Max-Min Fair Precoder Design for Non-Orthogonal Multiple Access

Ahmet Zahid Yalcin and Melda Yuksel IEEE, Senior Member

Abstract—In this paper, a downlink multiple input multiple output (MIMO) non-orthogonal multiple access (NOMA) wireless communication system is considered. In NOMA systems, the base station has unicast data for all users, and multiple users share the same resources. Maximizing the minimum cluster rate amongst multiple NOMA clusters through transmit precoding under quality-of-service (QoS) and total transmit power constraints is investigated. It is first shown that maximizing the minimum cluster rate and minimizing the maximum weighted cluster mean square error problems lead to the same optimal point. For the latter problem, Karush-Kuhn-Tucker (KKT) optimality conditions are derived and the expressions satisfied by the optimal receivers, minimum mean square error (MMSE) weights and the optimal precoders are obtained. Then, an iterative and low complexity transmit precoder design algorithm is proposed. Simulation results show that the proposed algorithm significantly outperforms orthogonal multiple access (OMA) and multi-user linear precoding (MULP) schemes.

Index Terms—Max-min fairness, mean square error, MIMO, NOMA, precoder design, quality-of-service.

I. INTRODUCTION

The demand for data traffic is steadily increasing and wireless networks of the next decade have to meet the high data rate requirements for many different applications [1]. To handle this high data rate, non-orthogonal multiple access (NOMA) is considered as a breakthrough technique, which enables simultaneous multiple access in the power domain for 5G wireless networks [2]. Specifically, downlink NOMA is an application of broadcast channels [3] and it relies on superposition coding (SPC) at the transmitter to transfer multiple data streams in the same resource block, and successive interference cancellation (SIC) at the receiver to cancel co-channel interference. NOMA has the potential to deliver higher system throughput [4], [5] and higher ergodic sum capacity [6], and to achieve better outage performance [7] compared to the existing orthogonal multiple access (OMA) techniques.

In NOMA systems, each user can have a dedicated precoding vector, or a cluster of users can share the same precoding vector. The former has the advantage of custom precoding for each user, but suffers from the rank constraints in the downlink multiuser MIMO broadcast channel [8]. The latter is not limited by rank, but messages are not individually precoded, so channel gain vectors and precoders are mismatched.

Assuming the transmit signals of each user are coded by a dedicated precoding vector, sum rate maximization, total power minimization and max-min fairness for NOMA systems are studied under different constraints and with different methods in the literature. The paper [8] solves the sum rate maximization problem by approximating the problem with a minorization-maximization algorithm (MMA). The paper [9] presents a precoding design for maximizing the sum rate of all users under decoding order and quality-of-service (QoS) constraints. Similarly, to maximize sum rate, [10] studies the channel state information (CSI) based singular value decomposition (CSI-SVD) precoding scheme. Total power minimization with QoS requirements and total power minimization under target interference level constraints are respectively investigated in [11] and [12]. Fairness in NOMA systems are studied in [13] and [14], but both are for single-antenna base stations, and the latter assumes only 2 users. Max-min fairness for non-clustered NOMA for a multiple antenna base station is studied in [11].

As mentioned above, in NOMA, a single precoder vector can be shared by a cluster of users. For this case, weighted sum rate optimization under a total power constraint when two users exist in each cluster is studied in [15]. For clustered downlink NOMA systems, a sub-optimal user clustering algorithm is proposed and the optimal power allocation policy that maximizes the weighted sum rate is derived in [16]. Similarly, [17] maximizes strong users’ sum rate subject to QoS constraints on weak users’ rates. Minimizing total transmission power for downlink clustered NOMA is studied in [18] and [19].

In this work, we study fairness for a downlink MIMO NOMA system, where the base station broadcasts precoded and superposed signals to clusters of users. To the best of our knowledge, there is no precoding study on fairness for clustered downlink NOMA. In our problem, the objective is to maximize the minimum cluster rate under QoS and total transmit power constraints. Due to the non-convexity of maximizing the minimum cluster rate, we use the equivalence between maximizing the minimum cluster rate and minimizing the maximum weighted cluster mean square error problems at the optimal point. We first prove that these two problems lead to the same optimal solution. Employing the KKT optimality conditions, we find the expressions the optimal receivers, minimum mean square error (MMSE) weights and the optimal precoders have to satisfy. Utilizing these expressions, we propose an iterative algorithm to evaluate the optimal precoders and the optimal receivers. We use the exponential penalty method to evaluate the Lagrange multipliers. We find that the proposed alternating optimization algorithm significantly improves the minimum cluster rate performance with respect to OMA and multi-user linear precoding (MULP) schemes.

Next, we explain the system model in Section II. We define the optimization problems in Section III and propose the precoder design in Section IV. We present the numerical results in Section V. Finally we provide conclusions and future
II. SYSTEM MODEL

In this paper we study a downlink multiuser MIMO system. The base station has $M$ transmit antennas and communicates with $K$ clusters. There are $L$ single antenna users in each cluster. Each user belongs to only one cluster.

The base station aims to send the data $s_{1,1}, \ldots, s_{K,L}$ to all $KL$ users. Here $s_{k,l}$ represents the data stream intended for the $l$-th user in the $k$-th cluster. All $s_{k,l}$ are independent and $\mathbb{E}\{s_{k,l}s_{k,l}^\dagger\} = \alpha_{k,l}$, where $\alpha_{k,l}$ is the ratio of power allocated to data stream for the $l$-th user in the $k$-th cluster. Moreover, $\sum_{l=1}^{L} s_{k,l} = 1$. To send all the messages, the base station superposes all the messages in a cluster and forms $s = [s_1, \ldots, s_K]^T \in \mathbb{C}^{K \times 1}$, where $s_k = \sum_{l=1}^{L} s_{k,l}$. When $p_k \in \mathbb{C}^{M \times 1}$ indicates the precoder vector for the $k$-th cluster, the base station transforms $s$ with the precoder matrix $P = [p_1, \ldots, p_K] \in \mathbb{C}^{M \times K}$. Then, the base station transmits $x \in \mathbb{C}^{M \times 1}$, which is equal to

$$x = Ps = \sum_{k=1}^{K} p_k s_k = \sum_{k=1}^{K} \sum_{l=1}^{L} p_k s_{k,l}. \quad (1)$$

The base station has an average total power constraint $E_{tx}$, which is written as

$$\mathbb{E}\{x^Hx\} = \text{Tr}(PP^H) \leq E_{tx}. \quad (2)$$

Then, the received signal at the $l$-th user in the $k$-th cluster becomes

$$y_{k,l} = h_{k,l}p_k \sum_{l=1}^{L} s_{l,k} + h_{k,l} \sum_{i=1,i \neq k}^{K} p_i s_i + n_{k,l}. \quad (3)$$

Here, $h_{k,l} \in \mathbb{C}^{1 \times M}$ is the effective channel gain vector of the $l$-th user in the $k$-th cluster. The effective channel gain is defined as $h_{k,l} = \tilde{h}_{k,l}/\sqrt{d_{k,l}}$, where $d_{k,l}$ is the distance between the $l$-th user in the $k$-th cluster and the base station, and $\rho$ is the path loss exponent. The entries in $\tilde{h}_{k,l}$ are independent and identically distributed (i.i.d.) and complex valued random variables. The noise component $n_{k,l}$ is a circularly symmetric complex Gaussian random variable with zero mean and unit variance, and $n_{k,l}$ are i.i.d. for all $k$ and $l$. The base station is informed about all effective channel gains $h_{k,l}$, while the receivers know only their own $h_{k,l}$.

