transmit power in the data phase and the total UL transmit power. In the next iteration, the required data rate in the training phase is reduced to lower the effect of pilot contamination, which enhances the data rate in the data phase. The new total UL transmit power is then calculated by solving (40) with respect to the new required QoS. If the total UL transmit power in the new iteration is higher than the previous one, it means that the allocated QoS in the training phase has been reduced to much, which causes excessive power in the data phase. In this case, $R_{\min}$ needs to be updated to raise the allocated QoS in the training phase and ease the burden in the data phase. Otherwise, if the total power in the new iteration is lower than the previous one, we can continuously reduce the allocated power in the training phase by updating $R_{\max}$. The iteration process stops when two bounds converges (when $\delta = R_{\max} - R_{\min}$ is lower than a threshold value $\delta_{\text{threshold}}$). The proposed procedure is summarized in Algorithm 2.

Algorithm 2 Bisection Algorithm for Power Minimization

Require: The maximum achievable rate in training phase $R_{\max}^{(\text{tp})}$ and the required QoS ξ
$\delta = \infty$;
$P_{\text{prev}}^{(\text{total})} = \infty$;
$R_{\max} = R_{\max}^{(\text{tp})}$;
$R_{\min} = 0$;
$R_{\text{req}}^{(\text{tp})} = R_{\max}$;
while $\delta > \delta_{\text{threshold}}$ **do**
  Solve (39) with respect to $\lambda_{\text{req}}^{(\text{tp})}$ calculated in (36).
  Calculate the required $R_{\text{req}}^{(\text{dp})}$ as in (42) and solve (40).
  $P_{\text{new}}^{(\text{total})} = \sum_{g=1}^{G} \sum_{k=1}^{K} P_{g,k}^{(\text{total})}$;
  if $P_{\text{new}}^{(\text{total})} \leq P_{\text{prev}}^{(\text{total})}$ **then**
    $R_{\max} = R_{\text{req}}^{(\text{tp})}$;
    $P_{\text{prev}}^{(\text{total})} = P_{\text{new}}^{(\text{total})}$;
  else
    $R_{\min} = R_{\text{req}}^{(\text{tp})}$;
  end if
  $R_{\text{req}}^{(\text{tp})} = (R_{\min} + R_{\max})/2$;
  $\delta = R_{\max} - R_{\min}$;
end while
return $P_{\text{new}}^{(\text{total})}$

V. SIMULATION RESULTS

In this section, numerical results are given to evaluate the performance of the multiuser massive MIMO system with time-offset pilots in terms of achievable QoS and power consumption. The results are also compared to results obtained with orthogonal pilots and non-orthogonal pilots. The performance is observed by changing the number of antennas, coherence interval and the required QoS. The massive MIMO system considered in simulation consists of one multi-antenna BS and 30 randomly-distributed users. In each iteration, the locations of 30 users are randomly generated within the 200 meters radius around the BS. Numerical results are averaged over 200 iterations. The large-scale fading coefficients are modeled according to the 3GPP LTE standard [40]. Specifically, the large scale fading is computed as $\beta_{g,k} = -131 - 42.8\log_{10} d_{g,k} + z_{l,k}$ (dB), where $d_{g,k}$ denotes the distance from the BS to the kth user of the gth group and $z_{g,k}$ is the standard deviation of the shadowing variable. The noise figure of 5dB translates to a noise variance of 96dBm. The simulation parameters are summarized in Table 1. In all simulation scenarios, the number of pilots for time-offset and non-orthogonal pilot methods is τ_p, whereas, in order to serve the same number of users, the orthogonal pilot method needs twice the number of pilots, i.e., $2\tau_p$.

TABLE 1. Simulation parameters.

Parameter	Value
Peak UL radio transmit power	23 dBm
Number of users	30
Shadowing standard deviation	10 dB
Penetration loss (indoor users)	20 dB
Noise figure	5 dB
Pathloss	$131 + 42.8\log_{10} d$

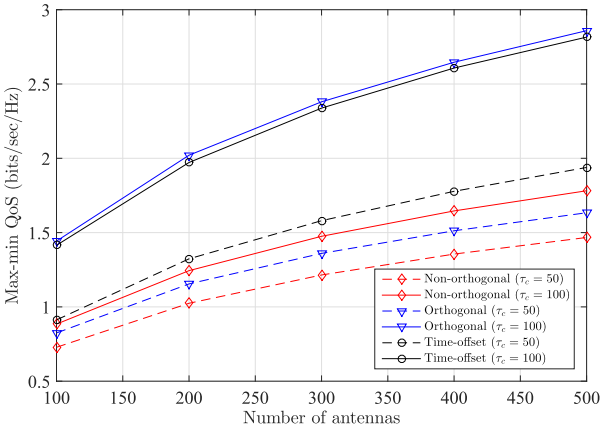


FIGURE 2. Max-min QoS versus the number of antennas ($N = 30$ users).

