# Globally Optimal Spectrum- and Energy-Efficient Beamforming for Rate Splitting Multiple Access

Bho Matthiesen, *Member, IEEE*, Yijie Mao, *Member, IEEE*, Armin Dekorsy, *Senior Member, IEEE*, Petar Popovski, *Fellow, IEEE*, Bruno Clerckx, *Fellow, IEEE*,

***Abstract*—Rate splitting multiple access (RSMA) is a promising non-orthogonal transmission strategy for next-generation wireless networks. It has been shown to outperform existing multiple access schemes in terms of spectral and energy efficiency when suboptimal beamforming schemes are employed. In this work, we fill the gap between suboptimal and truly optimal beamforming schemes and conclusively establish the superior spectral and energy efficiency of RSMA. To this end, we propose a successive incumbent transcending (SIT) branch and bound (BB) algorithm to find globally optimal beamforming solutions that maximize the weighted sum rate or energy efficiency of RSMA in Gaussian multiple-input single-output (MISO) broadcast channels. Numerical results show that RSMA exhibits an explicit globally optimal spectral and energy efficiency gain over conventional multi-user linear precoding (MU-LP) and power-domain non-orthogonal multiple access (NOMA). Compared to existing globally optimal beamforming algorithms for MU-LP, the proposed SIT BB not only improves the numerical stability but also achieves faster convergence. Moreover, for the first time, we show that the spectral/energy efficiency of RSMA achieved by suboptimal beamforming schemes (including weighted minimum mean squared error (WMMSE) and successive convex approximation) almost coincides with the corresponding globally optimal performance, making it a valid choice for performance comparisons. The globally optimal results provided in this work are imperative to the ongoing research on RSMA as they serve as benchmarks for existing suboptimal beamforming strategies and those to be developed in multi-antenna broadcast channels.**

***Index Terms*—Rate splitting multiple access (RSMA), rate splitting, global optimization, spectral efficiency, energy efficiency, multiple-input single-output (MISO), broadcast channel (BC), interference networks, next generation multiple access, non-orthogonal transmission**

## I. Introduction

Over the past few years, rate splitting multiple access (RSMA), built upon the concept of rate splitting (RS), has emerged as a promising non-orthogonal physical-layer transmission paradigm for interference management and multiple access in modern multi-antenna communication networks [2]–[4]. The key design principle of RSMA is to partially decode the multi-user interference and partially treat it as noise. This is done by splitting user messages into common and private parts, respectively, and then transmitting them employing superposition coding [5]. The common message is decoded by multiple users, while the private message is only decoded by the corresponding user employing successive interference cancellation (SIC). By flexibly adjusting the message splits according to the interference level, RSMA allows arbitrary combinations of joint decoding and treating interference as noise. Therefore, RSMA is a powerful interference management and multiple access strategy that softly bridges and subsumes existing schemes such as space division multiple access, which fully treats interference as noise, non-orthogonal multiple access (NOMA), which fully decodes interference, and orthogonal multiple access, which completely avoids interference by allocating orthogonal radio resources among users [3], [6].

The concept of RS was introduced forty years ago for the two-user single-input single-output interference channel [7], [8]. However, the use of RS as a building block of RSMA is motivated by the recent progress of RS-based network design in multi-antenna wireless networks: RS was first shown to achieve the optimal degrees of freedom (DoF) region of underloaded multiple-input single-output (MISO) broadcast channels (BCs) with partial channel state information at the transmitter (CSIT) in [9] and of overloaded MISO BCs with heterogeneous CSIT in [10]. The DoF benefits of RS subsequently motivated investigations of RSMA precoder design at finite signal-to-noise ratios (SNRs) [3], [6], [11]–[16]. There are two general lines of research for RSMA resource allocation, namely, low-complexity beamforming design [6], [17]–[19], [19]–[23] and beamforming optimization [3], [10]–[12], [14]–[16], [20], [24]–[29] mainly with respect to maximizing the spectral efficiency (SE) or energy efficiency (EE). Low-complexity beamforming approaches such as using random beamforming (RBF) [17]–[19], [30], matched beamforming (MBF) [6], [19]–[22], [31], or singular value decomposition (SVD) [11] for the common stream together with zero-forcing (ZF) [17], [18], [30] or reguralized-zero forcing beamforming (RZF)/minimum mean square error (MMSE) [19], [21], [22], [31] for the private streams have been widely studied in RSMA-aided MISO BC. Block diagonalization [18], [23] has been further investigated when the receivers are equipped with multiple antennas. Instead, precoder optimization strives to find optimal beamformers that

Parts of this paper were presented at 2021 IEEE International Conference on Acoustics, Speech and Signal Processing [1].

B. Matthiesen and A. Dekorsy are with the Department of Communications Engineering, University of Bremen, 28359 Bremen, Germany (e-mail: {matthiesen, dekorsy}@ant.uni-bremen.de). Y. Mao is with School of Information Science and Technology, ShanghaiTech University, 387433 Shanghai, China, (e-mail: maoyj@shanghaitech.edu.cn). P. Popovski is with the Department of Electronic Systems, Aalborg University, 9220 Aalborg, Denmark. He is also a U Bremen Excellence Chair in the Department of Communications Engineering, University of Bremen, 28359 Bremen, Germany (e-mail: petarp@es.aau.dk). B. Clerckx is with the Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, U.K. (e-mail: b.clerckx@imperial.ac.uk).

This work is supported in part by the German Research Foundation (DFG) under grant EXC 2077 (University Allowance), by the U.K. Engineering and Physical Sciences Research Council (EPSRC) under grants EP/N015312/1 and EP/R511547/1, and by the North-German Supercomputing Alliance (HLRN).

TABLE I
SURVEY OF EXISTING RSMA BEAMFORMING DESIGN APPROACHES.

| Beamforming approach | | Maximize SE | Maximize EE |
|---|---|---|---|
| Low complexity | RBF + ZF | [17], [18], [30] | — |
| | RBF + RZF | [19] | — |
| | MBF + ZF | [6] | [20] |
| | MBF + RZF | [19], [21], [22], [31] | — |
| | SVD + ZF | [11] | — |
| | MBF + BD | [23] | — |
| Suboptimal | WMMSE-based | [3], [10], [11], [13], [24] | — |
| | SCA-based | [13]–[16], [20], [29] | [12], [25], [32] |
| | ADMM-based | [26], [33], [34] | — |
| | SDR-based | [16], [29] | — |
| Globally optimal | SIT BB | This work | This work |

maximize achievable performance regions of RSMA. Existing beamforming design algorithms such as weighted minimum mean squared error (WMMSE) [3], [10], [11], [13], [24], successive convex approximation (SCA) [12]–[16], [20], [25], [25], [29], [32], alternating direction method of multipliers (ADMM) [26], [33], [34], and semidefinite relaxation (SDR) [16], [27]–[29] have been investigated for RSMA. SCA-based algorithms follow the classical idea of successively approximating the original non-convex problem with a sequence of convex approximations. WMMSE and ADMM-based algorithms are the block-wise alternative optimization where the variables are divided into blocks and the original problem is optimized alternatively with respect to a single block of variables while the rest of the blocks are held fixed. All of them could only guarantee a local optima of the original problem [35]. Though SDR has the potential to find the global optimum, existing works all focus on combining SDR with other efficient approaches such as SCA [16], [29], gradient-based approach [27], particle swarm optimization [28], heuristic approaches [36] in order to reduce the computational complexity. Therefore, the solutions obtained in these works cannot ensure global optimality. Table I summarizes the state-of-the-art beamforming design approaches that have been proposed for RSMA. None of them exhibits strong optimality guarantees. Hence, all performance analyses based on these approaches are incapable of conclusively establishing the superiority of RSMA over the previously mentioned multiplexing schemes. To the best of the authors knowledge, there is no existing work focusing on the globally optimal beamforming design of RSMA, and the maximum SE and EE performance achieved by RSMA remains unknown.

The goal of this paper is to bridge this gap and derive an algorithm to determine a globally optimal beamforming solution for RSMA with respect to weighted sum rate (WSR) and EE maximization. The corresponding optimization problem is related to joint multicast and unicast precoding that is known to be NP-hard [37], [38]. While several globally optimal algorithms for unicast beamforming [39], [40] and multicast beamforming [41] exist, joint solution methods are scarce. In particular, the procedure in [42] solves the power minimization problem and [43] maximizes the WSR for joint multicast and unicast beamforming. All these methods are based on branch and bound (BB) in combination with the second-order cone (SOC) transformation in [44]. However, as this transformation moves the complexity into the feasible set, pure BB methods are prone to numerical problems, see Section III. Instead, in this paper we design a successive incumbent transcending (SIT) BB algorithm to solve this beamforming problem with improved numerical stability and faster convergence. To the best of the authors knowledge, this is the first globally optimal solution algorithm for an instance of the joint unicast and multicast problem with respect to EE maximization.

To summarize, the contributions of this paper are:

1) We develop a numerical solver for the WSR and EE beamforming problem in RSMA with guaranteed convergence to a globally optimal solution. We emphasize the novelty of the globally optimal EE maximization method for an instance of the joint unicast and multicast beamforming problem.
2) We apply the successive incumbent transcending (SIT) principle to a MISO beamforming problem. The proposed algorithm incorporates multi-user linear precoding (MU-LP) and 2-user NOMA beamforming as special cases. It exhibits faster practical convergence and improved numerical stability over state-of-the-art solution methods for MU-LP beamforming.
3) From a theoretical perspective, we establish finite convergence to the optimal solution. This property does not hold for most BB-based beamforming algorithms.
4) Extensive numerical verification is done, both to assess the numerical properties of the proposed algorithm and to evaluate the performance of suboptimal state-of-the-art methods. In particular, we show that these methods, including WMMSE and SCA, are often close to the true optimum solution.

The paper organization continues as follows. In the next section, we define the system model, formally state the optimization problem and transform it into an equivalent form more suitable for numerical solution. In Section III, the mathematical fundamentals of the proposed algorithm are reviewed. These are applied in Section IV to derive the solution algorithm and prove its convergence. We close the paper with numerical experiments in Section V and a short discussion.

*Notation:* Scalars and functions are typeset in normal font $x$. The absolute value of $\cdot$ is $|\cdot|$ and $v(\mathrm{n})$ is the optimal value of the optimization problem in equation (n). $\Re\{\cdot\}$ and $\Im\{\cdot\}$ are the real and imaginary parts of a complex number, $j$ is the imaginary unit, and $\angle\cdot$ is the argument of $(\cdot)$. A vector $\boldsymbol{x}$ has components $[x_1, x_2, \dots]^T$ and is a column vector unless noted otherwise. The all-zero and all-ones vectors are denoted as $\mathbf{0}$ and $\mathbf{1}$, respectively. The operators $(\cdot)^T$, $(\cdot)^H$, and $\|\cdot\|$ are the transpose, the conjugate transpose and the Euclidean norm, respectively. Scalar operators are applied element-wise to vectors, where relational operators evaluate to true if they hold element-wise for all elements. A set is written as $\mathcal{X}$ and a family of sets as $\mathscr{X}$. The sets of real and complex numbers are denoted as $\mathbb{R}$ and $\mathbb{C}$. The notation $\mathcal{X} \setminus x$ is a shorthand for $\mathcal{X} \setminus \{x\}$. Let $\mathcal{X}$ and $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X}$. Then, $\mathrm{proj}_{\boldsymbol{x}} \mathcal{X} = \{\boldsymbol{x} : (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X} \text{ for some } \boldsymbol{y}\}$, i.e., the projection of $\mathcal{A}$ onto the $\boldsymbol{x}$ coordinates.