A. Achievable Data Rates

For the clustered NOMA system we investigate, the messages for different clusters will be treated as noise, while SIC will be carried out within a cluster to limit intra-cluster interference. Due to SIC, in the $k$-th cluster, the $l$-th user’s message is decoded at the $i$-th user, for which $l \leq i$. In other words, the first user in the cluster decodes its own message only, and the $L$-th user decodes all user’s messages within the cluster. Moreover, decoding is ordered and starts from

the first user’s message. Then, for this NOMA system, the signal to interference ratio (SINR) for decoding the $l$-th user’s message at the $i$-th user in the $k$-th cluster, $i = 1, \ldots, L, l = 1, \ldots, K$, can be written as

$$\gamma_{k,i \rightarrow l} = \alpha_{k,l} |h_{k,i} p_k|^2 r_{k,i \rightarrow l}^{-1}. \quad (4)$$

In the above equation, $r_{k,i \rightarrow l}$ is the effective noise variance and is defined as

$$r_{k,i \rightarrow l} = \sum_{j=1}^{L} \alpha_{k,j} |h_{k,i} p_k|^2 + I_{k,i} + 1. \quad (5)$$

where, $I_{k,i} = \sum_{i=1,i \neq k}^{K} |h_{k,i} p_i|^2$ and is equal to the inter-cluster interference at user-$i$ in the $k$-th cluster. Then, in the $k$-th cluster, the $i$-th user’s achievable rate for decoding the $l$-th user’s message is

$$R_{k,i \rightarrow l} = \log \left(1 + \gamma_{k,i \rightarrow l}\right). \quad (6)$$

Overall, the achievable rate for the $l$-th user’s message in the $k$-th cluster is defined as the minimum of all $R_{k,i \rightarrow l}$, and is denoted as

$$R_k = \min_{i,l \in \{1, \ldots, L\}} R_{k,i \rightarrow l}. \quad (7)$$

B. Error Variance Definitions

It is well known that mutual information and minimum mean square error (MMSE) are related [20], [21], and we can state $R_k$ in terms of error variances, assuming MMSE receivers are employed at the receivers.

To estimate the $l$-th user’s message, the $i$-th user in the $k$-th cluster employs the SIC receiver $V_{k,i \rightarrow l}$ on its equivalent received signal as

$$\hat{s}_{k,i \rightarrow l} = V_{k,i \rightarrow l}(y_{k,l} - h_{k,l}p_k \sum_{j=1}^{l-1} s_{j,k}) \quad (8)$$

and obtains its estimate $\hat{s}_{k,i \rightarrow l}$. Then, the MSE of the $i$-th user’s estimate of the $l$-th user’s message in the $k$-th cluster becomes

$$\varepsilon_{k,i \rightarrow l} = \mathbb{E}\{||\hat{s}_{k,i \rightarrow l} - s_{k,l}||^2\},$$

$$= |V_{k,i \rightarrow l}|^2 T_{k,i \rightarrow l} + \alpha_{k,l} - 2R \{\alpha_{k,l} |V_{k,i \rightarrow l}| h_{k,i} p_k\}, \quad (9)$$

where

$$T_{k,i \rightarrow l} = |h_{k,i} p_k|^2 \alpha_{k,l} + r_{k,i \rightarrow l}. \quad (10)$$

Given above, the optimal MMSE receiver can be written as

$$V_{k,i \rightarrow l}^{\text{mmse}} = \arg \min_{V_{k,i \rightarrow l}} \varepsilon_{k,i \rightarrow l},$$

$$= \frac{V_{k,i \rightarrow l}^H h_{k,i} T_{k,i \rightarrow l}^{-1}}{\alpha_{k,l}}. \quad (11)$$

When this MMSE receiver in (11) is employed, the resulting error variance expression in (9) becomes

$$\varepsilon_{k,i \rightarrow l}^{\text{mmse}} = \left(\frac{1}{\alpha_{k,l}} + |h_{k,i} p_k|^2 r_{k,i \rightarrow l}^{-1}\right)^{-1}. \quad (12)$$

1^In fact, the results can easily be extended to cover for unequal number of users in each group. However, to keep the notation simple we adhere to a fixed number of users in each cluster.
As the message for the \( l \)-th user has to be decoded by all users \( i \) for which \( i \geq l \) in the \( k \)-th cluster, we define \( \varepsilon_{k,l}^{\text{mmse}} \) as

\[
\varepsilon_{k,l}^{\text{mmse}} = \max_{i,l \in \{1, \ldots, L\}} \varepsilon_{k,i \to l}^{\text{mmse}}.
\] (13)

Note that, the rate and MMSE expressions in (6) and (12) are related, and we can write

\[
R_{k,i \to l} = \log \left( \alpha_{k,l} \varepsilon_{k,i \to l}^{\text{mmse}} \right). \tag{14}
\]

III. PROBLEM DEFINITIONS

A. Max-Min Fair Sum Rate

In this subsection, we define the max-min fair (MMF) sum rate optimization problem, which aims to find the optimal precoder matrix \( \mathbf{P} \), such that the minimum of each cluster rate is maximized subject to a total power constraint and a minimum rate constraint for each user. The cluster rate is defined as the sum of the rates of the users in the same cluster. Then, the optimization problem is stated as

\[
P_1: \max_{\mathbf{P}, \mathbf{R}} \min_k \left( \sum_{l=1}^{L} R_{k,l} \right) \tag{15a}
\]

s.t. \( R_{k,l} \leq R_{k,i \to l}, \forall k, l, i = \{1, \ldots, L\} \) \hspace{1cm} (15b)

\[
R_{k,l}^{\text{th}} \leq R_{k,l}, \forall k, l \tag{15c}
\]

\[
\text{Tr}(\mathbf{PP}^H) \leq E_{\text{tx}} \tag{15d}
\]

where, \( R_{k,l} \) is defined in (7), \( \mathbf{R} = [R_1, \ldots, R_1, \ldots, R_{K,1}, \ldots, R_{K,L}] \) and \( R_{k,l}^{\text{th}} \geq 0 \) is the threshold data rate that user-\( l \) in the \( k \)-th cluster has to satisfy.