Fig. 2 plots the maximum QoS that all users can be equally served by the BS. It can be seen that using time-offset pilots yields a far better performance compared to using non-orthogonal pilots. Moreover the performance gap between this two methods increases with the number of antennas, from about 0.5 bits/sec/Hz at $M = 100$ to almost 1 bit/sec/Hz at $M = 500$ for the case of $\tau_c = 100$ symbols. The reason is that, when M increases, the denominator in the SINR expression increases proportionally with M for non-orthogonal pilots [19], [24], while it is not the case with time-offset pilots, thanks to the power control algorithm represented in Section IV-B. Another remarkable observation is that the performance curves with time-offset pilots are just slightly below that with orthogonal pilots for the case

TABLE 2. Rate contribution from the training phase.

Coherence length (τ_c)	30	35	40	45	50	55	60	65	70	75
Percentage (%)	100	79.66	63.74	40.62	32.71	18.24	6.60	1.29	1.13	1.00

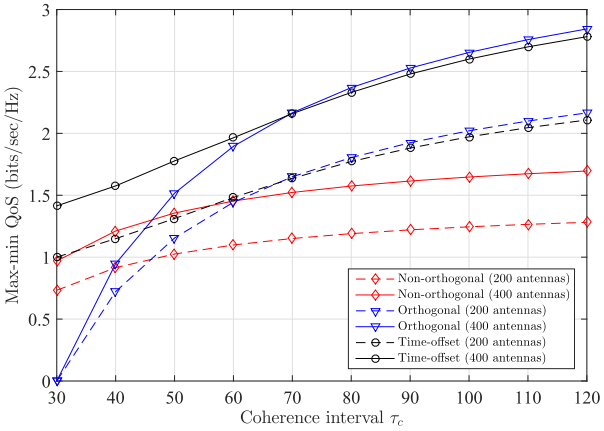


FIGURE 3. Max-min QoS versus coherence interval ($N = 30$ users).

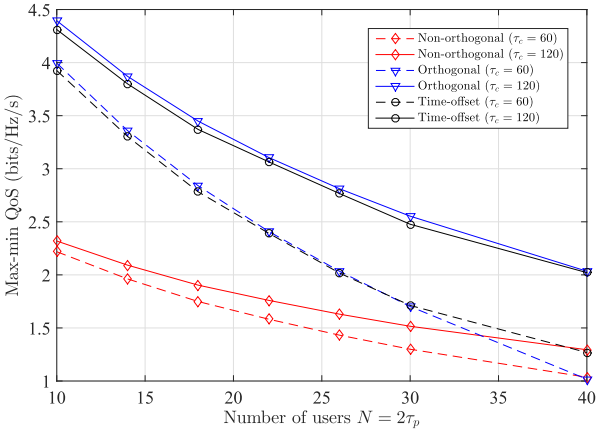


FIGURE 4. Max-min QoS versus the number of users ($M = 300$ antennas).

$\tau_c = 100$, while the performance curves with time-offset pilots are better when $\tau_c = 50$. This is because when the coherence interval is short, the orthogonal pilot method has to spend a larger portion of the coherence interval for channel estimation, while the time-offset pilot method has a much longer duration for data transmission.

The achievable rates that the BS can equally serve all users for different coherence intervals are illustrated in Fig. 3. Obviously, when $\tau_c \to K \times G$, there are no time slots available for data transmission in the orthogonal pilot case and the data rate goes to zero. On the other hand, non-orthogonal and time-offset pilot methods can still provide SEs of up to 1 and 1.5 bits/sec/Hz, respectively (when $M = 400$). When τ_c increases, the SE achieved with the non-orthogonal pilot method tends to asymptotically approach 1.7 bits/sec/Hz for 400 antennas and 1.3 bits/sec/Hz for 200 antennas due to pilot contamination. In contrast, the SEs achieved with time-offset and orthogonal pilot methods sharply increase with τ_c and reach up to 2.6 bit/sec/Hz when $\tau_c = 120$. It can also be seen that when the coherence interval is shorter than about 70 symbols, using time-offset pilots yields a better performance than using orthogonal pilots. The intersection value increases when the number of antennas goes up (at $\tau_c = 66$ for 200 antennas and $\tau_c = 70$ for 400 antennas).

The max-min QoS values versus the number of users for different coherence lengths are shown in Fig 4. The number of pilots is set as half of the number of users for time-offset and non-orthogonal pilots. The max-min QoS value decreases when the number of users increases because there are more interference sources. However, the time-offset pilot method still outperforms the non-orthogonal pilot method. This is because interference cancellation can be applied for data detection and better UL channel estimation can be obtained with time-offset pilots compared to non-orthogonal pilots. Remarkably, when $\tau_c = 60$ the time-offset pilot method eventually shows a better performance compared to orthogonal pilots when $N \geq 30$ users. Again, the reason is that with orthogonal pilots, the system has to spend a larger portion of time slots of pilots, which leaves a smaller number of time slots for data transmission. Furthermore, the contribution from the training phase to the total UL SE is presented in Table 2 for the case $N = 30$. When the coherence interval $\tau_c = 30$, it is obvious that the training phase contributes 100% of the total UL SE. The contribution in total uplink SE of the training phase decreases when τ_c increases. When $\tau_c = 65$ symbols, the SE contribution from the training phase approximately approaches zero, which means that no data is transmitted in the training phase. In such a case, the system is equivalent to the one that uses orthogonal pilots with the length of $2\tau_p$.