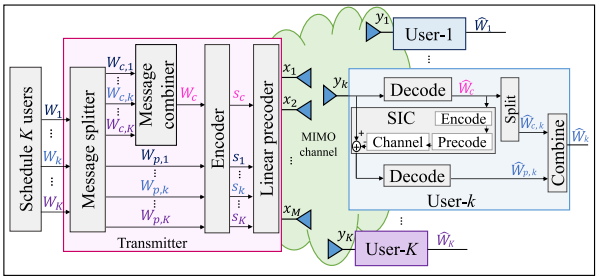



Fig. 1. 1-layer RS model for $K$ users, where the common stream $s_c$ is shared by all users.

## II. System Model & Problem Statement

Consider the downlink in a wireless network where an $M$ antenna base station (BS) serves $K$ single-antenna users. The received signal at user $k$, $k \in \mathcal{K} = \{1, \ldots, K\}$, for each channel use is $y_k = \boldsymbol{h}_k^H \boldsymbol{x} + n_k$, where the transmit signal $\boldsymbol{x} \in \mathbb{C}^{M\times 1}$ is subject to an average power constraint $P$, $\boldsymbol{h}_k$ is the complex-valued channel from the BS to user $k$, and $n_k$ is circularly symmetric complex white Gaussian noise with unit power at user $k$. We assume perfect channel state information at the transmitter and receivers.

The transmitter employs 1-layer RS [3], [11] as illustrated in Fig. 1, i.e., it splits the message $W_k$ intended for user $k$ into a common part $W_{c,k}$ and a private part $W_{p,k}$. Then, the common messages $W_{c,1}, \ldots, W_{c,K}$ are combined into a single message $W_c$ and the $K+1$ resulting messages are encoded into independent Gaussian data streams $s_c, s_1, \ldots, s_K$, each having unit power. These symbols are combined with linear precoding into the transmit signal $\boldsymbol{x} = \boldsymbol{p}_c s_c + \sum_{k\in\mathcal{K}} \boldsymbol{p}_k s_k$. Due to the average transmit power constraint, the precoders need to satisfy $\|\boldsymbol{p}_c\|^2 + \sum_{k\in\mathcal{K}} \|\boldsymbol{p}_k\|^2 \le P$.

Each receiver $k \in \mathcal{K}$ uses SIC to first recover $s_c$ and then $s_k$ from its received signal $y_k$. In particular, $s_c$ is decoded first by treating interference from all other streams as noise. This allows user $k$ to recover its desired common part $W_{c,k}$. Then, $s_c$ is cancelled from the received signal and user $k$ proceeds to decode $s_k$ to recover the desired private part $W_{p,k}$. These two messages are combined to obtain $W_k$.

Given a precoding scheme $\boldsymbol{p}_c, \boldsymbol{p}_1, \ldots, \boldsymbol{p}_K$, asymptotic error free decoding of $W_c$ and $W_{p,k}$ is possible if the rates of these messages satisfy

$$R_c \le \log(1+\gamma_{c,k}), \forall k \in \mathcal{K}, \quad R_{p,k} \le \log(1+\gamma_{p,k}) \tag{1}$$

with signal to interference plus noise ratios (SINRs)

$$\gamma_{c,k} = \frac{|\boldsymbol{h}_k^H \boldsymbol{p}_c|^2}{\sum\limits_{j\in\mathcal{K}} |\boldsymbol{h}_k^H \boldsymbol{p}_j|^2 + 1}, \quad \gamma_{p,k} = \frac{|\boldsymbol{h}_k^H \boldsymbol{p}_k|^2}{\sum\limits_{j\in\mathcal{K}\setminus k} |\boldsymbol{h}_k^H \boldsymbol{p}_j|^2 + 1}. \tag{2}$$

The rate $R_c$ is shared across the users, where user $k$ is allocated a portion $C_k$ corresponding to the rate of $W_{c,k}$, such that $\sum_{k\in\mathcal{K}} C_k = R_c$. The total rate of user $k$ is $R_k = C_k + R_{p,k}$.

Observe that this system model can be interpreted as an instance of the joint multicast and unicast beamforming problem. It includes several notable special cases. With $\boldsymbol{p}_c = 0$, we obtain MU-LP and for $\boldsymbol{p}_k = 0$, $k \in \mathcal{K}$, it is the multicast beamforming problem. It also includes 2-user NOMA [6].

### A. Problem Statement

We consider WSR maximization under minimum rate Quality of Service (QoS) constraints, i.e.,

$$\max_{\substack{\boldsymbol{p}_1,\ldots,\boldsymbol{p}_K,\boldsymbol{p}_c,\\ \boldsymbol{C},\boldsymbol{R}_c,\boldsymbol{R}_p,\boldsymbol{\gamma}_c,\boldsymbol{\gamma}_p}} \quad \sum_{k\in\mathcal{K}} u_k (C_k + R_{p,k}) \tag{3a}$$

$$\text{s.t.} \quad R_c, R_{p,k}, \gamma_{c,k} \text{ and } \gamma_{p,k} \text{ as in (1)–(2)} \tag{3b}$$

$$\sum_{k'\in\mathcal{K}} C_{k'} \le R_c \tag{3c}$$

$$C_k \ge \max\left\{0,\, R_k^{th} - R_{p,k}\right\},\ k \in \mathcal{K} \tag{3d}$$

$$\|\boldsymbol{p}_c\|^2 + \sum_{k\in\mathcal{K}} \|\boldsymbol{p}_k\|^2 \le P \tag{3e}$$

with nonnegative weight vector $\boldsymbol{u} = [u_1, \ldots, u_K]^T \neq \boldsymbol{0}$, and EE maximization

$$\max_{\substack{\boldsymbol{p}_1,\ldots,\boldsymbol{p}_K,\boldsymbol{p}_c,\\ \boldsymbol{C},R_c,\boldsymbol{R}_p,\boldsymbol{\gamma}_c,\boldsymbol{\gamma}_p}} \quad \frac{\sum_{k\in\mathcal{K}} C_k + R_{p,k}}{\mu\left(\|\boldsymbol{p}_c\|^2 + \sum_{k\in\mathcal{K}} \|\boldsymbol{p}_k\|^2\right) + P_c} \tag{4a}$$

$$\text{s.t.} \quad \text{(3b)–(3e)}, \tag{4b}$$

where $\mu \ge 0$ is the power amplifier inefficiency and $P_c > 0$ is the static circuit power consumption. For notational simplicity, we define $\boldsymbol{C} = [C_1, \ldots, C_k]^T$, $\boldsymbol{\gamma}_p = [\gamma_1, \ldots, \gamma_K]^T$, $\boldsymbol{\gamma}_c = [\gamma_{c,1}, \ldots, \gamma_{c,K}]^T$, $\boldsymbol{R}_p = [R_{p,1}, \ldots, R_{p,K}]^T$. Both problems can be combined into the equivalent optimization problem

$$\max_{\substack{\boldsymbol{p}_1,\ldots,\boldsymbol{p}_K,\\ \boldsymbol{p}_c,\boldsymbol{C},\boldsymbol{\gamma}_c,\boldsymbol{\gamma}_p}} \quad \frac{\sum_{k\in\mathcal{K}} u_k \left(C_k + \log(1+\gamma_{p,k})\right)}{\mu\left(\|\boldsymbol{p}_c\|^2 + \sum_{k\in\mathcal{K}} \|\boldsymbol{p}_k\|^2\right) + P_c} \tag{5a}$$

$$\text{s.t.} \quad \gamma_{c,k} \text{ and } \gamma_{p,k} \text{ as in (2)} \tag{5b}$$

$$\sum_{k'\in\mathcal{K}} C_{k'} \le \log(1+\gamma_{c,k}),\ \ k \in \mathcal{K} \tag{5c}$$

$$C_k \ge \max\left\{0, R_k^{th} - \log(1+\gamma_{p,k})\right\}, k \in \mathcal{K} \tag{5d}$$

$$\|\boldsymbol{p}_c\|^2 + \sum_{k\in\mathcal{K}} \|\boldsymbol{p}_k\|^2 \le P. \tag{5e}$$

Clearly, we obtain WSR maximization for $\mu = 0, P_c = 1$ and EE maximization for $\boldsymbol{u} = \boldsymbol{1}$.

A globally optimal solution of (5) can be obtained by solving

$$\max_{\substack{\boldsymbol{p}_c,\boldsymbol{p}_1,\ldots,\boldsymbol{p}_K,\\ \boldsymbol{C},\boldsymbol{\gamma}_p,s,\boldsymbol{d},\boldsymbol{e}}} \quad \frac{\sum_{k\in\mathcal{K}} u_k \left(C_k + \log(1+\gamma_{p,k})\right)}{\mu\left(\|\boldsymbol{p}_c\|^2 + \sum_{k\in\mathcal{K}} \|\boldsymbol{p}_k\|^2\right) + P_c} \tag{6a}$$

$$\text{s.t.} \quad \sqrt{\gamma_{p,k}}\left(\sum_{j\in\mathcal{K}\setminus k} |\boldsymbol{h}_k^H \boldsymbol{p}_j|^2 + 1\right)^{1/2} \le \boldsymbol{h}_k^H \boldsymbol{p}_k \tag{6b}$$

$$\sqrt{s}\left(\sum_{j\in\mathcal{K}} |\boldsymbol{h}_1^H \boldsymbol{p}_j|^2 + 1\right)^{1/2} \le \boldsymbol{h}_1^H \boldsymbol{p}_c \tag{6c}$$

$$\sqrt{s}\left(\sum_{j\in\mathcal{K}} |\boldsymbol{h}_k^H \boldsymbol{p}_j|^2 + 1\right)^{1/2} \le d_k, \forall k > 1 \tag{6d}$$

$$(e_k, d_k) \in \mathcal{C}, \forall k > 1 \tag{6e}$$

$$\Re\{\boldsymbol{h}_k^H \boldsymbol{p}_k\} \ge 0, \quad \Im\{\boldsymbol{h}_k^H \boldsymbol{p}_k\} = 0 \tag{6f}$$

$$\Re\{\boldsymbol{h}_1^H \boldsymbol{p}_c\} \ge 0, \quad \Im\{\boldsymbol{h}_1^H \boldsymbol{p}_c\} = 0 \tag{6g}$$

$$\forall k > 1 : d_k \ge 0,\ e_k = \boldsymbol{h}_k^H \boldsymbol{p}_c \tag{6h}$$

$$\sum_{k\in\mathcal{K}} C_k \le \log(1+s) \tag{6i}$$

$$\text{(5d) and (5e)} \tag{6j}$$

with

$$(e, d) \in \mathcal{C} = \{e \in \mathbb{C}, d \in \mathbb{R} : d \le |e|\} \tag{7}$$

instead of (5). A crucial observation is that this problem is a second-order cone program (SOCP) for fixed $s$, $\boldsymbol{\gamma}_p$ except for constraint (6h). Hence, the nonconvexity of (5) is only due to the SINR expressions and not due to the beamforming vectors. We will exploit this partial convexity in the final algorithm to limit the numerical complexity.