To solve this problem, we restate the problem by adding an auxiliary variable \( \bar{c} \) and convert \( P_1 \) to a new optimization problem as

\[
P_1': \max_{\mathbf{P}, \mathbf{R}, \bar{c}} \bar{c} \tag{16a}
\]

s.t. \( c \leq \sum_{l=1}^{L} R_{k,l}, \forall k \) \hspace{1cm} (16b)

\[
R_{k,l} \leq R_{k,i \to l}, \forall k, l, i = \{1, \ldots, L\} \tag{16c}
\]

\[
R_{k,l}^{\text{th}} \leq R_{k,l}, \forall k, l \tag{16d}
\]

\[
\text{Tr}(\mathbf{PP}^H) \leq E_{\text{tx}} \tag{16e}
\]

This problem is still non-convex and hard to solve. Next, we convert this problem once more into an equivalent WMMSE problem using the relation between rate and MMSE. To do that, we introduce the augmented weighted MSE defined as

\[
\xi_{k,i \to l} = b_{k,i \to l} \varepsilon_{k,i \to l} - \log(\alpha_{k,l} b_{k,i \to l}), \tag{17}
\]

where \( b_{k,i \to l} > 0 \) is introduced as the weight for MSE. We also define

\[
\xi_{k,l}^{\text{mmse}} = b_{k,i \to l} \varepsilon_{k,i \to l}^{\text{mmse}} - \log(\alpha_{k,l} b_{k,i \to l}), \tag{18}
\]

\[
\xi_{k,l}^{\text{mmse}} = \max_{i,l \in \{1, \ldots, L\}} \xi_{k,i \to l}^{\text{mmse}}. \tag{19}
\]

B. Min-Max Fair WMMSE

Similar to the MMF sum rate problem, the MMF WMMSE problem is defined as

\[
P_2: \min_{\mathbf{P}, \mathbf{E}} \max_k \left( \sum_{l=1}^{L} \xi_{k,l}^{\text{mmse}} \right) \tag{20a}
\]

s.t. \( \xi_{k,i \to l}^{\text{mmse}} \leq \xi_{k,l}^{\text{mmse}}, \forall k, l, i = \{1, \ldots, L\} \) \hspace{1cm} (20b)

\[
\xi_{k,l}^{\text{mmse}} \leq \xi_{k,l}^{\text{th}}, \forall k, l \tag{20c}
\]

\[
\text{Tr}(\mathbf{PP}^H) \leq E_{\text{tx}} \tag{20d}
\]

where \( \mathbf{E} = [\xi_{1,1}^{\text{mmse}}, \ldots, \xi_{1,L}^{\text{mmse}}, \ldots, \xi_{K,1}^{\text{mmse}}, \ldots, \xi_{K,L}^{\text{mmse}}] \) and \( \xi_{k,l}^{\text{th}} \) is the threshold error variance for the \( l \)-th user’s data in the \( k \)-th cluster.

The problem \( P_2 \) is not convex. To solve \( P_2 \), we restate \( P_2 \) as a smooth constrained optimization problem by introducing an auxiliary variable \( \bar{c} \) and write

\[
P_2': \min_{\mathbf{P}, \mathbf{E}, \bar{c}} \bar{c} \tag{21a}
\]

s.t. \( \sum_{l=1}^{L} \xi_{k,l}^{\text{mmse}} \leq \bar{c}, \forall k \) \hspace{1cm} (21b)

\[
\xi_{k,i \to l}^{\text{mmse}} \leq \xi_{k,l}^{\text{mmse}}, \forall k, l, i = \{1, \ldots, L\} \tag{21c}
\]

\[
\xi_{k,l}^{\text{mmse}} \leq \xi_{k,l}^{\text{th}}, \forall k, l \tag{21d}
\]

\[
\text{Tr}(\mathbf{PP}^H) \leq E_{\text{tx}} \tag{21e}
\]

C. Gradient Expressions for Both MMF Problems

In this subsection, we find the gradient expressions and KKT conditions for the MMF sum rate and the MMF WMMSE problems respectively defined in (16) and (21). Comparing the Lagrangian expressions for the two problems, we will prove that they are equivalent at the optimal solution point.

To solve the problems in (16) and (21), we write the Lagrangian expressions for both problems \( P_1' \) and \( P_2' \) respectively as

\[
f(\mathbf{P}, \mathbf{R}, c, \mu, \kappa, \psi, \lambda) = -c + \sum_{k=1}^{K} \mu_k c - \sum_{k=1}^{K} \sum_{l=1}^{L} \mu_k R_{k,l}
\]

\[+ \sum_{k=1}^{K} \sum_{l=1}^{L} \kappa_{k,l} (R_{k,l}^{\text{th}} - R_{k,l}) + \lambda (\text{Tr}(\mathbf{PP}^H) - E_{\text{tx}}) \]

\[+ \sum_{k=1}^{K} \sum_{l=1}^{L} \psi_{k,i \to l} (R_{k,l} - R_{k,i \to l}), \tag{22}\]

\[
g(\mathbf{P}, \mathbf{E}, \bar{c}, \bar{\mu}, \bar{\kappa}, \bar{\psi}, \bar{\lambda}) = \bar{c} - \sum_{k=1}^{K} \bar{\mu}_k \bar{c} + \sum_{k=1}^{K} \sum_{l=1}^{L} \bar{\mu}_k \xi_{k,l}^{\text{mmse}}
\]

\[+ \sum_{k=1}^{K} \sum_{l=1}^{L} \bar{\kappa}_{k,l} (\xi_{k,l}^{\text{mmse}} - \xi_{k,l}^{\text{th}}) + \bar{\lambda} (\text{Tr}(\mathbf{PP}^H) - E_{\text{tx}})
\]

\[+ \sum_{k=1}^{K} \sum_{l=1}^{L} \bar{\psi}_{k,i \to l} (\xi_{k,i \to l}^{\text{mmse}} - \xi_{k,l}^{\text{mmse}}), \tag{23}\]

where \( \{\mu, \kappa, \psi, \lambda\} \) and \( \{\bar{\mu}, \bar{\kappa}, \bar{\psi}, \bar{\lambda}\} \) denote the Lagrange multiplier sets for \( f \) and \( g \) respectively. We derive \( \nabla_{\mathbf{P}} f, \nabla_{\mathbf{P}} g \) in Appendix A and state them in (24) and (25) as
\[ \nabla_{\mathbf{p}_k} f(\mathbf{P}, \mathbf{R}, \mathbf{c}, \mu, \kappa, \psi, \lambda) = \sum_{l=1}^{L} \sum_{i=1}^{L} \left[ -\sum_{l=1}^{L} \sum_{j=1}^{L} \alpha_{k,j}^{*} \psi_{k,i}^{*} - c^{*} \right] \psi_{k,i} - \sum_{l=1}^{L} \sum_{i=1}^{L} \psi_{l,i}^{*} \psi_{k,i} \mathbf{h}_{k,i}^{H} \mathbf{p}_{k} + \lambda \mathbf{p}_{k} \]

\[ - \sum_{l=1}^{L} \sum_{i=1}^{L} \psi_{l,i}^{*} \psi_{k,i} \mathbf{h}_{k,i}^{H} \mathbf{p}_{k} + \sum_{t \neq k}^{L} \sum_{l=1}^{L} \sum_{i=1}^{L} \psi_{t,i}^{*} \psi_{k,i} \mathbf{h}_{k,i}^{H} \mathbf{p}_{t} \mathbf{h}_{l,i}^{H} \mathbf{h}_{l,i} \mathbf{p}_{k} + \lambda \mathbf{p}_{k} \]