Fig. 5 plots the cumulative distribution function (CDF) of the max-min QoS of all three pilot methods, where the number of antennas is 500 for the non-orthogonal pilot method, and 300 for the other two methods. For the time-offset pilot method, proposed and random group assignments are considered. As can be seen, by performing group assignment according to large-scale fading information as described at the end of Section III, the time-offset pilot method can provide a max-min QoS of more than 4.2 bits/Hz/s, which is far better than when group assignment is performed randomly. In addition, the figure also shows that the time-offset pilot method outperforms the non-orthogonal pilot method when employing the same number of pilots.

Fig. 6 illustrates the optimal values of per-user average transmit power at the required QoS of 1.5 bits/sec/Hz with $\tau_c = 80$ and 120 symbols. As can be seen, the

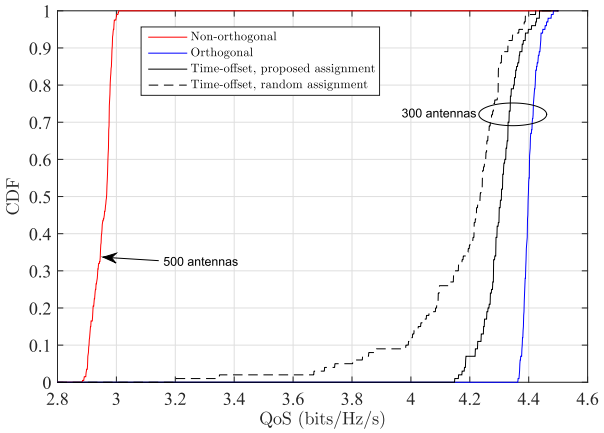


FIGURE 5. Cumulative distribution function of the max-min QoS ($N = 10$ users, $\tau_c = 120$).

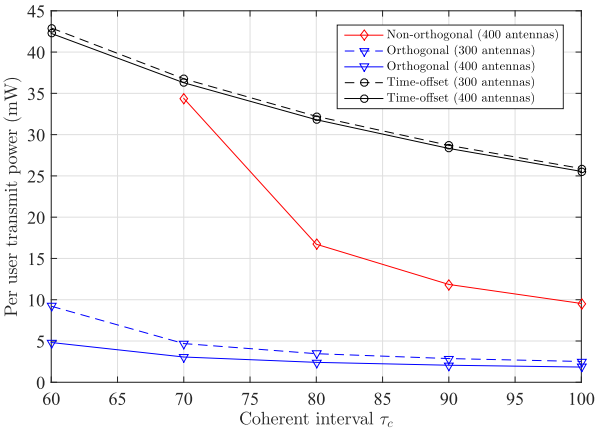


FIGURE 7. Per-user average transmit power versus coherence interval (QoS=1.5 bit/Hz/s, $N = 30$ users).

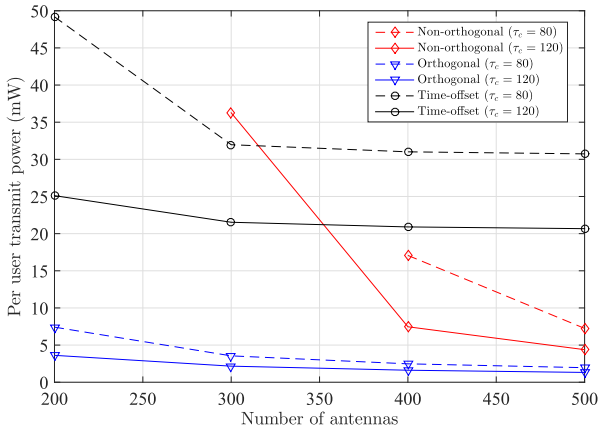


FIGURE 6. Per-user average transmit power versus the number of antennas (QoS=1.5 bit/Hz/s, $N = 30$ users).

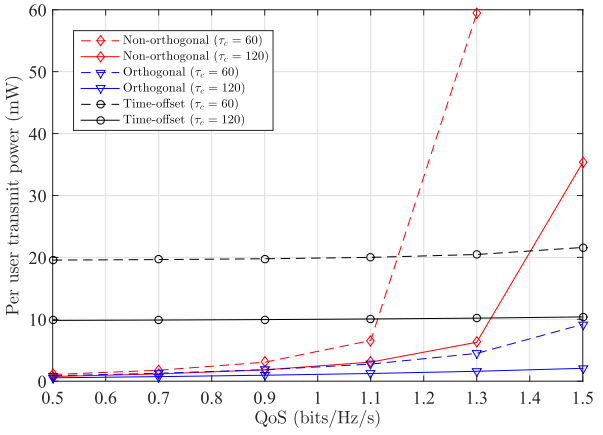


FIGURE 8. Per-user average transmit power versus required QoS level ($M = 300$ antennas, $N = 30$ users).

non-orthogonal pilot method enjoys a significant transmit power reduction when the number of antennas increases. However, the power minimization problem is not always feasible with non-orthogonal pilots. Specifically, when $\tau_c = 80$, the problem is feasible for the number of antennas larger than 400, whereas for $\tau_c = 120$ the minimum number of antennas required to have a feasible problem is 300. Similarly, the power consumption in the case of orthogonal pilots decreases when the number of antennas increases, but much slower. The same trend can be observed for time-offset pilots, where the power consumption drops noticeably when the number of antennas increases from 200 to 300.