Another equivalent variant of (5) that is interesting in its own right is

$$\max_{\substack{\boldsymbol{p}_c,\boldsymbol{p}_1,\ldots,\boldsymbol{p}_K,\\ \boldsymbol{C},\boldsymbol{\gamma}_p,s}} \quad \frac{\sum_{k\in\mathcal{K}} u_k \left(C_k + \log(1+\gamma_{p,k})\right)}{\mu\left(\|\boldsymbol{p}_c\|^2 + \sum_{k\in\mathcal{K}} \|\boldsymbol{p}_k\|^2\right) + P_c} \tag{8a}$$

$$\text{s.t.} \quad \gamma_{p,k} \leq \frac{|\boldsymbol{h}_k^H \boldsymbol{p}_k|^2}{\sum_{j\in\mathcal{K}\setminus k} |\boldsymbol{h}_k^H \boldsymbol{p}_j|^2 + 1} \tag{8b}$$

$$s \leq \frac{|\boldsymbol{h}_k^H \boldsymbol{p}_c|^2}{\sum_{j\in\mathcal{K}} |\boldsymbol{h}_k^H \boldsymbol{p}_j|^2 + 1}, \ k \in \mathcal{K} \tag{8c}$$

$$(5\text{d}), (5\text{e}) \text{ and } (6\text{i}). \tag{8d}$$

This problem is obtained as a side product when showing the equivalence of (5) and (6), which is established next.

*Proposition 1:* Let $\boldsymbol{x}^\star = (\boldsymbol{p}_c^\star, \boldsymbol{p}_1^\star, \ldots, \boldsymbol{p}_K^\star, \boldsymbol{C}^\star, \boldsymbol{\gamma}_p^\star)$. A point $(\boldsymbol{x}^\star, s^\star)$ solves (8) if $(\boldsymbol{x}^\star, \boldsymbol{\gamma}_c^\star)$ solves (5) and $s^\star = \min_k \gamma_{c,k}^\star$. Conversely, the point $(\boldsymbol{x}^\star, \boldsymbol{\gamma}_c^\star)$ solves (5) if $(\boldsymbol{x}^\star, s^\star)$ solves (8) and $\gamma_{c,k}^\star = \frac{|\boldsymbol{h}_k^H \boldsymbol{p}_c^\star|^2}{\sum_{j\in\mathcal{K}} |\boldsymbol{h}_k^H \boldsymbol{p}_j^\star|^2 + 1}$ for all $k \in \mathcal{K}$. Moreover, if $(\boldsymbol{x}^\star, s^\star, \boldsymbol{d}^\star, \boldsymbol{e}^\star)$ solves (6), then $(\boldsymbol{x}^\star, s^\star)$ solves (8) and $(\boldsymbol{x}^\star, \boldsymbol{\gamma}_c^\star)$ solves (5), where $s^\star$ and $\boldsymbol{\gamma}_c^\star$ are as before.

*Proof:* See Appendix A. ∎

*Corollary 1:* Problems (5), (6), (8) have the same optimal value.

In the next section, we introduce some mathematical preliminaries before we develop a solution algorithm for (6) in Section IV.

## III. Mathematical Background

Problem (6) is an NP-hard nonconvex optimization problem. To see this, consider problem (8) for $\mu = 0$, $u_k = 1$, and fix all variables except $\boldsymbol{p}_c$, $\boldsymbol{C}$ and $s$. Then, it is equivalent to

$$\max_{\boldsymbol{p}_c} \min_k \{|\tilde{\boldsymbol{h}}_k^H \boldsymbol{p}_c|^2\} \quad \text{s.t.} \quad \|\boldsymbol{p}_c\|^2 \leq \tilde{P}. \tag{9}$$

This is known as multicast beamforming and shown to be NP-hard in [37]. Our solution approach for this part of the problem relies on the so-called "argument cuts" proposed in [41]. The introduction of the auxiliary variables $d_k$ and $e_k$ in (6) is motivated by this approach and collects most of the nonconvexity due to $\boldsymbol{p}_c$ in (6e). The idea is to add a box constraint on $\arg(e_k)$ to $\mathcal{C}$ and then optimize over its convex envelope to obtain a bound on (6) suitable for a BB procedure.

The remaining nonconvexity in (6) stems from $\gamma_{p,k}$ and $s$ in (6b)–(6d). Previous global optimization algorithms for such problems rely on BB procedures with SOCP bounding [39], [40], [42], [43]. However, this leads to an infinite algorithm where the convergence to the global optimal solution cannot be guaranteed in a finite number of iterations.[1] This is because the difficulty in solving (6) is due to the feasible set, while BB works best if the nonconvexity is mostly due to the objective. Please refer to [46]–[48] for a detailed discussion of this topic. For the solution of (6), finite convergence can be obtained by adding a line search procedure to every iteration of the BB procedure that recovers a feasible point and requires the solution of several SOC feasibility problems [39, Alg. 3]. Hence, finite convergence in BB procedures comes at the cost of increased computational complexity. Moreover, the auxiliary SOCP that is solved in every iteration of the BB procedure is numerically challenging as the feasible set can become very small. This leads to numerical problems even with commercial state-of-the-art solvers like Mosek [49]. A computationally more tractable modification is proposed in [39, §2.2.2] that comes at the price of much harder feasible point acquisition in the BB procedure.

Instead, we design an algorithm based on the SIT scheme [47], [48], [50], [51] and combine it with a branch reduce and bound (BRB) procedure. The resulting algorithm is numerically stable, has proven finite convergence, solves EE maximization and is the first global optimization algorithm specifically designed for RSMA. Practically, it outperforms algorithms for similar problems as will be verified in Section V. To better illustrate the core principles of SIT, we first consider the following general optimization problem

$$\max_{(\boldsymbol{x},\boldsymbol{\xi})\in\mathcal{D}} \quad f(\boldsymbol{x},\boldsymbol{\xi}) \quad \text{s.t.} \quad g_i(\boldsymbol{x},\boldsymbol{\xi}) \leq 0, \ i = 1,\ldots,n \tag{10}$$

with continuous, real-valued functions $f, g_1, \ldots, g_n$ and nonempty feasible set. Further, assume that $f$ is concave,[2] $g_1, \ldots, g_n$ are convex in $\boldsymbol{\xi}$ for fixed $\boldsymbol{x}$, and $\mathcal{D}$ is a closed convex set. Depending on the structure of $g_1, \ldots, g_n$ in $\boldsymbol{x}$ this problem might be quite hard to solve for BB methods [47], [52].[3] By exchanging the objective and constraints of (10), we obtain the so-called SIT dual

$$\min_{(\boldsymbol{x},\boldsymbol{\xi})\in\mathcal{D}} \quad \max_i \{g_i(\boldsymbol{x},\boldsymbol{\xi})\} \quad \text{s.t.} \quad f(\boldsymbol{x},\boldsymbol{\xi}) \geq \delta. \tag{11}$$

Observe that the optimal value of (10) is greater than or equal to $\delta$ if the optimal value of (11) is less than or equal to zero. Conversely, if the optimal value of (11) is greater than zero, the optimal value of (10) is less than $\delta$. Hence, the optimal solution of (10) can be obtained by solving a sequence of (11) with increasing $\delta$. Since the feasible set of (11) is closed and convex, it can be solved much easier by BB than (10).

Obtaining the exact optimal solution to continuous real-valued optimization problems is often computationally infeasible, even for linear or convex problems. A widely employed practice is to accept any feasible point with objective value within a prescribed tolerance $\eta$ of the exact optimal value as a solution. That is, a point $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{\xi}})$ is called an $\eta$-optimal solution of (10) if, for all feasible points $(\boldsymbol{x}, \boldsymbol{\xi})$,

$$f(\bar{\boldsymbol{x}}, \bar{\boldsymbol{\xi}}) \geq f(\boldsymbol{x}, \boldsymbol{\xi}) - \eta. \tag{12}$$

Likewise, the constraints in (10) can be hard to satisfy numerically. The most common approach is to relax them

[1] While this often does not lead to problems in practice, slower convergence might be observed in infinite algorithms. In addition, finiteness is an important theoretical aspect that differentiates a "computational method" from an "algorithm" [45].

[2] Although this assumption does not hold for (6), it will be established later that the SIT approach is still applicable. This is because the sole purpose of this convexity assumption is to obtain a convex feasible set in (11).

[3] This is also true for outer approximation methods like the Polyblock algorithm [52].

by $\varepsilon$. However, for nonconvex feasible sets this can lead to completely wrong solutions [47], [50], [52]. Instead, for the SIT scheme, the constraints are tightened by $\varepsilon$, i.e., the problem to be solved is

$$\max_{(\boldsymbol{x},\boldsymbol{\xi})\in\mathcal{D}} \; f(\boldsymbol{x},\boldsymbol{\xi}) \quad \text{s.t.} \quad g_i(\boldsymbol{x},\boldsymbol{\xi}) \leq -\varepsilon,\ i=1,\ldots,n \tag{13}$$

for some $\varepsilon > 0$. Any point in this feasible set is denoted as $\varepsilon$-essential feasible and a solution of this problem satisfying (12) is called essential $(\varepsilon,\eta)$-optimal solution of (10). This constraint tightening removes numerically instable points from the feasible set and is necessary to ensure finite convergence of the SIT scheme.

The outlined duality between (10) and (11) is formalized in the following lemma.

*Lemma 1:* For every $\varepsilon > 0$, the $\varepsilon$-essential optimal value of (10) is less than $\delta$ if and only of the optimal value of (11) is greater than or equal to $-\varepsilon$.

*Proof:* Direct consequence of [50, Prop. 1]. ∎

We refer to (10) as the primal problem and (11) as the dual problem.[4]

### A. Successive Incumbent Transcending Algorithm

The discussion above leads to the SIT algorithm as stated in Algorithm 1. The core problem is Step 1, which is implemented by solving (11) with a modified rectangular BB procedure. Such a procedure has exponential computational complexity in the number of optimization variables. Since (11) is a convex optimization problem for fixed $\boldsymbol{x}$, the SIT BB procedure should only operate on the nonconvex variables $\boldsymbol{x}$ and employ a convex solver for $\boldsymbol{\xi}$ to limit the computational complexity.