(24)

\[ \nabla_{\mathbf{p}_k} g(\mathbf{P}, \mathbf{R}, \bar{\mu}, \bar{\kappa}, \bar{\psi}, \bar{\lambda}) = \sum_{l=1}^{L} \sum_{i=1}^{L} \left[ -\sum_{l=1}^{L} \sum_{j=1}^{L} \alpha_{k,j}^{*} \psi_{k,i}^{*} - c^{*} \right] \psi_{k,i} - \sum_{l=1}^{L} \sum_{i=1}^{L} \psi_{l,i}^{*} \psi_{k,i} \mathbf{h}_{k,i}^{H} \mathbf{p}_{k} + \lambda \mathbf{p}_{k} \]

\[ - \sum_{l=1}^{L} \sum_{i=1}^{L} \psi_{l,i}^{*} \psi_{k,i} \mathbf{h}_{k,i}^{H} \mathbf{p}_{k} + \sum_{t \neq k}^{L} \sum_{l=1}^{L} \sum_{i=1}^{L} \psi_{t,i}^{*} \psi_{k,i} \mathbf{h}_{k,i}^{H} \mathbf{p}_{t} \mathbf{h}_{l,i}^{H} \mathbf{h}_{l,i} \mathbf{p}_{k} \]

(25)

**D. KKT Conditions for Both Problems and Equivalence**

For the MMF sum rate problem, in addition to \( \nabla_{\mathbf{p}_k} f \), we calculate \( \partial c f \) and \( \partial R_{k,i} f \) as

\[ \partial c f = -1 + \sum_{k=1}^{K} \mu_k, \]

(26)

\[ \partial R_{k,i} f = -\mu_k - \kappa_{k,i} + \sum_{l=1}^{L} \psi_{l,i}^{*} \psi_{k,i}, \]

(27)

Then, due to the KKT conditions, a local optimum must satisfy \( \nabla_{\mathbf{p}_k} f = 0 \), \( \partial R_{k,i} f = 0 \), \( \partial c f = 0 \) and the following complementary slackness conditions

\[ \mu_k^{*} \left[ c^{*} - \sum_{l=1}^{L} R_{k,l}^{*} \right] = 0, \]

(28)

\[ \kappa_{k,i}^{*} \left[ R_{k,l}^{*} - R_{k,i}^{*} \right] = 0, \]

(29)

\[ \psi_{k,i}^{*} \left[ R_{k,l}^{*} - R_{k,i}^{*} \right] = 0, \]

(30)

\[ \lambda^{*} \left[ \mathbf{Tr}(\mathbf{P}^{*} \mathbf{P}^{*H}) - E_{tx} \right] = 0, \]

(31)

where \( * \) is used to denote the optimal value. In this problem, we assume the total power constraint is always satisfied with equality, as increasing the power will always increase the achievable rates. Similarly, for the MMF WMMSE problem

\[ \partial c g = 1 - \sum_{k=1}^{K} \mu_k \]

(32)

\[ \partial R_{k,i} g = \mu_k + \kappa_{k,i} - \sum_{l=1}^{L} \psi_{k,i}^{*} \psi_{l,i} \]

(33)

Then, a local optimum must satisfy \( \nabla_{\mathbf{p}_k} g = 0 \), \( \partial R_{k,i} g = 0 \), and \( \partial c g = 0 \). At the locally optimum point \( \left( \mathbf{P}^{*}, \xi_{k,l}^{mms}, c^{*}, \mu_k, \kappa_{k,i}^{*}, \psi_{k,i}^{*}, \lambda^{*} \right) \), due to complementary slackness, we must have

**IV. Iterative Precoder Design**

The problem defined in (23) is hard to solve and there are no closed form expressions for the optimal precoders. Instead, in this section we propose an iterative precoder design algorithm.

The optimization problem in (23) assumes that the optimal MMSE receiver defined in (11) is employed at all users, and finds the optimal precoders at the transmitter. Below, we first define a generalized problem which allows for arbitrary receivers \( V_{k,i} \) that attain \( \xi_{k,i} \) in (2).
\[ P_3 : \ \min_{\mathbf{P}, \mathbf{E}, \mathbf{c}, \mathbf{V}, \mathbf{b}} \ \bar{c} \]  
\[ \text{s.t.} \ \sum_{i=1}^{L} \xi_{k,l} \leq \bar{c}, \ \forall k,l \]  
\[ \xi_{k,i} \leq \xi_{k,i}^{\text{th}}, \ \forall k, i = \{1, \ldots, L\} \]  
\[ \xi_{k,l} \leq \xi_{k,l}^{\text{th}}, \ \forall k, l \]  
\[ \text{Tr}(\mathbf{PP}^H) \leq E_{tx}. \]  

When \( \theta_k, \Gamma_k, \eta_{k,i} \) and \( \beta \) denote the Lagrange multipliers for (39) and \( \mathbf{V} \) and \( \mathbf{b} \) consist of all receivers \( V_{k,i=1} \) and weights \( b_{k,i} \) respectively, the Lagrangian objective function of (39) is defined as

\[ h(\mathbf{P}, \mathbf{R}, \mathbf{c}, \theta, \Gamma, \eta, \beta) = \bar{c} - \sum_{k=1}^{K} \theta_k \bar{c} + \sum_{k=1}^{K} \sum_{l=1}^{L} \theta_k \xi_{k,l} \]
\[ + \sum_{k=1}^{K} \sum_{l=1}^{L} \Gamma_k(l) (\xi_{k,l} - \xi_{k,l}^{\text{th}}) + \beta (\text{Tr}(\mathbf{PP}^H) - E_{tx}) \]
\[ + \sum_{k=1}^{K} \sum_{l=1}^{L} \sum_{i=1}^{L} \eta_{k,i} \bar{\xi}_{k,i} (\bar{\xi}_{k,i} - \xi_{k,i}). \]  

(40)

Studying the KKT conditions for (40), similar to the analysis in the previous section, we can state the following theorem.

**Theorem 1:** For the optimization problem defined in (39), the following data receiver \( V_{k,i=1} \), the Lagrange multiplier \( \beta \), and the transmit precoders \( \mathbf{P} \) defined in (43) satisfy the KKT conditions.

\[ V_{k,i=1} = \alpha_{k,i} \mathbf{p}_k^H h_{k,i}^H T_{k,i=1}, \]  

(41)

\[ \beta = \frac{1}{E_{tx}} \left[ \sum_{k=1}^{K} \sum_{l=1}^{L} \sum_{i=1}^{L} \bar{\xi}_{k,i} b_{k,i} V_{k,i=1} \right]^2, \]  

(42)

\[ \mathbf{p}_k = \left[ \beta I + \sum_{l=1}^{L} \sum_{i=1}^{L} \sum_{j=1}^{L} \alpha_{k,j} \bar{\xi}_{k,i} b_{k,i} h_{k,i}^H V_{k,i=1} | h_{k,i}^H |^2 h_{k,i} \right. \]
\[ + \sum_{l=1}^{L} \sum_{i=1}^{L} \sum_{j=1}^{L} \bar{\xi}_{k,i} b_{k,i} h_{k,i}^H V_{k,i=1} | h_{k,i}^H |^2 h_{k,i} \]
\[ \left. \sum_{l=1}^{L} \sum_{i=1}^{L} \sum_{j=1}^{L} \eta_{k,i} \bar{\xi}_{k,i} b_{k,i} h_{k,i}^H V_{k,i=1} | h_{k,i}^H |^2 h_{k,i} \right]^{-1}. \]  

(43)

**Proof:** The proof is provided in Appendix B.

**Remark 1:** The receiver \( V_{k,i=1} \) in (41) is exactly equal to the MMSE receiver \( V_{k,i=1}^{\text{mmse}} \) given in (1). \( \square \)