The impact of coherence interval on the optimal power allocation is illustrated in Fig. 7. With the required QoS of 1.5 bits/sec/Hz and the number of antennas is 200 or 400, the power consumptions of all three pilot methods reduce when the coherence interval increases. Specifically, the power consumption of the time-offset pilot method decreases by 10 mW when the coherence interval increases from 60 to 120 symbols in both cases. With non-orthogonal pilots, the power minimization problem is infeasible with $M = 300$ antennas. When $M = 400$, this problem is solvable only when $\tau_c \geq 70$ symbols. In contrast, the transmit power reduction of the orthogonal pilot method is not very significant. This reduction in power consumption can be explained as a result of the lower required SINR when there are more time slots for data transmission.

Fig. 8 compares the change in per-user transmit power of the three pilot methods with respect to different required QoS levels when the BS has 300 antennas and the coherence interval is set at 60 and 120 symbols. Obviously, the transmit power increases when the required QoS increases. It can be seen that the slope of the curve under the non-orthogonal pilot method is much sharper than that of the two other methods. Noticeably, the curve with the time-offset pilot method only increases slightly when the required QoS increases from 0.5 to 1.5 bits/sec/Hz. The same tendency can also be observed in the case of the orthogonal pilot method but the change is larger.

Finally, Fig. 9 compare the sum SE of all users in the system between the orthogonal and proposed time-offset pilot methods. Although the per-user UL SE is lower with the

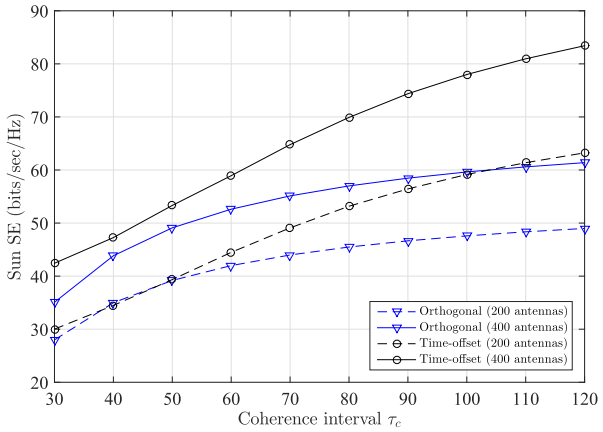


FIGURE 9. Comparison of the sum SE: $N = 15$ users with orthogonal pilots and $N = 30$ users with time-offset pilots.

time-offset pilot method than the orthogonal pilot method, with a fixed number of pilots sequences (here $\tau_p = 15$), the time-offset pilot method can serve twice the number of users (30 users) as compared to the conventional orthogonal pilot method (15 users). As a result, the sum SE is significantly larger with the time-offset pilot method than the orthogonal pilot method.

VI. CONCLUSION

This work investigated performance of time-offset pilots in the UL of a single-cell multiuser massive MIMO system. It is shown that the correlated interference, a consequence of the correlation between pilots of one group and UL data of the other group, can be effectively removed by applying successive interference cancellation. We further formulate power control problems for two different cost functions: max-min QoS and total power minimization. Due to the signomial constraints, these two problems are NP-hard and hence very computationally demanding. Therefore, we proposed algorithms to find the suboptimal solutions based on the bisection method, which solve a series of GPs. Numerical results have shown that the time-offset pilot method provides a far better performance than the non-orthogonal pilot method. The time-offset pilot is also better than the orthogonal pilot method when the coherence interval is short.

APPENDIX

From the received signal in (26), a lower bound on the UL ergodic SE of the qth user in the first group can be obtained based on the definition of the mutual information between the original base-band signal $x_{1,q}^{(\mathrm{tp})}$ and the received signal (after multiplied with the corresponding combining vector) $s_{1,q}$:

$$R_{1,q}^{(\mathrm{tp})} \geq I\left(x_{1,q}^{(\mathrm{tp})}; s_{1,q}, \hat{\mathcal{H}}\right), \tag{43}$$

where $\hat{\mathcal{H}}$ denote the knowledge of channel estimation at the BS. Under the input distribution $x_{1,q} \sim \mathcal{CN}(0, 1)$, the mutual information can be equivalently expressed as

$$\begin{aligned} I\left(x_{1,q}^{(\mathrm{tp})}; s_{1,q}, \hat{\mathcal{H}}\right) &= h(x_{1,q}^{(\mathrm{tp})}) - h\left(x_{1,q}^{(\mathrm{tp})} | s_{1,q}, \hat{\mathcal{H}}\right) \\ &= \log_2(\pi e) - h\left(x_{1,q}^{(\mathrm{tp})} | s_{1,q}, \hat{\mathcal{H}}\right), \end{aligned} \tag{44}$$