**Algorithm 1** SIT Algorithm [52, §7.5.1].

- **Step 0** Initialize $(\bar{\boldsymbol{x}},\bar{\boldsymbol{\xi}})$ with the best known nonisolated feasible solution and set $\delta = f(\bar{\boldsymbol{x}},\bar{\boldsymbol{\xi}}) + \eta$; otherwise do not set $(\bar{\boldsymbol{x}},\bar{\boldsymbol{\xi}})$ and choose $\delta \leq f(\boldsymbol{x},\boldsymbol{\xi})\ \forall(\boldsymbol{x},\boldsymbol{\xi}) \in \mathcal{D}$.
- **Step 1** Check if (10) has a nonisolated feasible solution $(\boldsymbol{x},\boldsymbol{\xi})$ satisfying $f(\boldsymbol{x},\boldsymbol{\xi}) \geq \delta$; otherwise, establish that no such $\varepsilon$-essential feasible $(\boldsymbol{x},\boldsymbol{\xi})$ exists and go to Step 3.
- **Step 2** Update $(\bar{\boldsymbol{x}},\bar{\boldsymbol{\xi}}) \leftarrow (\boldsymbol{x},\boldsymbol{\xi})$ and $\delta \leftarrow f(\bar{\boldsymbol{x}},\bar{\boldsymbol{\xi}})+\eta$. Go to Step 1.
- **Step 3** Terminate: If $(\bar{\boldsymbol{x}},\bar{\boldsymbol{\xi}})$ is set, it is an essential $(\varepsilon,\eta)$-optimal solution; else Problem (10) is $\varepsilon$-essential infeasible.

The general idea of BB is to relax the feasible set and then subsequently partition this relaxed set in such a way that upper and lower bounds on the objective value in each partition can be computed efficiently. As the partition is successively refined, these bounds approach each other until the optimal value is found. For a rectangular BB procedure, the feasible set is relaxed into an initial box

$$\mathcal{M}_0 = [\boldsymbol{r}^0,\boldsymbol{s}^0] = \{\boldsymbol{x} : r_i^0 \leq x_i \leq s_i^0\} \tag{14}$$

satisfying $\mathcal{M}_0 \supseteq \operatorname{proj}_{\boldsymbol{x}} \mathcal{D}$. Further, a bounding function $\beta(\mathcal{M})$, $\mathcal{M} \subseteq \mathcal{M}_0$, with $\beta(\mathcal{M}) = \infty$ if $\operatorname{proj}_{\boldsymbol{x}} \mathcal{F} \cap \mathcal{M} = \emptyset$ and

$$\beta(\mathcal{M}) \leq \min_{(\boldsymbol{x},\boldsymbol{\xi})\in\mathcal{F},\boldsymbol{x}\in\mathcal{M}} \max_i \{g_i(\boldsymbol{x},\boldsymbol{\xi})\}, \tag{15}$$

[4]In this paper, the concept of duality is used with respect to the SIT dual as discussed in this section and not in terms of Lagrange duality theory.

otherwise is required, where $\mathcal{F} = \{(\boldsymbol{x},\boldsymbol{\xi}) \in \mathcal{D} : f(\boldsymbol{x},\boldsymbol{\xi}) \geq \delta\}$ is the feasible set of (11). The algorithm subsequently partitions the relaxed feasible set $\mathcal{M}_0$ into smaller boxes and stores the current partition of $\mathcal{M}_0$ in a set $\mathscr{R}_k$. In iteration $k$, the algorithm uses best-first selection to determine the next branch, i.e.,

$$\mathcal{M}_k \in \arg\min\{\beta(\mathcal{M}) \,|\, \mathcal{M} \in \mathscr{R}_k\}, \tag{16}$$

and then replaces $\mathcal{M}_k = [\boldsymbol{r}^k,\boldsymbol{s}^k]$ by two new subrectangles

$$\mathcal{M}^- = \{\boldsymbol{x} : r_j \leq x_j \leq v_j,\ r_i \leq x_i \leq s_i\ (i \neq j)\} \tag{17a}$$
$$\mathcal{M}^+ = \{\boldsymbol{x} : v_j \leq x_j \leq s_j,\ r_i \leq x_i \leq s_i\ (i \neq j)\} \tag{17b}$$

with $\boldsymbol{v} = \frac{1}{2}(\boldsymbol{s}+\boldsymbol{r})$ and $j \in \arg\max_j s_j - r_j$. For each of these new boxes, a lower bound on the objective value is computed using the bounding function $\beta(\mathcal{M})$. To ensure convergence, the bounding needs to be consistent with branching, i.e., $\beta(\mathcal{M})$ has to satisfy

$$\beta(\mathcal{M}) - \min_{\substack{(\boldsymbol{x},\boldsymbol{\xi})\in\mathcal{F},\\ \boldsymbol{x}\in\mathcal{M}}} \max_i \{g_i(\boldsymbol{x},\boldsymbol{\xi})\} \to 0 \ \text{ as } \max_{\boldsymbol{x},\boldsymbol{y}\in\mathcal{M}} \|\boldsymbol{x}-\boldsymbol{y}\| \to 0, \tag{18}$$

and a dual feasible point $\boldsymbol{x}^k \in \operatorname{proj}_{\boldsymbol{x}} \mathcal{F} \cap \mathcal{M}_k$ is required if $\beta(\mathcal{M}_k) < \infty$. Suitable pruning and termination rules that ensure convergence can be obtained from the following lemma that is adapted from [52, Prop. 7.14] and [48, Prop. 5.9].

*Lemma 2:* Let $\varepsilon > 0$ be given and define $g(\boldsymbol{x},\boldsymbol{\xi}) = \max_i\{g_i(\boldsymbol{x},\boldsymbol{\xi})\}$. Let $\beta(\mathcal{M})$ satisfy (15) and (18) and $\mathcal{M}_k$ be as in (16). Then, as $\max_{\boldsymbol{x},\boldsymbol{y}\in\mathcal{M}_k} \|\boldsymbol{x}-\boldsymbol{y}\| \to 0$ for $k \to \infty$, either $g(\boldsymbol{x}^k,\boldsymbol{\xi}^*) < 0$ for some $k$ and $(\boldsymbol{x}^k,\boldsymbol{\xi}^*) \in \mathcal{F}$ or $\beta(\mathcal{M}_k) > -\varepsilon$ for some $k$. In the former case, $(\boldsymbol{x}^k,\boldsymbol{\xi}^*)$ is a nonisolated feasible solution of (10) satisfying $f(\boldsymbol{x}^k,\boldsymbol{\xi}^*) \geq \delta$. In the latter case, no $\varepsilon$-essential feasible solution $(\boldsymbol{x},\boldsymbol{\xi})$ of (10) exists such that $f(\boldsymbol{x},\boldsymbol{\xi}) \geq \delta$.

*Proof:* Please refer to [48, Prop. 5.9]. ∎

This suggests a BB procedure with pruning criterion $\beta(\mathcal{M}) < -\varepsilon$ and termination criterion

$$0 > \min_{\boldsymbol{\xi}} \; g(\boldsymbol{x}^k,\boldsymbol{\xi}) \ \text{ s.t. } \ (\boldsymbol{x}^k,\boldsymbol{\xi}) \in \mathcal{F}. \tag{19}$$

In the following section, we apply this approach to find the solution of (6) and explicitly incorporate the outlined BB procedure into Algorithm 1.

## IV. Globally Optimal Beamforming

We design a globally optimal solution algorithm for (6) based on the fundamentals in the previous section. There, we have seen that the SIT algorithm requires a bounding function, a feasible point in each iteration and an initial box $\mathcal{M}_0$. These aspects will be discussed after identifying and discussing the SIT dual. At the end of this section, we state the complete algorithm and establish its convergence. We also derive a reduction procedure in Section IV-D that is essential for practical convergence, although it is not strictly necessary from a theoretical perspective.

The SIT dual should contain all of the problem's nonconvexity in the objective function. Following the discussion in

Section II-A, the nonconvexity in (6) is due to (6b)–(6e) and we obtain the SIT dual as

$$\min_{\substack{\boldsymbol{p}_c,\boldsymbol{p}_1,\ldots,\boldsymbol{p}_K,\\ \boldsymbol{C},\boldsymbol{\gamma}_p,s,\boldsymbol{d},\boldsymbol{e}}} \max\Bigg[\sqrt{s}\left(\sum\nolimits_{j\in\mathcal{K}}|\boldsymbol{h}_1^H\boldsymbol{p}_j|^2+1\right)^{1/2}-\boldsymbol{h}_1^H\boldsymbol{p}_c,$$
$$\max_{k>1}\left\{\sqrt{s}\left(\sum\nolimits_{j\in\mathcal{K}}|\boldsymbol{h}_k^H\boldsymbol{p}_j|^2+1\right)^{1/2}-d_k\right\},$$
$$\max_{k\in\mathcal{K}}\left\{\sqrt{\gamma_{p,k}}\left(\sum\nolimits_{j\in\mathcal{K}\setminus k}|\boldsymbol{h}_k^H\boldsymbol{p}_j|^2+1\right)^{1/2}-\boldsymbol{h}_k^H\boldsymbol{p}_k\right\},$$
$$\max_{k>1}\left\{d_k-|e_k|\right\}\Bigg] \tag{20a}$$
$$\text{s.t.}\quad \frac{\sum_{k\in\mathcal{K}}u_k\left(C_k+\log(1+\gamma_{p,k})\right)}{\mu\left(\|\boldsymbol{p}_c\|^2+\sum_{k\in\mathcal{K}}\|\boldsymbol{p}_k\|^2\right)+P_c}\geq\delta \tag{20b}$$
$$\text{(5d), (5e), (6f)–(6i).} \tag{20c}$$

Observe that (20b) is equivalent to the SOC

$$\sum_{k\in\mathcal{K}}u_k\left(C_k+\log(1+\gamma_{p,k})\right)$$
$$\geq\delta\left(\mu\left(\|\boldsymbol{p}_c\|^2+\sum_{k\in\mathcal{K}}\|\boldsymbol{p}_k\|^2\right)+P_c\right) \tag{21}$$

since the denominator in (20b) is positive. First, smoothen the objective by using the epigraph form with auxiliary variable $t$, and successively convert the pointwise maximum expressions to smooth constraints. Then, the new constraints $d_k-|e_k|\leq t$, for $k>1$, are equivalent to $(e_k,d_k-t)\in\mathcal{C}$. Introducing auxiliary variables $\alpha_k\in[0,2\pi]$ and constraints $\alpha_k=\angle e_k$ for $k>1$ leads to the equivalent optimization problem

$$\min_{\substack{\boldsymbol{p}_1,\ldots,\boldsymbol{p}_k,\\ \boldsymbol{p}_c,\boldsymbol{C},\boldsymbol{\gamma}_p,\\ s,\boldsymbol{d},\boldsymbol{e},t,\boldsymbol{\alpha}}} t \tag{22a}$$
$$\text{s.t.}\quad \sqrt{s}\big(\sum_{j\in\mathcal{K}}|\boldsymbol{h}_1^H\boldsymbol{p}_j|^2+1\big)^{1/2}-\boldsymbol{h}_1^H\boldsymbol{p}_c\leq t \tag{22b}$$
$$\sqrt{s}\big(\sum_{j\in\mathcal{K}}|\boldsymbol{h}_k^H\boldsymbol{p}_j|^2+1\big)^{1/2}-d_k\leq t,\ k>1 \tag{22c}$$
$$\sqrt{\gamma_{p,k}}\big(\sum_{j\in\mathcal{K}\setminus k}|\boldsymbol{h}_k^H\boldsymbol{p}_j|^2+1\big)^{1/2}-\boldsymbol{h}_k^H\boldsymbol{p}_k,\ k\in\mathcal{K} \tag{22d}$$
$$(e_k,d_k-t,\alpha_k)\in\tilde{\mathcal{C}},\ k>1 \tag{22e}$$
$$\text{(5d), (5e), (6f)–(6i), (21)} \tag{22f}$$

with

$$\tilde{\mathcal{C}}=\{e\in\mathbb{C},d\in\mathbb{R},\alpha\in\mathbb{R}:d\leq|e|,\ \angle e=\alpha\}. \tag{23}$$

Note that this is a convex optimization problem for fixed $(\boldsymbol{\gamma}_p,s,\boldsymbol{\alpha})$. Hence, we design the BRB procedure to operate on these variables.