**Remark 2:** When the optimal MMSE receiver \( V_{k,i=1}^{\text{mmse}} \) and the weights \( b_{k,i} \) in (38) are substituted in \( \xi_{k,i} \) of (17), then \( \xi_{k,i} \) becomes equal to \( \xi_{k,i}^{\text{mmse}}. \)

Utilizing Theorem 1 we propose solving for the receivers (41), the Lagrange multiplier (42) and the precoders (43) in an iterative fashion in Algorithm 1. However, calculating the Lagrange multipliers set \( \{\theta_k, \Gamma_k, \eta_{k,i}\} \) for problem (39) in Algorithm 1 is not trivial. In (22), an exponential penalty method is suggested to solve min-max type problems. According to the exponential penalty method, in each iteration of the algorithm, we update \( \{\theta_k, \Gamma_k, \eta_{k,i}\} \) as

\[ \theta_k = \frac{\exp \left\{ \frac{1}{\nu} \left( \sum_{l=1}^{L} \xi_{k,l} - \bar{c} \right) \right\}}{\sum_{k=1}^{K} \exp \left\{ \frac{1}{\nu} \left( \sum_{l=1}^{L} \xi_{k,l} - \bar{c} \right) \right\}}, \]  

(44)

\[ \Gamma_k = \exp \left\{ \frac{1}{\nu} \left( \xi_{k,i} - \xi_{k,i}^{\text{th}} \right) \right\}, \]  

(45)

\[ \eta_{k,i} = \exp \left\{ \frac{1}{\nu} \left( \xi_{k,i} - \xi_{k,i}^{\text{th}} \right) \right\}, \]  

(46)

In the above equations, \( \nu \) is a constant and as long as \( \nu \geq (\log KL)/\epsilon \), the solution is \( \epsilon \)-optimal. Note that, this choice satisfies the KKT conditions on \( \{\theta_k, \Gamma_k, \eta_{k,i}\} \) since \( \sum_{k=1}^{K} \theta_k = 1 \), and \( \sum_{l=1}^{L} \eta_{k,i} = \Gamma_k + \theta_k \geq 0, \Gamma_k \geq 0, \eta_{k,i} \geq 0 \).

In each iteration, the algorithm increases the objective function since there is a total power constraint. Thus, the proposed WMMSE algorithm converges to an upper limit. This limit is within an \( \epsilon \) neighborhood of a local optimum, as the algorithm utilizes the equations found via the KKT conditions, and the exponential penalty method is employed. Following similar steps as in [21 Section IV-A] and [23], one can prove convergence in full detail.

Note that, the expressions given in Theorem 1 are valuable as they significantly accelerate Algorithm 1. When these expressions are used, the algorithm ends in a couple of minutes. However, if in each iteration of the algorithm, an optimization toolbox such as CVX [24] is used, the algorithm takes hours or days.

**V. Numerical Results**

In this section, we present numerical results to evaluate the performance of the proposed transmission strategy MMF WMMSE NOMA. In particular, we compute the proposed transmit precoder using Algorithm 1 and compare it with orthogonal transmission and multiuser linear precoding.

In the simulations, the entries in \( \mathbf{h}_{\mathbf{k}} \) are assumed to be i.i.d. circularly symmetric complex Gaussian random variables with zero mean and unit variance, and the users are uniformly distributed in a circular region. We assume that the path loss...
exponent $\rho = 4$. The effective channel gain magnitudes are ordered and $|h_{k,L}| > |h_{k,L-1}| > \ldots > |h_{k,1}|$. It means that we name the user with the smallest effective channel gain magnitude as the first user in a cluster and the $L$-th user has the largest channel gain magnitude. The users are clustered according to the user selection scheme proposed in [16, Algorithm 1, Figure 3]. In this user selection scheme, the aim is to put users, which have highly different effective channel gain magnitudes, $|h_{k,l}|$, in the same cluster. Thus, the base station sorts the channel gains of all users. For $L = 2$, the base station then puts the user with the highest effective channel gain magnitude in the same cluster with the worst user in the first cluster. Then the second best and second worst users are put in the second cluster. The remaining clusters are formed in a similar fashion. For $L = 3$, the users are ordered, and divided into 3 groups, as the first $K$ best users, the second best $K$ users and the worst $K$ users. The best users in the first and second groups are then clustered with the worst user in the last group. Similarly, the second best users in the first and second groups are clustered with the second worst user in the last group. All clusters are formed in a similar fashion.

In this work, we assume a simple power allocation scheme. The total power is equally distributed among the clusters. However, within each cluster, the power level of each user is inversely proportional to its effective channel gain magnitude squared, $|h_{k,l}|^2$. This power allocation scheme supports fairness as it tries to balance out the rate terms $R_{k,l}$. The presented results are averaged over 1000 channel realizations. In Algorithm 1, the maximum number of iterations is limited to 200, and $\Gamma$ is set to $10^{-3}$ and $\Delta$ is set to 3. $\Delta$ is used to tune the algorithm to satisfy the rate constraint $R_{k,l}$. The transmit signal to noise ratio (SNR) is defined as $E_{tx}/\sigma^2$. Here $\sigma^2$ is the noise variance and set to 1. In the following simulation results, we consider algorithm convergence and MMF WMMSE results of Algorithm 1 for different settings.

A. Convergence Analysis

Fig. 1 shows the convergence behavior of the proposed alternating optimization scheme given by Algorithm 1 for three different initial points. In the figure, the initial precoder matrix, $P^{\text{init}}$, can be the scaled identity matrix, the singular value decomposition (SVD) based precoder or a randomly chosen matrix. If it is chosen as the scaled identity matrix, then $P^{\text{init}} = qI$, where $q$ is selected so as to satisfy the transmit power constraint. On the other hand, the SVD based initial precoder is calculated as follows. Let $H_k$ be the channel matrix of the $k$-th cluster; i.e. $H_k = [h_{k,1}, \ldots, h_{k,L}]$, with the singular value decomposition $H_k = USV^T$. Then $P^{\text{init}} = U(:,1)$ where $U(:,1)$ denotes the first column of $U$. In the figure, $M = 4$, $K = 4$, $L = 2$, $R_{k,l}^h = 0.2$, $\forall k,l$, and the total transmit power is set to 10 dB. The figure confirms that the proposed algorithm converges fast and to the same optimal point for all three initial points.

B. Orthogonal Multiple Access and Multiuser Linear Precoding

In this subsection, we describe the two multiple access schemes we use for comparison: orthogonal multiple access (OMA) and multiuser linear precoding (MULP). To make a fair comparison, we have to define the MMF rates for OMA and MULP, MMF OMA and MULP respectively.