where $h(x_{1,q}^{(\mathrm{tp})})$ is the differential entropy and $h\left(x_{1,q}^{(\mathrm{tp})} | s_{1,q}, \hat{\mathcal{H}}\right)$ is the conditional entropy. Because of the fact that the entropy does not change when subtracting a known variable, $h\left(x_{1,q}^{(\mathrm{tp})} | s_{1,q}, \hat{\mathcal{H}}\right)$ can be bounded from above as

$$\begin{aligned} h\left(x_{1,q}^{(\mathrm{tp})} | s_{1,q}, \hat{\mathcal{H}}\right) &= h\left(x_{1,q}^{(\mathrm{tp})} - \alpha s_{1,q} | s_{1,q}, \hat{\mathcal{H}}\right) \\ &\leq h(x_{1,q}^{(\mathrm{tp})} - \alpha s_{1,q}) \\ &\leq \log_2\left(\pi e \mathbb{E}\left\{|x_{1,q}^{(\mathrm{tp})} - \alpha s_{1,q}|^2\right\}\right), \end{aligned} \tag{45}$$

where α is a deterministic scalar. The best upper bound for $h\left(x_{1,q}^{(\mathrm{tp})} | s_{1,q}, \hat{\mathcal{H}}\right)$ can be found by minimizing the expectation in (45) with respect to α. Since the UL data signals of users in the system are mutually independent, calculating the statical average $\mathbb{E}\left\{|x_{1,q}^{(\mathrm{tp})} - \alpha s_{1,q}|^2\right\}$ over $x_{g,k}^{(\mathrm{tp})}$ $(g = 1, 2)$ leads to a quadratic function of α:

$$\begin{aligned} &\mathbb{E}\left\{|x_{1,q}^{(\mathrm{tp})} - \alpha s_{1,q}|^2\right\} \\ &= 1 - 2\alpha \mathbb{E}\left\{\sum_{l=1}^{L} v_{1,q}^H h_{1,q}\right\} \sqrt{\rho_{1,q}^{(\mathrm{d})}} \\ &\quad + \alpha^2 \Bigg[\sum_{k=1}^{K} \left(\rho_{1,k}^{(\mathrm{d})} \mathbb{E}\left\{\left|\boldsymbol{v}_{1,q}^H \boldsymbol{h}_{1,k}\right|^2\right\} + \rho_{2,k}^{(\mathrm{p})} \mathbb{E}\left\{\left|\boldsymbol{v}_{1,q}^H \boldsymbol{h}_{2,k}\right|^2\right\}\right) \\ &\quad - \mathbb{E}\left\{\left|\hat{\Upsilon}_{1,q}^{(\mathrm{IP})}\right|^2\right\} + \sigma^2 \mathbb{E}\left\{\|\boldsymbol{v}_{1,q}\|^2\right\}\Bigg]. \end{aligned} \tag{46}$$

The minimum value of this quadratic function can be easily obtained as:

$$\mathbb{E}\left\{|x_{1,q}^{(\mathrm{tp})} - \alpha s_{1,q}|^2\right\} \geq \frac{1}{1 + \mathrm{SINR}_{1,q}^{(\mathrm{tp})}}. \tag{47}$$

where $\mathrm{SINR}_{1,q}^{(\mathrm{tp})}$ is defined in (48), as shown at the bottom of this page. By choosing this value, the tightest lower bound of $I\left(x_{1,q}^{(\mathrm{tp})}; s_{1,q}, \hat{\mathcal{H}}\right)$ is obtained. Finally, plugging the result from

$$\mathrm{SINR}_{1,q}^{(\mathrm{tp})} = \frac{\rho_{1,q}^{(\mathrm{d})} \left|\mathbb{E}\left\{\boldsymbol{v}_{1,q}^H \boldsymbol{h}_{1,q}\right\}\right|^2}{\sum_{k=1}^{K} \left(\rho_{1,k}^{(\mathrm{d})} \mathbb{E}\left\{\left|\boldsymbol{v}_{1,q}^H \boldsymbol{h}_{1,k}\right|^2\right\} + \rho_{2,k}^{(\mathrm{p})} \mathbb{E}\left\{\left|\boldsymbol{v}_{1,q}^H \boldsymbol{h}_{2,k}\right|^2\right\}\right) - \mathbb{E}\left\{\left|\hat{\Upsilon}_{1,q}^{(\mathrm{IP})}\right|^2\right\} - \rho_{1,q}^{(\mathrm{d})} \left|\mathbb{E}\left\{\boldsymbol{v}_{1,q}^H \boldsymbol{h}_{1,q}\right\}\right|^2 + \sigma^2 \mathbb{E}\left\{\|\boldsymbol{v}_{1,q}\|^2\right\}} \tag{48}$$

(44) to (47) into (43) we obtain the lower bound for UL SE that the qth user of the first group can achieve as in *Theorem 1*.

With the MRC, the combining vector for the qth user of the first group is $\boldsymbol{v}_{1,q} = \hat{\boldsymbol{h}}_{1,q}$, and we can calculate the closed-form SINR expression as follows.