Relating this to the previous section, we can identify the nonconvex variables $\boldsymbol{x}=(\boldsymbol{\gamma}_p,s,\boldsymbol{\alpha})$, the convex variables $\boldsymbol{\xi}=(\boldsymbol{p}_c,\boldsymbol{p}_1,\ldots,\boldsymbol{p}_K,\boldsymbol{C},\boldsymbol{\gamma}_p,\boldsymbol{d},\boldsymbol{e})$, the dual feasible set $\mathcal{F}$ as

$$\Big\{\boldsymbol{\gamma}_p,s,\boldsymbol{\alpha},\boldsymbol{p}_c,\boldsymbol{p}_1,\ldots,\boldsymbol{p}_K,\boldsymbol{C},\boldsymbol{\gamma}_p,\boldsymbol{d},\boldsymbol{e}:\angle\boldsymbol{e}=\boldsymbol{\alpha},$$
$$\boldsymbol{\alpha}\in[0,2\pi]^{K-1},\ \text{and (5d), (5e), (6f)–(6i), (21)}\Big\}. \tag{24}$$

and the dual objective $\tilde{g}(\boldsymbol{x},\boldsymbol{\xi})=\max_i\{g_i(\boldsymbol{x},\boldsymbol{\xi})\}$ as the function

$$\tilde{g}:(\boldsymbol{\gamma}_p,s,\boldsymbol{\alpha})\mapsto\min_{\substack{\boldsymbol{p}_1,\ldots,\boldsymbol{p}_K,\\ \boldsymbol{p}_c,\boldsymbol{C},\boldsymbol{d},\boldsymbol{e},t}} t\quad\text{s.t.}\quad\text{(22b)–(22f).} \tag{25}$$

### A. Bounding Procedure

A bounding function $\beta(\mathcal{M})$ that satisfies (18) is required. We obtain it by adding suitable box constraints to (22) and then relaxing it adequately. First, observe that the objective of (20) is increasing in $(\boldsymbol{\gamma}_p,s)$. Hence, a lower bound on $[\underline{\boldsymbol{\gamma}}_p,\bar{\boldsymbol{\gamma}}_p]\times[\underline{s},\bar{s}]$ is obtained by setting $\boldsymbol{\gamma}_p=\underline{\boldsymbol{\gamma}}_p$ and $s=\underline{s}$. This leaves the nonconvexity in $\tilde{\mathcal{C}}$. Consistent bounding over this set is achieved by using argument cuts [41], i.e., we introduce box constraints on $\boldsymbol{\alpha}$, i.e., $\boldsymbol{\alpha}\in[\underline{\boldsymbol{\alpha}},\bar{\boldsymbol{\alpha}}]$, and replace $\tilde{\mathcal{C}}$ by its convex envelope. For $\bar{\alpha}_k-\underline{\alpha}_k\leq\pi$, this envelope is

$$\sin(\underline{\alpha}_k)\Re\{e_k\}-\cos(\underline{\alpha}_k)\Im\{e_k\}\leq 0 \tag{26a}$$
$$\sin(\bar{\alpha}_k)\Re\{e_k\}-\cos(\bar{\alpha}_k)\Im\{e_k\}\geq 0 \tag{26b}$$
$$a_k\Re\{e_k\}+b_k\Im\{e_k\}\geq(d_k-t)(a_k^2+b_k^2) \tag{26c}$$

and $(e_k,d_k)\in\mathbb{C}\times\mathbb{R}$ otherwise [41, Prop. 1], where $a_k=\frac{1}{2}\left(\cos(\underline{\alpha}_k)+\cos(\bar{\alpha}_k)\right)$, and $b_k=\frac{1}{2}\left(\sin(\underline{\alpha}_k)+\sin(\bar{\alpha}_k)\right)$. Then, the bounding problem for a box $\mathcal{M}=[\underline{\boldsymbol{\gamma}}_p,\bar{\boldsymbol{\gamma}}_p]\times[\underline{s},\bar{s}]\times[\underline{\boldsymbol{\alpha}},\bar{\boldsymbol{\alpha}}]$ is

$$\min_{\substack{\boldsymbol{p}_c,\boldsymbol{p}_1,\ldots,\boldsymbol{p}_K,\\ \boldsymbol{C},\boldsymbol{\gamma}_p,s,\boldsymbol{d},\boldsymbol{e},t}} t \tag{27a}$$
$$\text{s.t.}\quad \underline{\sqrt{s}}\big(\sum_{j\in\mathcal{K}}|\boldsymbol{h}_1^H\boldsymbol{p}_j|^2+1\big)^{1/2}-\boldsymbol{h}_1^H\boldsymbol{p}_c\leq t \tag{27b}$$
$$\underline{\sqrt{s}}\big(\sum_{j\in\mathcal{K}}|\boldsymbol{h}_k^H\boldsymbol{p}_j|^2+1\big)^{1/2}-d_k\leq t,\ k>1 \tag{27c}$$
$$\underline{\sqrt{\gamma_{p,k}}}\big(\sum_{j\in\mathcal{K}\setminus k}|\boldsymbol{h}_k^H\boldsymbol{p}_j|^2+1\big)^{1/2}-\boldsymbol{h}_k^H\boldsymbol{p}_k,\ k\in\mathcal{K} \tag{27d}$$
$$\text{(26a)–(26c)},\ k\in\mathcal{I}_{\mathcal{M}} \tag{27e}$$
$$\boldsymbol{\gamma}_p\in[\underline{\boldsymbol{\gamma}}_p,\bar{\boldsymbol{\gamma}}_p],\quad s\in[\underline{s},\bar{s}], \tag{27f}$$
$$\text{(5d), (5e), (6f)–(6i), (21),} \tag{27g}$$

where, with a slight abuse of notation,

$$\mathcal{I}_{\mathcal{M}}=\Big\{k\in\mathcal{K}:k>1\wedge\max_{\alpha,\bar{\alpha}\in\mathcal{M}}|\bar{\alpha}_k-\underline{\alpha}_k|\leq\pi\Big\}. \tag{28}$$

Define the bounding function $\beta(\mathcal{M})$ such that it takes the optimal value of (27) if (27) is feasible and $\infty$ otherwise. This is a suitable bounding function to solve (20) with a BB procedure over the nonconvex variables $\boldsymbol{\gamma}_p,s,\boldsymbol{\alpha}$.

*Lemma 3:* The bounding function $\beta(\mathcal{M})$ computed from (27) is consistent with respect to (22), i.e., it satisfies (18) with $g_i(\boldsymbol{x},\boldsymbol{\xi})$ and $\mathcal{F}$ as identified in (24) and (25).

*Proof:* We need to show that $\beta(\mathcal{M})$ asymptotically approaches the optimal value of (27) on $\mathcal{M}$ as $\mathcal{M}$ shrinks to a singleton, i.e., $\mathcal{M}\to\{\boldsymbol{z}^*\}$ with $\boldsymbol{z}^*=(\boldsymbol{\gamma}_p^*,s,\boldsymbol{\alpha}^*)$. Observe that these problems only differ in the constraints (22b)–(22e) and (27b)–(27e). Asymptotically, (22b)–(22d) and (27b)–(27d) are equivalent since $\underline{\boldsymbol{\gamma}}_p,\boldsymbol{\gamma}_p\to\boldsymbol{\gamma}_p^*$ and $\underline{s},s\to s^*$.

For the remaining constraints, note that $\mathcal{I}_{\mathcal{M}} = \{k \in \mathcal{K} : k > 1\}$ as $\underline{\boldsymbol{\alpha}}, \bar{\boldsymbol{\alpha}} \to \boldsymbol{\alpha}^*$. Further, (26a) and (26b) asymptotically evaluate to

$$\sin(\alpha_k^*)\Re\{e_k\} = \cos(\alpha_k^*)\Im\{e_k\} \tag{29}$$

and (26c) to

$$\cos(\alpha_k^*)\Re\{e_k\} + \sin(\alpha_k^*)\Im\{e_k\} \geq d_k - t. \tag{30}$$

Recall that constraint (22e) is $d_k - t \leq |e_k|$ and $\angle e_k = \alpha_k^*$. The second equation is equivalent to

$$|e_k| = \frac{\Re\{e_k\}}{\cos(\alpha_k^*)} = \frac{\Im\{e_k\}}{\sin(\alpha_k^*)} \tag{31}$$

and, hence, asymptotically the same as (29). Finally, with $\Re\{e_k\} = |e_k|\cos(\alpha_k^*)$ and $\Im\{e_k\} = |e_k|\cos(\alpha_k^*)$, the left-hand side of (30) is

$$\cos^2(\alpha_k^*)|e_k| + \sin^2(\alpha_k^*)|e_k| = |e_k|. \tag{32}$$

This establishes the lemma. ∎

Observe that problem (27) depends on $\boldsymbol{\gamma}_p$ and $s$ only through constraints (5d), (6i), (21), (27f). These are (convex) exponential cone constraints that can be transformed into affine functions of $(\boldsymbol{\gamma}_p, s)$ by substituting $s' = \log(1+s)$ and $\gamma'_{p,k} = \log(1+\gamma_{p,k})$. This leads to an equivalent optimization problem with considerably reduced computational complexity. In particular, we can solve the following SOCP

$$\min_{\substack{\boldsymbol{p}_1,\dots,\boldsymbol{p}_K,\boldsymbol{p}_c,\\ \boldsymbol{C},\boldsymbol{\gamma}'_p,s',\boldsymbol{d},\boldsymbol{e},t}} t \tag{33a}$$

$$\text{s.t.} \quad C_k \geq \max\left\{0,\, R_k^{th} - \gamma'_{p,k}\right\},\ k \in \mathcal{K} \tag{33b}$$

$$\sum\nolimits_{k\in\mathcal{K}} C_k \leq s' \tag{33c}$$

$$\sum\nolimits_{k\in\mathcal{K}} u_k\left(C_k + \gamma'_{p,k}\right) \geq \delta\Big(\mu\Big(\|\boldsymbol{p}_c\|^2 + \sum_{k\in\mathcal{K}}\|\boldsymbol{p}_k\|^2\Big) + P_c\Big) \tag{33d}$$

$$\gamma'_{p,k} \in [\log(1+\underline{\gamma}_{p,k}), \log(1+\bar{\gamma}_{p,k})], k \in \mathcal{K} \tag{33e}$$

$$s' \in [\log(1+\underline{s}), \log(1+\bar{s})], \tag{33f}$$

$$\text{(5e), (6f)–(6h), (27b)–(27e)} \tag{33g}$$

instead of (27) to compute the bounding function $\beta(\mathcal{M})$.