1) OMA: In OMA, the transmission time is divided into $L$ equal slots and the base station transmits data to $K$ users in each time slot. In other words, the base station communicates with one user from each cluster in each time slot. For each time slot-$l$, the input data vector is denoted as $s_{l,\text{OMA}} = [s_{l,1}, \ldots, s_{l,K}]^T \in \mathbb{C}^{K \times 1}$. We assume all $s_{k,l}$ are independent and $E\{s_{k,l}s_{k,l}^H\} = 1$. The input data vector $s_{l,\text{OMA}}$ is linearly processed in time slot-$l$ by a precoder matrix $P_{l,\text{OMA}} = [P_{l,\text{OMA}}^{1}, \ldots, P_{l,\text{OMA}}^{K}] \in \mathbb{C}^{M \times K}$, where the precoding vector $P_{k,l,\text{OMA}} \in \mathbb{C}^{M \times 1}$ is dedicated to the $k$-th user in the time slot-$l$. The overall transmit data vector $x_{l,\text{OMA}} = P_{l,\text{OMA}} s_{l,\text{OMA}}$. Then, the SINR at user-$k$ in time slot-$l$ is given by

$$\gamma_{k,l,\text{OMA}} = \frac{|h_{k,l}P_{k,l,\text{OMA}}|^2}{\sum_{i=1,i \neq k}^{K} |h_{k,i}P_{i,l,\text{OMA}}|^2 + 1}.$$ 

and the corresponding rate expression is calculated as $R_{k,l,\text{OMA}} = \frac{1}{2} \log(1 + \gamma_{k,l,\text{OMA}})$. To find the MMF OMA rate, we solve the following optimization problem for $\forall k = 1, 2, \ldots, K$:

$$\min_{P_{l,\text{OMA}}^*} R_{k,l,\text{OMA}}^* \quad \text{s.t.} \quad R_{k,l,\text{OMA}}^* \leq R_{k,l}^h, \quad \text{Tr}(P_{l,\text{OMA}}^*P_{l,\text{OMA}}^H) \leq E_{tx},$$

and define $R_{k,l,\text{OMA}}^*$ as the rates attained with the precoder $P_{l,\text{OMA}}^*$ at all the $K$ users served at time slot $l$. Then, the MMF OMA rate $R_{\text{OMA}}$ can be calculated as

$$R_{\text{OMA}} = \min_{k,k=1,2,\ldots,K} \sum_{l=1}^{L} R_{k,l,\text{OMA}}^*.$$
In other words, at each time slot $l$, the optimization problem maximizes the minimum rate over $K$ users served in that time slot. The sum cluster rate is calculated as the sum of the rates of all users in the cluster and the MMF rate is the minimum of all such cluster rates.

2) MULP: In MULP precoding, the base station transmits data to all $KL$ users simultaneously. The input data vector is denoted as $s_{MULP}^T = [s_{1,1}, \ldots, s_{1,L}, \ldots, s_{K,1}, \ldots, s_{K,L}] \in \mathbb{C}^{KL \times 1}$. We assume all $s_{k,l}$ are independent and $E\{s_{k,l}^*s_{k,l}\} = 1$. The input data vector $s_{MULP}$ is linearly processed by a precoder matrix $P_{MULP} = [P_{MULP,1}^T, \ldots, P_{MULP,L}^T, \ldots, P_{K,1}^T, \ldots, P_{K,L}^T] \in \mathbb{C}^{M \times KL}$, where the precoding vector $P_{k,l}^T \in \mathbb{C}^{M \times 1}$ is dedicated to the $l$-th user of the $k$-th cluster. We also define $P_{MULP} = [P_{1,1}^T, \ldots, P_{1,L}^T, \ldots, P_{K,1}^T, \ldots, P_{K,L}^T] \in \mathbb{C}^{M \times K}$. Then, the overall transmit data vector $x_{MULP}^T = P_{MULP}^T s_{MULP}$. The SINR at user $l$ in the $k$-th cluster is given by

$$\gamma_{k,l}^{MULP} = \frac{|h_{k,l} P_{k,l}^{MULP}|^2}{\sum_{j \neq l}^L |h_{k,j} P_{k,j}^{MULP}|^2 + \sum_{i \in [K]} \sum_{l=1}^L |h_{i,l} P_{i,l}^{MULP}|^2 + 1},$$

and the corresponding rate expression is calculated as $R_{k,l}^{MULP} = \log (1 + \gamma_{k,l}^{MULP})$. Then, we solve the following problem for $\forall k, k = 1, 2, \ldots, K$

$$P_{MULP}^{L,k,l} = \arg \max_{P_{k,l}^{MULP}} \min_k R_{k,l}^{MULP}$$

s.t. $R_{l,MULP}^{th} \leq R_{k,l}^{MULP}$

and $\text{Tr}(P P^H) \leq E_{tx}$. (50c)

When $R_{l,MULP}^{L,k,l}$ is defined as the rates attained with the precoder $P_{MULP}^{L,k,l}$ at the $l$-th user of each cluster, the MMF MULP rate $R_{MULP}^{MULP}$ can be calculated as

$$R_{MULP}^{MULP} = \min_{k,l} \sum_{l=1}^L R_{k,l}^{MULP}^{L,k,l}.$$

To solve both optimization problems stated for OMA and MULP, we first find their equivalent weighted MMSE problems as done in Section III and solve them in an iterative fashion as in Section IV. However, instead of solving for the KKT conditions, in each iteration of the algorithm, we employ the CVX toolbox [24].

C. Comparisons with MMF WMMSE NOMA

Fig. 2 compares the proposed MMF WMMSE NOMA algorithm with MMF OMA and MMF MULP schemes for $M = 4, K = 4, L = 2$, and $R_{l,MULP}^{th} = 0.2$ bits. In this figure, we show the minimum cluster rate for each of the precoding schemes. As shown in the figure, proposed NOMA algorithm significantly outperforms MMF OMA and MMF MULP schemes. At 20 dB, the NOMA scheme achieves 6 bits/channels use, while OMA and MULP respectively attain approximately 5.8 and 3.4 bits/channel use. Due to the lack of degrees of freedom (DoF), MMF MULP converges for high transmit SNR. While OMA enjoys full DoF, it suffers from time division and is inferior to the proposed NOMA scheme.

The MMF WMMSE NOMA reaps the benefits of full DoF and superposition coding and becomes superior to all.

In Fig. 3, we investigate the effect of number of users in each cluster. We simulate a system with $M = K = 3, L = 2$ and $M = K = 3, L = 3$ for the rate constraint $R_{l,MULP}^{th} = 0.2$ bits for all users. We observe that proposed NOMA algorithm outperforms OMA and MULP schemes for both cases and increasing the number of users in each cluster decreases the MMF rates for all schemes. As the number of users in each cluster increase, it is harder to attain a large rate for all users.

Finally, Fig. 4 shows the impact of the user distances in the same cluster. The simulation is run for $M = 2, K = 2, L = 2$ for the given rate constraint $R_{l,MULP}^{th} = 0.2$ bits and transmit SNR is set to 10dB. For this figure, the simulation setting is slightly different from above. One of the users is uniformly located on a disk with radius $d$, $0 < d < 1$. The other user is
assumed to be uniformly located on an annulus in between radii $d$ and 1. As a result, as $d$ decreases, $1/d$ increases and the effective channel gain magnitudes the users observe become more imbalanced on average. We observe that when the asymmetry in the system increases, the MMF rate of the proposed NOMA algorithm increases significantly. The MULP scheme performs the worst as it suffers from DoF limitations.