The expected squared norm of the Rayleigh-distributed channel between the BS and the qth user is

$$\mathbb{E}\left\{\|\boldsymbol{v}_{1,q}\|^2\right\} = \mathbb{E}\left\{\|\hat{\boldsymbol{h}}_{1,q}\|^2\right\} = \gamma_{1,q}M. \tag{49}$$

and

$$\begin{aligned}\mathbb{E}\left\{\boldsymbol{v}_{1,q}^H\boldsymbol{h}_{1,q}\right\} &= \mathbb{E}\left\{\hat{\boldsymbol{h}}_{1,q}^H(\hat{\boldsymbol{h}}_{1,q}+\boldsymbol{e}_{1,q})\right\}\\ &= \mathbb{E}\left\{\|\hat{\boldsymbol{h}}_{1,q}\|^2\right\} = \gamma_{1,q}M,\end{aligned} \tag{50}$$

The expectation $\mathbb{E}\left\{\left|\boldsymbol{v}_{1,q}^H\boldsymbol{h}_{2,k}\right|^2\right\}$ is:

$$\begin{aligned}&\mathbb{E}\left\{\left|\boldsymbol{v}_{1,q}^H\boldsymbol{h}_{2,k}\right|^2\right\}\\ &= \gamma_{1,q}\beta_{2,k}M + \frac{\rho_{2,k}^{(\mathrm{d})}\rho_{1,q}^{(\mathrm{p})}\tau_p\beta_{1,q}^2}{\left(\rho_{1,q}^{(\mathrm{p})}\tau_p\beta_{1,q} + \sum_{k=1}^{K}\rho_{2,k}^{(\mathrm{d})}\beta_{2,k} + \sigma^2\right)^2}\\ &\quad\times\mathbb{E}\left\{\left\|\boldsymbol{h}_{2,k}\right\|^4\right\}\\ &= \gamma_{1,q}\beta_{2,k}M + \frac{\rho_{2,k}^{(\mathrm{d})}\rho_{1,q}^{(\mathrm{p})}\tau_p\beta_{1,q}^2}{\left(\rho_{1,q}^{(\mathrm{p})}\tau_p\beta_{1,q} + \sum_{k=1}^{K}\rho_{2,k}^{(\mathrm{d})}\beta_{2,k} + \sigma^2\right)^2}\\ &\quad\times\beta_{2,k}^2\frac{\Gamma(M+2)}{\Gamma(M)}\\ &= \gamma_{1,q}\beta_{2,k}M + \frac{\rho_{2,k}^{(\mathrm{d})}}{\rho_{1,q}^{(\mathrm{p})}}\tau_p\gamma_{1,q}^2\left(\frac{\beta_{2,k}}{\beta_{1,q}}\right)^2M(M+1).\end{aligned} \tag{51}$$

Consider the interference within the first group, when $k = q$, one has:

$$\begin{aligned}&\mathbb{E}\left\{\left|\boldsymbol{v}_{1,q}^H\boldsymbol{h}_{1,q}\right|^2\right\}\\ &= \mathbb{E}\left\{\left|\boldsymbol{v}_{1,q}^H\left(\hat{\boldsymbol{h}}_{1,q}+\boldsymbol{e}_{1,q}\right)\right|^2\right\}\\ &= \mathbb{E}\left\{\left|\boldsymbol{v}_{1,q}^H\hat{\boldsymbol{h}}_{1,q}\right|^2\right\} + \mathbb{E}\left\{\left|\boldsymbol{v}_{1,q}^H\boldsymbol{e}_{1,q}\right|^2\right\}\\ &= \gamma_{1,q}^2(M+M^2) + \gamma_{g',q}\left(\beta_{1,q}-\gamma_{1,q}\right)M\\ &= \left(\gamma_{1,q}M\right)^2 + \beta_{1,q}\gamma_{1,q}M.\end{aligned} \tag{52}$$

In the case when $k \neq q$, one has:

$$\mathbb{E}\left\{\left|\boldsymbol{v}_{1,q}^H\boldsymbol{h}_{1,k}\right|^2\right\} = \gamma_{1,q}\beta_{1,k}M. \tag{53}$$

With $\hat{\Upsilon}_{1,q}^{(IP)}$ being defined as in (17), the reduced amount of interference is:

$$\mathbb{E}\left\{\left|\hat{\Upsilon}_{1,q}^{(\mathrm{IP})}\right|^2\right\} = \sum_{k=1}^{K}\rho_{2,k}^{(\mathrm{p})}\frac{\rho_{2,k}^{(\mathrm{d})}}{\rho_{1,q}^{(\mathrm{p})}\tau_p}\gamma_{1,q}^2\left(\frac{\beta_{2,k}}{\beta_{1,q}}\right)^2M^2. \tag{54}$$

Substituting (49) to (54) into (48), one obtains the SINR as in (21).

REFERENCES

[1] F. Fernandes, A. Ashikhmin, and T. L. Marzetta, "Inter-cell interference in noncooperative TDD large scale antenna systems," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 192–201, Feb. 2013.

[2] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, Jan. 2013.

[3] X. Gao, O. Edfors, F. Rusek, and F. Tufvesson, "Massive MIMO performance evaluation based on measured propagation data," *IEEE Trans. Wireless Commun.*, vol. 14, no. 7, pp. 3899–3911, Jul. 2015.