### B. Feasible Point

In every iteration, a dual feasible point $\boldsymbol{x}^k$ is required that satisfies $\boldsymbol{x}^k \in \operatorname{proj}_{\boldsymbol{x}} \mathcal{F} \cap \mathcal{M}_k$ whenever $\operatorname{proj}_{\boldsymbol{x}} \mathcal{F} \cap \mathcal{M}_k \neq \emptyset$ (or, equivalently, $\beta(\mathcal{M}_k) < \infty$). If this point satisfies (19), it is nonisolated primal feasible according to Lemma 2 and can be used to update $\delta$ in Algorithm 1. Such a point $(\boldsymbol{\gamma}_p^k, s^k, \boldsymbol{\alpha}^k)$ can be obtained from the optimal solution $(\boldsymbol{\gamma}_p^\star, s^\star, \boldsymbol{e}^\star, \dots)$ of the bounding problem (33) as

$$\gamma_{p,i}^k = 2^{\gamma_{p,i}^\star} - 1, i \in \mathcal{K}, \qquad s^k = 2^{s^\star} - 1 \tag{34a}$$

and $\boldsymbol{\alpha}^k \in \operatorname{proj}_{\boldsymbol{\alpha}} \mathcal{M}_k = [\underline{\boldsymbol{\alpha}}^k, \bar{\boldsymbol{\alpha}}^k]$. At first glance, a sensible choice for $\boldsymbol{\alpha}$ seems to be $\alpha_i^k = \angle e_i^\star$. However, preliminary numerical experiments show that this point leads to very slow convergence. Much better results are obtained by using the corner point of $\operatorname{proj}_{\boldsymbol{\alpha}} \mathcal{M}^k$ closest to $\angle \boldsymbol{e}^\star$, i.e.,

$$\alpha_i^k = \operatorname*{arg\,min}_{\alpha\in\{\underline{\alpha}_i^k, \bar{\alpha}_i^k\}} |\alpha - \angle e_i^\star|. \tag{34b}$$

It is easily verified that this point satisfies $(\boldsymbol{\gamma}_p^k, s^k, \boldsymbol{\alpha}^k) \in \operatorname{proj}_{(\boldsymbol{\gamma}_p, s, \boldsymbol{\alpha})} \mathcal{F} \cap \mathcal{M}_k$. If further $\tilde{g}(\boldsymbol{\gamma}_p^k, s^k, \boldsymbol{\alpha}^k) \leq 0$, it is primal feasible and the solution of (25) achieves a primal objective value greater than or equal to $\delta$.

*Lemma 4:* Let $\boldsymbol{z}$ be a solution of (33) for some $\delta$. Obtain $\boldsymbol{x}^k = (\boldsymbol{\gamma}_p^k, s^k, \boldsymbol{\alpha}^k)$ from $\boldsymbol{z}$ as in (34). Compute $\tilde{g}(\boldsymbol{\gamma}_p^k, s^k, \boldsymbol{\alpha}^k)$ as in (25) and let $(t^\star, \boldsymbol{y}^\star) = (t^\star, \boldsymbol{p}_1^\star, \dots, \boldsymbol{p}_K^\star, \boldsymbol{p}_c^\star, \boldsymbol{C}^\star, \boldsymbol{d}^\star, \boldsymbol{e}^\star)$ be a solution of the accompanying optimization problem. If $t^\star \leq 0$, $(\boldsymbol{x}^k, \boldsymbol{y}^\star)$ is a primal feasible point with primal objective value greater than or equal to $\delta$. Then, $(\boldsymbol{y}^\star, \boldsymbol{\gamma}_p^\star, s^\star)$ with $\gamma_{p,k}^\star = \frac{|\boldsymbol{h}_k^H \boldsymbol{p}_k^\star|^2}{\sum_{j\in\mathcal{K}\setminus k} |\boldsymbol{h}_k^H \boldsymbol{p}_j^\star|^2 + 1}$ for all $k \in \mathcal{K}$ and $s^\star = \min_{k\in\mathcal{K}} \frac{|\boldsymbol{h}_k^H \boldsymbol{p}_c^\star|^2}{\sum_{j\in\mathcal{K}} |\boldsymbol{h}_k^H \boldsymbol{p}_j^\star|^2 + 1}$ is a feasible point of (6) and achieves a primal objective value greater than or equal to that of $(\boldsymbol{x}^k, \boldsymbol{y}^\star)$. The primal objective value can be further improved (while preserving primal feasibility) by updating $\boldsymbol{C}^\star$ to a solution of

$$\max_{\boldsymbol{C}} \quad \sum_{k\in\mathcal{K}} u_k C_k \tag{35a}$$

$$\text{s.t.} \quad \frac{\sum_{k\in\mathcal{K}} u_k\left(C_k + \log(1+\gamma_{p,k}^\star)\right)}{\mu\left(\|\boldsymbol{p}_c^\star\|^2 + \sum_{k\in\mathcal{K}}\|\boldsymbol{p}_k^\star\|^2\right) + P_c} \geq \delta \tag{35b}$$

$$\sum_{k\in\mathcal{K}} C_k \leq \log(1+s^\star) \tag{35c}$$

$$C_k \geq \max\left\{0,\, R_k^{th} - \log(1+\gamma_{p,k}^\star)\right\}, \forall k \in \mathcal{K}. \tag{35d}$$

*Proof:* Observe that $\boldsymbol{\gamma}_p^k$ and $s^k$ are such that the set defined by (5d), (5e), (6f)–(6i), (21) is nonempty. Further, if $t \leq 0$, every point satisfying (22b)–(22d) also meets (6b)–(6d). Since $\tilde{\mathcal{C}} \subseteq \mathcal{C}$, (22e) implies that $(e_k, d_k) \in \mathcal{C}$ if $t \leq 0$. Hence, $(\boldsymbol{y}^\star, \boldsymbol{\gamma}_p^k, s^k)$ is a feasible solution of (6) if $t^\star \leq 0$.

From the proof of Proposition 1, we know that every point that satisfies (6b)–(6h) also satisfies (8b) and (8c). This implies $\boldsymbol{\gamma}_p^k \leq \boldsymbol{\gamma}_p^\star$, $s^k \leq s^\star$ and, hence, $(\boldsymbol{\gamma}_p^\star, \boldsymbol{s}^\star)$ satisfies (5d), (6b)–(6i). Thus, $(\boldsymbol{\gamma}^\star, \boldsymbol{\gamma}_p^\star, s^\star)$ is a feasible point of (6).

Clearly, $(\boldsymbol{x}^k, \boldsymbol{y}^\star)$ satisfies (21). Hence,

$$\delta \leq \frac{\sum_{i\in\mathcal{K}} u_i(C_i^\star + \log(1+\gamma_{p,i}^k))}{\mu\left(\|\boldsymbol{p}_c^\star\|^2 + \sum_{i\in\mathcal{K}}\|\boldsymbol{p}_i^\star\|^2\right) + P_c} \leq \frac{\sum_{i\in\mathcal{K}} u_i(C_i^\star + \log(1+\gamma_{p,i}^\star))}{\mu\left(\|\boldsymbol{p}_c^\star\|^2 + \sum_{i\in\mathcal{K}}\|\boldsymbol{p}_i^\star\|^2\right) + P_c}. \tag{36}$$

Clearly, any $\boldsymbol{C}$ that is feasible in (35) is also feasible in (6) and maximizes (36) in $\boldsymbol{C}$ (with all other variables fixed). ∎

### C. Initial Box

The BB procedure requires an initial box $\mathcal{M}_0 = [\underline{\boldsymbol{\gamma}}_p^0, \bar{\boldsymbol{\gamma}}_p^0] \times [\underline{s}^0, \bar{s}^0] \times [\underline{\boldsymbol{\alpha}}^0, \bar{\boldsymbol{\alpha}}^0]$ that contains the nonconvex dimensions of the dual feasible set $\operatorname{proj}_{(\boldsymbol{\gamma}_p, s, \boldsymbol{\alpha})} \mathcal{F}$. As $\boldsymbol{\alpha}$ is already constrained by box constraints, we have $[\underline{\boldsymbol{\alpha}}^0, \bar{\boldsymbol{\alpha}}^0] = [0, 2\pi]^{K-1}$. For $\bar{\boldsymbol{\gamma}}_p^0$, observe that $\bar{\gamma}_{p,k}^0 \geq \max_{\boldsymbol{\gamma}_p, s, \boldsymbol{\alpha}\in\mathcal{F}} \gamma_{p,k}$ but also

$$\bar{\gamma}_{p,k}^0 \geq \max_{\boldsymbol{p}_c, \boldsymbol{p}_1, \dots, \boldsymbol{p}_K, \boldsymbol{C}, \gamma_c, \boldsymbol{\gamma}_p} \gamma_{p,k} \quad \text{s. t.} \quad \text{(5b)–(5e)}. \tag{37}$$

This nonconvex optimization problem can be relaxed to

$$\max_{\boldsymbol{p}_k} \ |\boldsymbol{h}_k^H \boldsymbol{p}_k|^2 \quad \text{s.t.} \quad \|\boldsymbol{p}_k\|^2 \le P. \tag{38}$$

The solution to (38) is $\boldsymbol{p}_k^\star = \sqrt{P}\frac{\boldsymbol{h}_k}{\|\boldsymbol{h}_k\|}$ [53, §5.3.2] and, hence, $\bar{\gamma}_{p,k}^0 = P\|\boldsymbol{h}_k\|^2$. Likewise, the upper bound $\bar{s}^0$ for $s$ needs to satisfy

$$\bar{s}^0 \ge \max_{\boldsymbol{p}_c} \ \min_{k\in\mathcal{K}} \ |\boldsymbol{h}_k^H \boldsymbol{p}_c|^2 \quad \text{s.t.} \quad \|\boldsymbol{p}_c\|^2 \le P. \tag{39}$$

This is an NP-hard optimization problem as discussed in Section III. Exchanging maximum and minimum leads to the relaxed problem

$$\bar{s}^0 = \min_{k\in\mathcal{K}} \ \max_{\|\boldsymbol{p}_c\|^2\le P} \ |\boldsymbol{h}_k^H \boldsymbol{p}_c|^2 = \min_{k\in\mathcal{K}} P\|\boldsymbol{h}_k\|^2. \tag{40}$$

An obvious lower bound on $\boldsymbol{\gamma}_p$ and $s$ is 0. For $\boldsymbol{\gamma}_p$, we can exploit the QoS constraints to obtain a possibly tighter initial box. Similar to the upper bound, $\underline{\gamma}_{p,k}$ needs to be less than or equal to the minimum of $\gamma_{p,k}$ over (5b)–(5e). The optimal solution to this problem either meets (5d) with equality or is zero. In the first case, this is equivalent to

$$\min_{\boldsymbol{p}_c,\boldsymbol{p}_1,\ldots,\boldsymbol{p}_K,\boldsymbol{C}} \quad 2^{R_k^{th}-C_k} - 1 \tag{41a}$$

$$\text{s.t.} \quad \sum_{k'\in\mathcal{K}} C_{k'} \le \log\left(1 + \min_{k\in\mathcal{K}} \frac{|\boldsymbol{h}_k^H \boldsymbol{p}_c|^2}{\sum\limits_{j\in\mathcal{K}} |\boldsymbol{h}_k^H \boldsymbol{p}_j|^2 + 1}\right) \tag{41b}$$

$$\boldsymbol{C} \ge \boldsymbol{0}, \quad \|\boldsymbol{p}_c\|^2 + \sum_{k\in\mathcal{K}} \|\boldsymbol{p}_k\|^2 \le P. \tag{41c}$$