VI. CONCLUSION

In this paper, we study precoder and receiver design for maximizing the minimum cluster rate in downlink MIMO non-orthogonal multiple access (NOMA). In NOMA, users’ messages in each cluster are combined using superposition coding and then precoded at the transmitter. To cancel intra-cluster interference, receivers in the same cluster employ successive interference cancellation. We prove that maximizing the minimum cluster rate, and minimizing the maximum cluster MMSE are equivalent at the optimal solution, and solve the problem via alternating optimization. We show that the NOMA scheme significantly outperforms orthogonal multiple access (OMA) and multi-user linear precoding (MULP). The effects of the number of users in each cluster and the discrepancy between the user distances within a cluster are also investigated. It is of future interest to study the proposed algorithm’s performance in mmWave massive MIMO systems with limited feedback.

APPENDIX A

In this appendix, we derive $\nabla_{p_k} f(P, R, c, \mu, \psi, \lambda)$ and $\nabla_{p_k} g(P, R, \bar{c}, \bar{\mu}, \bar{\psi}, \bar{\lambda})$. The Lagrangian objective function $f$ is given in (22). First, we need $\nabla_{p_k} R_{t,i} - I$ for both $t = k$ and $t \neq k$. Note that $\nabla \log X = (\nabla X) X^{-1}$, and $\nabla X (X^H) AX = AX$ [25] ch E.3.

First, assume $t = k$. Using (13),

$$\nabla_{p_k} R_{k,i} - I = \nabla_{p_k} [e_{mmse}^{-1}]_{e_{mmse}}.$$  \hspace{1cm} (51)

The gradient $\nabla_{p_k} [e_{mmse}^{-1}]$ is calculated by applying the chain rule as

$$\nabla_{p_k} [e_{mmse}^{-1}] =$$

$$\nabla_{p_k} \left( \frac{1}{\alpha_{k,t}} + p_k h_k^H r_{k,i}^{-1} h_k, p_k \right) =$$

$$p_k h_k^H \frac{\partial (r_{k,i}^{-1})}{\partial (p_k^H p_k^H)} h_k, p_k + \frac{\partial (p_k^H h_k^H)}{\partial (p_k^H p_k^H)} r_{k,i}^{-1} h_k, p_k$$

$$+ p_k h_k^H \frac{\partial (h_k, p_k)}{\partial (p_k^H p_k^H)}$$

$$= -p_k h_k^H r_{k,i}^{-2} \frac{\partial (r_{k,i}^{-1})}{\partial (p_k^H p_k^H)} h_k, p_k + e_m h_k^H r_{k,i}^{-1} h_k, p_k.$$  \hspace{1cm} (52)

where $e_m$ is the unity column vector with 1 at the $m^{th}$ element and zeros elsewhere and is of size $M \times 1$. The last term is 0 since $\frac{\partial p_k^H}{\partial p_k^H} = 0$. On the other hand, we have

$$\nabla_{p_k} r_{k,i} - I = \frac{\partial (r_{k,i}^{-1})}{\partial (p_k^H p_k^H)} = \sum_{j=l+1}^{L} \alpha_{k,j} h_k, p_k e_m H_{k,i}.$$  \hspace{1cm} (53)

as $\nabla X (X^H) = XA$ [25]. Then, substituting (53) in (52), we obtain

$$\nabla_{p_k} e_{mmse}^{-1} = e_m h_k^H r_{k,i}^{-1} h_k, p_k$$

$$- \sum_{j=l+1}^{L} \alpha_{k,j} h_k^H h_k, p_k e_m h_k^H h_k, p_k.$$  \hspace{1cm} (54)

Using (53) and (55), \hspace{1cm} (51) becomes

$$\nabla_{p_k} R_{k,i} - I = h_k^H r_{k,i}^{-1} h_k, p_k e_{mmse}^{-1}$$

$$- \sum_{j=l+1}^{L} \alpha_{k,j} h_k^H h_k, p_k e_{mmse}^{-1} p_k h_k^H r_{k,i}^{-2} h_k, p_k.$$  \hspace{1cm} (56)

Next, we compute $\nabla_{p_k} R_{t,i} - I$, for $t \neq k$

$$\nabla_{p_k} R_{t,i} - I = p_t^H h_t^H \nabla_{p_k} [r_{t,i}^{-1}]_{e_{mmse}} h_t, p_k e_{mmse}^{-1}. \hspace{1cm} (57)$$

Using $\nabla (X^{-1}) = -X^{-1} \nabla (X) X^{-1}$ \hspace{1cm} (26), we can write

$$\nabla_{p_k} [r_{t,i}^{-1}]_{e_{mmse}} = -r_{t,i}^{-1} \nabla_{p_k} [r_{t,i}^{-1}]_{e_{mmse}}.$$  \hspace{1cm} (58)

Then we compute

$$\nabla_{p_k} [r_{t,i}^{-1}]_{e_{mmse}} = h_t, p_k e_{m} h_t^H.$$  \hspace{1cm} (59)

By combining (57), (58) and (59) we have

$$\nabla_{p_k} = R_{t,i} - I = -e_{m} h_t^H r_{t,i}^{-2} h_t, p_k e_{mmse}^{-1} p_t^H h_t^H h_t, p_k.$$  \hspace{1cm} (60)

\hspace{1cm} (26) Note that, the gradient of a function $f(x)$ with respect to its variable $x$ is denoted as $\nabla_{x=1} f(x)$ and its $m^{th}$ element is defined as $[\nabla_{x=1} f(x)]_m = \nabla_{x=1} f(x) = \frac{\partial f(x)}{\partial x^m}$, where $*$ indicates conjugation.
Overall, for \( t \neq k \), we have

\[
\nabla_{p_k} R_{t,i} = -h_{k,i}^H \frac{2}{t,i} h_{t,i} p_k e^{\text{mmse}} p_k^H h_{k,i}^H h_{t,i} p_k.
\]  (61)

The gradient of last term in (22) is

\[
\nabla_{p_k} \lambda \text{Tr}([PP^H]) = \lambda p_k.
\]  (62)

Combining (56), (61) and (62), we find \( \nabla_{p_k} f(P, R, c, \mu, \psi, \lambda) \) as stated in (24). The gradient \( \nabla_{p_k} g(P, R, c, \mu, \psi, \lambda) \) is computed in a similar manner as \( \nabla_{p_k} f(P, R, c, \mu, \psi, \lambda) \) to be found as in (25).