[4] E. Björnson, E. G. Larsson, and T. L. Marzetta, "Massive MIMO: Ten myths and one critical question," *IEEE Commun. Mag.*, vol. 54, no. 2, pp. 114–123, Feb. 2016.

[5] 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.

[6] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO versus small cells," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1834–1850, Mar. 2017.

[7] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 61, no. 4, pp. 1436–1449, Apr. 2013.

[8] T. V. Chien and E. Björnson, "Massive MIMO communications," in *5G Mobile Communications*. Cham, Switzerland: Springer, 2017, pp. 77–116.

[9] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, *Fundamentals of Massive MIMO*. Cambridge, U.K: Cambridge Univ. Press, 2016.

[10] R. R. Müller, L. Cottatellucci, and M. Vehkaperä, "Blind pilot decontamination," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 773–786, Oct. 2014.

[11] 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.

[12] H. Q. Ngo, T. L. Marzetta, and E. G. Larsson, "Analysis of the pilot contamination effect in very large multicell multiuser MIMO systems for physical channel models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Prague, Czech Republic, 2011, pp. 3464–3467.

[13] O. Elijah, C. Y. Leow, T. A. Rahman, S. Nunoo, and S. Z. Iliya, "A comprehensive survey of pilot contamination in massive MIMO—5G system," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 905–923, 2nd Quart., 2016.

[14] N. Krishnan, R. D. Yates, and N. B. Mandayam, "Uplink linear receivers for multi-cell multiuser MIMO with pilot contamination: Large system analysis," *IEEE Trans. Wireless Commun.*, vol. 13, no. 8, pp. 4360–4373, Aug. 2014.

[15] J. C. Shen, J. Zhang, and K. B. Letaief, "Downlink user capacity of massive MIMO under pilot contamination," *IEEE Trans. Wireless Commun.*, vol. 14, no. 6, pp. 3183–3193, Jun. 2015.

[16] E. Björnson, E. de Carvalho, J. H. Sørensen, E. G. Larsson, and P. Popovski, "A random access protocol for pilot allocation in crowded massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 4, pp. 2220–2234, Apr. 2017.

[17] J. Li, D. Wang, P. Zhu, J. Wang, and X. You, "Downlink spectral efficiency of distributed massive MIMO systems with linear beamforming under pilot contamination," *IEEE Trans. Veh. Technol.*, vol. 67, no. 2, pp. 1130–1145, Feb. 2018.

[18] T. V. Chien, E. Björnson, and E. G. Larsson, "Joint pilot design and uplink power allocation in multi-cell massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 2000–2015, Mar. 2018.

[19] E. Björnson, E. G. Larsson, and M. Debbah, "Massive MIMO for maximal spectral efficiency: How many users and pilots should be allocated?" *IEEE Trans. Wireless Commun.*, vol. 15, no. 2, pp. 1293–1308, Feb. 2016.

[20] H. Al-Salihi, T. V. Chien, L. A. Tuan, and M. R. Nakhai, "A successive optimization approach to pilot design for multi-cell massive MIMO systems," in *IEEE Commun. Lett.*, vol. 22, no. 5, pp. 1086–1089, May 2018.

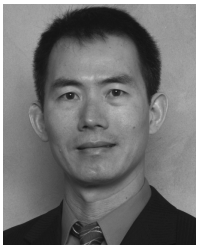
[21] W. A. W. M. Mahyiddin, P. A. Martin, and P. J. Smith, "Performance of synchronized and unsynchronized pilots in finite massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 14, no. 12, pp. 6763–6776, Dec. 2015.

[22] I. Atzeni, J. Arnau, and M. Debbah, "Fractional pilot reuse in massive MIMO systems," in *Proc. IEEE Int. Conf. Commun. Workshop (ICCW)*, London, U.K., Jun. 2015, pp. 1030–1035.

[23] X. Zhu, Z. Wang, L. Dai, and C. Qian, "Smart pilot assignment for massive MIMO," *IEEE Commun. Lett.*, vol. 19, no. 9, pp. 1644–1647, Sep. 2015.