Clearly, the optimal solution to this problem is equivalent to the solution of

$$\max_{\boldsymbol{p}_c,\boldsymbol{p}_1,\ldots,\boldsymbol{p}_K,\boldsymbol{C}} \quad \log\left(1 + \min_{k\in\mathcal{K}} \frac{|\boldsymbol{h}_k^H \boldsymbol{p}_c|^2}{\sum\limits_{j\in\mathcal{K}} |\boldsymbol{h}_k^H \boldsymbol{p}_j|^2 + 1}\right) \tag{42a}$$

$$\text{s.t.} \quad \|\boldsymbol{p}_c\|^2 + \sum_{k\in\mathcal{K}} \|\boldsymbol{p}_k\|^2 \le P. \tag{42b}$$

The optimal choice for $\boldsymbol{p}_1,\ldots,\boldsymbol{p}_K$ is $\boldsymbol{0}$. Optimizing over $\boldsymbol{p}_c$ is equivalent to (39). Thus, an upper bound to the optimal $C_k$ is $\log(1+\bar{s})$ and a lower bound on the optimal value of (41) is $2^{R_k^{th}-\log(1+\bar{s})} - 1$. Hence,

$$\underline{\gamma}_{p,k} = \max\left\{0, \frac{2^{R_k^{th}}}{1+\bar{s}} - 1\right\}. \tag{43}$$

### *D. Reduction Procedure*

The convergence criterion (18) implies that the quality of the bound $\beta(\mathcal{M})$ improves as the diameter of $\mathcal{M}$ shrinks. Since tighter bounds lead to faster convergence, it is beneficial to reduce the size of $\mathcal{M}$ prior to bounding if possible at reasonable computational cost. It is important that such a reduced box $\mathcal{M}' \subseteq \mathcal{M}$ still contains all solution candidates, i.e., $\mathcal{M} \cap \tilde{\mathcal{F}} = \mathcal{M}' \cap \tilde{\mathcal{F}}$ or, equivalently, $(\mathcal{M} \setminus \mathcal{M}') \cap \tilde{\mathcal{F}} = \emptyset$, where $\tilde{\mathcal{F}} = \mathrm{proj}_{(\boldsymbol{\gamma}_p,s,\boldsymbol{\alpha})} \mathcal{F}$.

A suitable reduction is derived in the lemma below. Preliminary numerical experiments have shown that this procedure is essential to ensure convergence within reasonable time.

*Lemma 5:* Let $\mathcal{M} = [\underline{\boldsymbol{\gamma}}_p, \bar{\boldsymbol{\gamma}}_p] \times [\underline{s}, \bar{s}] \times [\underline{\boldsymbol{\alpha}}, \bar{\boldsymbol{\alpha}}]$, $\mathcal{M}' = [\underline{\boldsymbol{\gamma}}'_p, \bar{\boldsymbol{\gamma}}'_p] \times [\underline{s}', \bar{s}'] \times [\underline{\boldsymbol{\alpha}}, \bar{\boldsymbol{\alpha}}]$, and $\tilde{\mathcal{F}} = \mathrm{proj}_{(\boldsymbol{\gamma}_p,s,\boldsymbol{\alpha})} \mathcal{F}$. Then, $\mathcal{M} \cap \tilde{\mathcal{F}} = \mathcal{M}' \cap \tilde{\mathcal{F}}$ if

$$\begin{aligned}
\underline{\gamma}'_{p,k} &= \max\{\underline{\gamma}_{p,k}, \underline{\gamma}''_{p,k}\} \\
\bar{\gamma}'_{p,k} &= \min\left\{\bar{\gamma}_{p,k},\ \underline{\gamma}'_{k,p} + \frac{\|\boldsymbol{h}_k\|^2}{\delta\mu}(U - \delta W')\right\} \\
\underline{s}' &= \max\left\{\underline{s},\ 2^{\max\left\{\frac{W\delta - U}{\max_{k\in\mathcal{K}}\{u_k\}},\ V\right\}}(1+\bar{s}) - 1\right\} \\
\bar{s}' &= \min\left\{\bar{s},\ \underline{s}' + \frac{\min_k \|\boldsymbol{h}_k\|^2}{\delta\mu}(U - \delta W')\right\}
\end{aligned}$$

with

$$\underline{\gamma}''_{p,k} = \begin{cases} 2^{\max\{\frac{W\delta - U}{u_k},\ V\}}(1+\bar{\gamma}_{p,k}) - 1, & k \in \mathcal{I} \\ \max\{\ 2^{\frac{W\delta - U}{u_k}}(1+\bar{\gamma}_{p,\kappa}), 2^{V + R_k^{th}}\} - 1, & k \notin \mathcal{I} \end{cases}$$

and

$$\begin{aligned}
\mathcal{I} &= \{k \in \mathcal{K} : R_k^{th} - \log(1+\bar{\gamma}_{p,k}) > 0\} \\
U &= \max_{k\in\mathcal{K}}\{u_k\}\log(1+\bar{s}) + \sum_{k\in\mathcal{K}} u_k \log(1+\bar{\gamma}_{p,k}) \\
V &= \sum_{k\in\mathcal{I}} \left(R_k^{th} - \log(1+\bar{\gamma}_{p,k})\right) - \log(1+\bar{s}) \\
W &= \mu\Big(\underline{s} \max_k \|\boldsymbol{h}_k\|^{-2} + \textstyle\sum_{k\in\mathcal{K}} \underline{\gamma}_{p,k}\|\boldsymbol{h}_k\|^{-2}\Big) + P_c \\
W' &= \mu\Big(\underline{s}' \max_k \|\boldsymbol{h}_k\|^{-2} + \textstyle\sum_{k\in\mathcal{K}} \underline{\gamma}'_{p,k}\|\boldsymbol{h}_k\|^{-2}\Big) + P_c.
\end{aligned}$$

*Proof:* Due to monotonicity, a necessary condition for $\mathcal{M} \cap \tilde{\mathcal{F}} \neq \emptyset$ is that (5d), (6i), (21) hold when evaluated at $(\bar{\boldsymbol{\gamma}}_p, \bar{s}, \bar{\boldsymbol{\alpha}})$. For (21), this implies

$$\max_{k\in\mathcal{K}} \sum_{k\in\mathcal{K}} u_k C_k + \sum_{k\in\mathcal{K}} u_k \log(1+\bar{\gamma}_{p,k}) \tag{44a}$$

$$\ge \delta\left(\mu\left(\|\boldsymbol{p}_c\|^2 + \sum_{k\in\mathcal{K}} \|\boldsymbol{p}_k\|^2\right) + P_c\right) \tag{44b}$$

$$\ge \delta \min_{\boldsymbol{p}_c,\boldsymbol{p}_1,\ldots,\boldsymbol{p}_K} \left\{\mu\left(\|\boldsymbol{p}_c\|^2 + \sum_{k\in\mathcal{K}} \|\boldsymbol{p}_k\|^2\right) + P_c\right\} \tag{44c}$$

$$\ge \delta\left(\mu\left(\min \|\boldsymbol{p}_c\|^2 + \sum_{k\in\mathcal{K}} \min \|\boldsymbol{p}_k\|^2\right) + P_c\right) \tag{44d}$$

where the minimum in (44c) and (44d) is such that $\boldsymbol{\gamma}_p \in \mathcal{M}$, i.e., for all $\kappa = 1,\ldots,K$,

$$\min_{\boldsymbol{p}_c,\ldots,\boldsymbol{p}_K} \|\boldsymbol{p}_\kappa\|^2 \quad \text{s.t.} \quad \forall k : \underline{\gamma}_{p,k} \le \frac{|\boldsymbol{h}_k^H \boldsymbol{p}_k|^2}{\sum\limits_{j\in\mathcal{K}\setminus k} |\boldsymbol{h}_k^H \boldsymbol{p}_j|^2 + 1} \le \bar{\gamma}_{p,k}.$$

This can be relaxed to

$$\min_{\boldsymbol{p}_c,\ldots,\boldsymbol{p}_K} \|\boldsymbol{p}_k\|^2 \quad \text{s.t.} \quad \underline{\gamma}_{p,k} \le |\boldsymbol{h}_k^H \boldsymbol{p}_k|^2. \tag{45}$$

After transforming (45) into a SOCP, an optimal solution can be readily obtained from the Karush-Kuhn-Tucker conditions as $\boldsymbol{p}_\kappa^\star = \sqrt{\underline{\gamma}_{p,\kappa}}\frac{\boldsymbol{h}_\kappa}{\|\boldsymbol{h}_\kappa\|^2}$ with optimal value $\frac{\underline{\gamma}_{p,\kappa}}{\|\boldsymbol{h}_\kappa\|^2}$. Likewise, a lower bound for $\|\boldsymbol{p}_c\|^2$ is obtained as $\frac{\underline{s}}{\min_k \|\boldsymbol{h}_k\|^2}$. Combining this with (44) and (6i), we get the necessary condition $U \ge W\delta$.

Let $\mathcal{M}'' = [\underline{\boldsymbol{\gamma}}'_p, \bar{\boldsymbol{\gamma}}_p] \times [\underline{s}', \bar{s}] \times [\underline{\boldsymbol{\alpha}}, \bar{\boldsymbol{\alpha}}]$. It follows from $U \ge W\delta$, that every dual feasible $\gamma_{p,\kappa}$ in $\mathcal{M}$ satisfies

$$W\delta \le U - u_\kappa \log(1+\bar{\gamma}_{p,\kappa}) + u_\kappa \log(1+\gamma_{p,\kappa}). \tag{46}$$

This is equivalent to

$$\gamma_{p,\kappa} \geq 2^{\frac{W\delta - U}{u_\kappa}}(1+\bar{\gamma}_{p,\kappa}) - 1. \tag{47}$$

Hence, every $\gamma_{p,\kappa} \in \mathrm{proj}_{\gamma_{p,\kappa}}(\mathcal{M} \cap \tilde{\mathcal{F}})$ needs to satisfy (47). From the initial remark, we further observe that (5d) and (6i) can only hold if $V \leq 0$. Thus, every $\gamma_{p,\kappa} \in \mathcal{M} \cap \tilde{\mathcal{F}}$ with $\kappa \in \mathcal{I}$ satisfies

$$\log(1+\gamma_{p,\kappa}) \geq V + \log(1+\bar{\gamma}_{p,\kappa}) \tag{48}$$

$$\Leftrightarrow \quad \gamma_{p,\kappa} \geq 2^V(1+\bar{\gamma}_{p,\kappa}) - 1, \tag{49}$$

and every $\gamma_{p,\kappa} \in \mathrm{proj}_{\gamma_{p,\kappa}}(\mathcal{M} \cap \tilde{\mathcal{F}})$ with $\kappa \notin \mathcal{I}$ satisfies

$$\log(1+\gamma_{p,\kappa}) \geq V + R^{th}_\kappa \tag{50}$$

$$\Leftrightarrow \quad \gamma_{p,\kappa} \geq 2^{V+R^{th}_\kappa} - 1. \tag{51}$$

This establishes $\underline{\gamma}'_p$. The lower bound $\underline{s}'$ for $s$ is obtained analogously.