**APPENDIX B**

In this appendix, we prove Theorem 1. Taking the derivative of the objective function \( h \) in (10) with respect to \( V_{k,i} \), then equating it to zero, we obtain

\[
\alpha_{k,l} p_k^H h_{k,i}^H = \sum_{j=1}^{L} \alpha_{k,j} h_{k,i} p_k^H h_{k,i}^H V_{k,i} \quad l = 1
\]

\[
+ \sum_{t=1, t \neq k}^{K} h_{k,i} p_k^H h_{k,i}^H V_{k,i} + V_{k,i}.
\]  (63)

Then, when \( \eta_{k,i} > 0 \),

\[
V_{k,i} = \alpha_{k,l} p_k^H h_{k,i}^H V_{k,i}^{-1}.
\]  (64)

Secondly, taking the gradient of (10) with respect to \( p_k^H \), and equating it to zero, we have the following equation

\[
\sum_{l=1}^{L} \sum_{i=1}^{L} \eta_{k,i} V_{k,i}^* \alpha_{k,l} = \sum_{l=1}^{L} \sum_{i=1}^{L} \sum_{j=1}^{L} \alpha_{k,j} \eta_{k,i} V_{k,i}^* V_{k,i} V_{k,i}^* h_{k,i} p_k
\]

\[
+ \sum_{l=1}^{L} \sum_{i=1}^{L} \sum_{t=1, t \neq k}^{K} \eta_{k,i} V_{k,i}^* |V_{k,i}|^2 h_{k,i} p_k + \beta p_k.
\]  (65)

Then,

\[
\beta = \frac{1}{E_{tx}} \sum_{k=1}^{K} \sum_{l=1}^{L} \sum_{i=1}^{L} \eta_{k,i} V_{k,i}^* |V_{k,i}|^2.
\]  (67)

To calculate \( \beta \), we post-multiply both sides of (65) by \( V_{k,i}^* \) and perform

\[
\sum_{k=1}^{K} \sum_{l=1}^{L} \sum_{i=1}^{L} |V_{k,i}| \eta_{k,i} V_{k,i}^* |V_{k,i}|^2 h_{k,i}.
\]  (66)

are also equal to each other. As we assume that the power constraint in (2) is satisfied with equality we can find that

\[
\beta = \frac{1}{E_{tx}} \left[ \sum_{k=1}^{K} \sum_{l=1}^{L} \sum_{i=1}^{L} \eta_{k,i} V_{k,i}^* |V_{k,i}|^2 \right].
\]  (67)

**REFERENCES**

[1] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, “What will 5G be?” *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[2] L. Dai, B. Wang, Y. Yuan, S. Han, C. L. and Z. Wang, “Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends,” *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74–81, Sep. 2015.

[3] T. Cover, “Broadcast channels,” *IEEE Trans. Inf. Theory*, vol. 18, no. 1, pp. 2–14, Jan. 1972.

[4] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Hirachi, “Non-orthogonal multiple access (NOMA) for cellular future radio access,” in *IEEE 77th Vehicular Technol. Conf. (VTC)*, Spring, Jun. 2013, pp. 1–5.

[5] A. Benjebbour, A. Li, Y. Saito, Y. Kishiyama, A. Harada, and T. Nakamura, “System-level performance of downlink NOMA for future LTE enhancements,” in *2013 IEEE Globecom Workshops (GC Wkshps)*, Dec. 2013, pp. 66–70.

[6] M. Zeng, A. Yadav, O. A. Dobre, G. I. Tsiropoulos, and H. V. Poor, “Capacity comparison between MIMO-NOMA and MIMO-OMA with multiple users in a cluster,” *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2413–2424, Oct. 2017.

[7] Z. Ding, D. Yang, P. Fan, and H. V. Poor, “On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users,” *IEEE Signal Process. Lett.*, vol. 21, no. 12, pp. 1501–1505, Dec. 2014.

[8] M. F. Hanif, Z. Ding, T. Ratnarajah, and G. K. Karagiannidis, “A minirontization-maximization method for optimizing sum rate in the downlink of non-orthogonal multiple access systems,” *IEEE Trans. Signal Process.*, vol. 64, no. 1, pp. 76–88, Jan. 2016.

[9] F. Zhu, Z. Lu, J. Zhu, J. Wang, and Y. Huang, “Beamforming design for downlink non-orthogonal multiple access systems,” *IEEE Access*, vol. 6, pp. 10956–10965, 2018.

[10] C. Chen, W. Cai, X. Cheng, L. Yang, and Y. Jin, “Low complexity beamforming and user selection schemes for 5G MIMO systems,” *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2706–2722, Dec. 2017.

[11] F. Alavi, K. Cumanan, Z. Ding, and A. G. Burr, “Beamforming techniques for nonorthogonal multiple access in 5G cellular networks,” *IEEE Trans. Veh. Technol.*, vol. 67, no. 10, pp. 9474–9487, Oct. 2018.

[12] Z. Chen, Z. Ding, P. Xu, and X. Dai, “Optimal precoding for a QoS optimization problem in two-user MISO-NOMA downlink,” *IEEE Commun. Lett.*, vol. 20, no. 6, pp. 1263–1266, Jun. 2016.

[13] S. Timotheou and I. Krikidis, “Fairness for non-orthogonal multiple access in 5G systems,” *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1647–1651, Oct. 2015.

[14] J. Choi, “Power allocation for sum-rate and max-min rate proportional fairness in NOMA,” *IEEE Communications Letters*, vol. 20, no. 10, pp. 2055–2058, Oct 2016.

[15] X. Sun, D. Duran-Herrmann, Z. Zhong, and Y. Yang, “Non-orthogonal multiple access with weighted sum-rate optimization for downlink broadcast channel,” in *IEEE Military Commun. Conf.*, Oct. 2015, pp. 1176–1181.

[16] M. S. Ali, H. Tabassum, and E. Fossien, “Dynamic user clustering and power allocation for uplink and downlink non-orthogonal multiple access (NOMA) systems,” *IEEE Access*, vol. 4, pp. 6325–6343, 2016.

[17] X. Sun, C. Shen, Y. Xu, S. M. Al-Basit, Z. Ding, N. Yang, and Z. Zhong, “Joint beamforming and power allocation design in downlink non-orthogonal multiple access systems,” in *2016 IEEE Globecom Workshops (GC Wkshps)*, Dec. 2016, pp. 1–6.

[18] J. Choi, “Minimum power multicast beamforming with superposition coding for multiresolution broadcast and application to NOMA systems,” *IEEE Trans. Commun.*, vol. 63, no. 3, pp. 791–800, Mar. 2015.

[19] Z. Liu, L. Lei, N. Zhang, G. Kang, and S. Chatzinotas, “Joint beamforming and power optimization with iterative user clustering for MISO-NOMA systems,” *IEEE Access*, vol. 5, pp. 6872–6884, 2017.

[20] D. Guo, S. Shamai, and S. Verdú, “Mutual information and minimum mean-square error in Gaussian channels,” *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1261–1282, Apr. 2005.
[21] S. S. Christensen and R. Agarwal and E. D. Carvalho and J. M. Cioffi, “Weighted sum-rate maximization using weighted MMSE for MIMO-BC beamforming design,” IEEE Trans. Wireless Commun., vol. 7, no. 12, pp. 4792–4799, Dec. 2008.

[22] S. P. X.S. Li, “Solving the finite min-max problem via an exponential penalty method,” Comput. Tech., vol. 8, pp. 3–15, 2003.

[23] J. Kaleva, A. Tili, and M. Juntti, “Decentralized sum rate maximization with QoS constraints for interfering broadcast channel via successive convex approximation,” IEEE Trans. Signal Process., vol. 64, no. 11, pp. 2788–2802, Jun. 2016.

[24] M. Grant and S. Boyd, “CVX: Matlab software for disciplined convex programming, version 2.1,” [http://cvxr.com/cvx] Mar. 2014.

[25] T. Moon and W. Stirling, Mathematical Methods and Algorithms for Signal Processing. Prentice Hall, 2000. [Online]. Available: https://books.google.com.tr/books?id=1xUfAQAAIAAJ

[26] K. B. Petersen and M. S. Pedersen, “The matrix cookbook.” November 2012, version 20121115.