[24] H. V. Cheng, E. Björnson, and E. G. Larsson, "Optimal pilot and payload power control in single-cell massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 65, no. 9, pp. 2363–2378, May 2017.
[25] D. Hu, L. He, and X. Wang, "Semi-blind pilot decontamination for massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 525–536, Jan. 2016.
[26] D. Kong, D. Qu, K. Luo, and T. Jiang, "Channel estimation under staggered frame structure for massive MIMO system," *IEEE Trans. Wireless Commun.*, vol. 15, no. 2, pp. 1469–1479, Feb. 2016.
[27] S. Jin, X. Wang, Z. Li, K.-K. Wong, Y. Huang, and X. Tang, "On massive MIMO zero-forcing transceiver using time-shifted pilots," *IEEE Trans. Veh. Technol.*, vol. 65, no. 1, pp. 59–74, Jan. 2016.
[28] B. Sun, Y. Zhou, L. Tian, and J. Shi, "Successive interference cancellation based channel estimation for massive MIMO systems," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Singapore, Dec. 2017, pp. 1–6.
[29] L. Wu, Z. Zhang, J. Dang, and H. Liu, "Enhanced time-shifted pilot based channel estimation in massive MIMO systems with finite number of antennas," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, Paris, France, May 2017, pp. 222–227.
[30] X. Xia, K. Xu, D. Zhang, Y. Xu, and Y. Wang, "Beam-domain full-duplex massive MIMO: Realizing co-time co-frequency uplink and downlink transmission in the cellular system," *IEEE Trans. Veh. Technol.*, vol. 66, no. 10, pp. 8845–8862, Oct. 2017.
[31] X. Xia, K. Xu, Y. Wang, and Y. Xu, "A 5G-enabling technology: Benefits, feasibility, and limitations of in-band full-duplex mMIMO," *IEEE Veh. Technol. Mag.*, vol. 13, no. 3, pp. 81–90, Sep. 2018.
[32] K. Xu, Z. Shen, Y. Wang, X. Xia, and D. Zhang, "Hybrid time-switching and power splitting swipt for full-duplex massive MIMO systems: A beam-domain approach," *IEEE Trans. Veh. Technol.*, vol. 67, no. 8, pp. 7257–7274, Aug. 2018.
[33] T. V. Chien, E. Björnson, and E. G. Larsson, "Joint power allocation and user association optimization for massive MIMO systems," in *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6384–6399, Sep. 2016.
[34] S. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ, USA: Prentice Hall, 1995.
[35] J. Li, E. Björnson, T. Svensson, T. Eriksson, and M. Debbah, "Joint precoding and load balancing optimization for energy-efficient heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 10, pp. 5810–5822, Jun. 2015.
[36] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.
[37] H. V. Cheng, E. Björnson, and E. G. Larsson, "Uplink pilot and data power control for single cell massive MIMO systems with MRC," in *Proc. Int. Symp. Wireless Commun. Syst. (ISWCS)*, Brussels, Belgium, Aug. 2015, pp. 396–400.
[38] M. Chiang, C. W. Tan, D. P. Palomar, D. O'neill, and D. Julian, "Power control by geometric programming," *IEEE Trans. Wireless Commun.*, vol. 6, no. 7, pp. 2640–2651, Jul. 2007.
[39] M. Grant, S. Boyd, and Y. Ye. (2015). *CVX: MATLAB Software for Disciplined Convex Programming*. [Online]. Available: http://www.stanford.edu/boyd/cvx
[40] *Further Advancements for E-UTRA Physical Layer Aspects (Release 9)*, document 3GPP TS 36.814, Mar. 2010.

THE KHAI NGUYEN received the B.Eng. degree from the School of Electronics and Telecommunications, Hanoi University of Science and Technology, Vietnam, in 2017. He is currently pursuing the M.S. degree with the University of Saskatchewan. His research fields include microprocessors, convex optimization, and future 5G mobile networks. His recent works concentrated in power allocation and user association for Massive MIMO.



HA H. NGUYEN (M'01–SM'05) received the B.Eng. degree from the Hanoi University of Technology, Hanoi, Vietnam, in 1995, the M.Eng. degree from the Asian Institute of Technology, Bangkok, Thailand, in 1997, and the Ph.D. degree from the University of Manitoba, Winnipeg, MB, Canada, in 2001, all in electrical engineering. He held adjunct appointments at the Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, MB, Canada, and was a Senior Visiting Fellow with the School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, NSW, Australia. He joined the Department of Electrical and Computer Engineering, University of Saskatchewan, Saskatoon, SK, Canada, in 2001, and became a Full Professor, in 2007. He is a coauthor, with editor Shwedyk, of the textbook *A First Course in Digital Communications* (Cambridge University Press). His research interests include the areas of communication theory, wireless communications, and statistical signal processing. He is a Fellow of the Engineering Institute of Canada and a Registered Member of the Association of Professional Engineers and Geoscientists of Saskatchewan. He was the Co-Chair of the Multiple Antenna Systems and Space-Time Processing Track, the IEEE Vehicular Technology Conferences (Fall 2010, Ottawa, ON, Canada and Fall 2012, Quebec, QC, Canada), the Lead Co-Chair of the Wireless Access Track, the IEEE Vehicular Technology Conferences (Fall 2014, Vancouver, BC, Canada), the Lead Co-Chair of the Multiple Antenna Systems and Cooperative Communications Track, the IEEE Vehicular Technology Conference (Fall 2016, Montreal, QC, Canada), and the Technical Program Co-Chair of the Canadian Workshop on Information Theory (St. John's, NL, Canada, 2015). He was an Associate Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, from 2007 to 2011. He currently serves as an Associate Editor of the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY and the IEEE WIRELESS COMMUNICATIONS LETTERS.

TIEN HOA NGUYEN received the Dipl.-Ing. degree in electronics and communication engineering from Hanover University and the Ph.D. degree from the Department Wireless Communication Technique, Hanoi University of Science and Technology, in 2016. He was with the Research and Development Department, Bosch, involved in image processing as well as development of SDR-based drivers. He had also three years experiments with MIMOon's Research and Development Team, to develop embedded signal processing and radio modules for 4G and 5G mobile networks. He is currently a Lecture with the Hanoi University of Science and Technology. His research interests include resource allocation in cognitive radios, massive MIMO, and vehicular communication systems.