Further, every $\gamma_{p,\kappa} \in \mathrm{proj}_{\gamma_{p,\kappa}}(\mathcal{M}'' \cap \mathcal{F})$ satisfies

$$\delta\Big(W' + \frac{\mu}{\|\boldsymbol{h}_\kappa\|^2}\left(\gamma_{\kappa,p} - \underline{\gamma}'_{\kappa,p}\right)\Big) \leq U \tag{52}$$

$$\Leftrightarrow \quad \gamma_{k,p} \leq \underline{\gamma}'_{\kappa,p} + \frac{\|\boldsymbol{h}_\kappa\|^2}{\delta\mu}(U - \delta W'). \tag{53}$$

Similarly, every $s \in \mathrm{proj}_s(\mathcal{M}'' \cap \mathcal{F})$ satisfies

$$\delta\Big(W' + \frac{\mu}{\min_k \|\boldsymbol{h}_\kappa\|^2}\left(s - \underline{s}'\right)\Big) \leq U \tag{54}$$

$$\Leftrightarrow \quad s \leq s' + \frac{\min_k \|\boldsymbol{h}_\kappa\|^2}{\delta\mu}(U - \delta W'). \tag{55}$$

Hence, the upper bounds on $\boldsymbol{\gamma}_p$ and $\boldsymbol{s}$ can be reduced to $\bar{\gamma}'_p$ and $\bar{s}'$, respectively. ∎

*Corollary 2:* Let $\mathcal{M}$, $\tilde{\mathcal{F}}$, $U$, $V$, and $W$ be as in Lemma 5. Then, $\mathcal{M}$ is infeasible, i.e., $\mathcal{M} \cap \tilde{\mathcal{F}} = \emptyset$, if $V > 0$ or $U < W\delta$.

*Proof:* From the proof of Lemma 5, we know that every $\mathcal{M} \cap \tilde{\mathcal{F}} \neq \emptyset$ satisfies $V \leq 0$ and $U \geq W\delta$. ∎

### *E. Algorithm and Convergence*

The complete algorithm is stated in Algorithm 2. It is essentially a BRB procedure [48], [52] that solves the SIT dual of (6) and updates the constant $\delta$ whenever a primal feasible point is encountered.

The algorithm is initialized in Step 0. The initial box $\mathcal{M}_0$ is computed as discussed in Section IV-C. The set $\mathscr{R}_k$ holds the current partition of the feasible set, $\delta_k$ is the current best value adjusted by the tolerance $\eta$, and $\bar{\boldsymbol{x}}^k$ is the current best solution (CBS). If a primal feasible solution $\boldsymbol{y}^0$ is known, it can be used to hot start the algorithm where the variables $\boldsymbol{\gamma}_p^0$ and $s$ are initialized from $\boldsymbol{y}^0$ as in Lemma 4. Observe that $\boldsymbol{y}^0$ needs to satisfy (6f) and (6g). In Step 1, the box most likely to contain a good feasible solution is selected as $\mathcal{M}_k$ and bisected. The new boxes are stored in $\mathscr{P}_k$ and reduced according to Lemma 5 in Step 2. The reduced boxes replace the original boxes in $\mathscr{P}_k$. In Step 3, bounds for each box in $\mathscr{P}_k$ are computed, infeasibility is detected, and dual feasible points are obtained from the bounding problem (cf. Sections IV-A and IV-B). For each of these dual feasible points, primal feasibility is checked in Step 4. If true, a primal feasible point is recovered as established in Lemma 4 and the corresponding primal objective value is computed. Should any of these feasible points achieve a higher objective value than the CBS, the CBS and $\delta_k$ are updated in Step 5. Boxes that cannot contain primal $\varepsilon$-essential feasible solutions are pruned in Step 6. If the partition $\mathscr{R}_k$ contains undecided boxes, the algorithm is continued in Step 7.

**Algorithm 2** SIT Algorithm for (6)

**Step 0** **(Initialization)** Set $\varepsilon, \eta > 0$. Let $k = 1$ and $\mathscr{R}_0 = \{\mathcal{M}_0\}$ with $\mathcal{M}_0$ as in Section IV-C. If an initial feasible solution $\boldsymbol{y}^0 = (\boldsymbol{p}_c^0, \ldots, \boldsymbol{p}_K^0)$ is available, set $\delta_0 = \eta + v(5)|_{\boldsymbol{y}_0}$ and initialize $\bar{\boldsymbol{x}}^0 = (\boldsymbol{\gamma}_p^0, s^0, \boldsymbol{\alpha}^0)$ with (2), $s^0 = \min_{i\in\mathcal{K}} \gamma_{c,i}^0$, and $\alpha_i^0 = \angle \boldsymbol{h}_i^H \boldsymbol{p}_c^0$, $i > 1$. Otherwise, do not set $\bar{\boldsymbol{x}}^0$ and choose $\delta_0 = 0$.

**Step 1** **(Branching)** Let

$$\mathcal{M}_k = [\boldsymbol{r}^k, \boldsymbol{s}^k] = \arg\min\{\beta(\mathcal{M}) \,|\, \mathcal{M} \in \mathscr{R}_{k-1}\}.$$

Bisect $\mathcal{M}_k$ via $(\boldsymbol{v}^k, j_k)$ where $j_k \in \arg\max_j s_j^k - r_j^k$ and $\boldsymbol{v}^k = \frac{1}{2}(\boldsymbol{s}^k + \boldsymbol{r}^k)$ as in (17) and set $\mathscr{P}_k = \{\mathcal{M}_-^k, \mathcal{M}_+^k\}$.

**Step 2** **(Reduction)** Replace each box $\mathcal{M} \in \mathscr{P}_k$ with $\mathcal{M}'$ as in Lemma 5.

**Step 3** **(Bounding)** For each reduced box $\mathcal{M} \in \mathscr{P}_k$, perform a preliminary feasibility check with Corollary 2 and, if necessary, solve (33). If infeasible, set $\beta(\mathcal{M}) = \infty$. Otherwise, set $\beta(\mathcal{M})$ to the optimal value of (33) and obtain a dual feasible point $\boldsymbol{x}(\mathcal{M})$ as in (34).

**Step 4** **(Feasible Point)** For each $\mathcal{M} \in \mathscr{P}_k$, if $\beta(\mathcal{M}) \leq 0$ compute $\tilde{g}(\boldsymbol{x}(\mathcal{M}))$ as in (25). If $\tilde{g}(\boldsymbol{x}(\mathcal{M})) \leq 0$, $\boldsymbol{x}(\mathcal{M})$ is primal feasible. Recover $\boldsymbol{x}'(\mathcal{M})$ from the solution of (25) with $\boldsymbol{\gamma}'_p$, $s'$ as in Lemma 4 and $\alpha'_i = \angle e_i^*$, $i > 1$, with $\boldsymbol{e}^*$ as in Lemma 4. Compute the primal objective value $f(\mathcal{M}) = \sum_{j\in\mathcal{K}} u_k(\tilde{C}_j^\star + \log(1+\gamma'_{p,j}))$, where $\tilde{\boldsymbol{C}}^*$ is the optimal solution of (35). If $\beta(\mathcal{M}) > 0$ or $\tilde{g}(\boldsymbol{x}(\mathcal{M})) > 0$, set $f(\mathcal{M}) = -\infty$.

**Step 5** **(Incumbent)** Let $\mathcal{M}' \in \arg\max\{f(\mathcal{M}) : \mathcal{M} \in \mathscr{P}_k\}$. If $f(\mathcal{M}') > \delta_{k-1} - \eta$, set $\bar{\boldsymbol{x}}^k = \boldsymbol{x}'(\mathcal{M}')$ and $\delta_k = f(\mathcal{M}') + \eta$. Otherwise, set $\bar{\boldsymbol{x}}^k = \bar{\boldsymbol{x}}^{k-1}$ and $\delta_k = \delta_{k-1}$.

**Step 6** **(Pruning)** Delete every $\mathcal{M} \in \mathscr{P}_k$ with $\beta(\mathcal{M}) > -\varepsilon$. Let $\mathscr{P}'_k$ be the collection of remaining sets and set $\mathscr{R}_k = \mathscr{P}'_k \cup (\mathscr{R}_{k-1} \setminus \{\mathcal{M}_k\})$.

**Step 7** **(Termination)** Terminate if $\mathscr{R} = \emptyset$: If $\bar{\boldsymbol{x}}^k$ is not set, then (6) is $\varepsilon$-essential infeasible; else $\bar{\boldsymbol{x}}^k$ is an essential $(\varepsilon, \eta)$-optimal solution of (6). Otherwise, update $k \leftarrow k+1$ and return to Step 1.

Convergence of the algorithm follows from the previous discussion and is formally established next.

*Theorem 1:* Algorithm 2 converges in finitely many steps to the $(\varepsilon, \eta)$-optimal solution of (6) or establishes that no such solution exists.

*Proof:* Lemma 5 ensures that no feasible solution candidates with objective values greater than $\delta_k$ are lost in Step 2. The bisection in Step 1 in exhaustive [52, Cor. 6.2]. Hence, $\max_{\boldsymbol{x},\boldsymbol{y}\in\mathcal{M}_k} \|\boldsymbol{x} - \boldsymbol{y}\| \to 0$ as $k \to \infty$. Then, by virtue of Lemma 3 and the observation that (27) and (33) are equivalent, Step 3 satisfies the convergence criterion in Lemma 2. Lemma 4 establishes that the point in Step 4 is primal feasible and suitable with respect to Lemma 3. It follows that, for fixed $\delta_k$, after a finite number of iterations, either a primal feasible point is found or all boxes are pruned in Step 6 and the algorithm is terminated in Step 7. Hence, from Lemma 1, $\delta_k$ is, upon termination, either a $(\varepsilon, \eta)$-optimal solution of (6) or, if $\delta_k$ was not set with some primal feasible point, the problem

is $\varepsilon$-essential infeasible.

It is established in [47, App. C] that updating $\delta_k$ with encountered primal feasible points does not invalidate the bounds in $\mathscr{R}_k$. Hence, restarting the procedure upon updating $\delta_k$ in Step 4 is not necessary to ensure correct convergence. Finally, observe that the primal objective is bounded above by the global optimum and below by zero. Hence, the initialization of $\delta_0$ in Step 0 is valid. Moreover, the sequence $\{\delta_k\}_k$ converges to a value between $v(6)$ and $v(6)+\eta$. For $\eta > 0$, this sequence is clearly finite. ■

## V. Numerical Evaluation

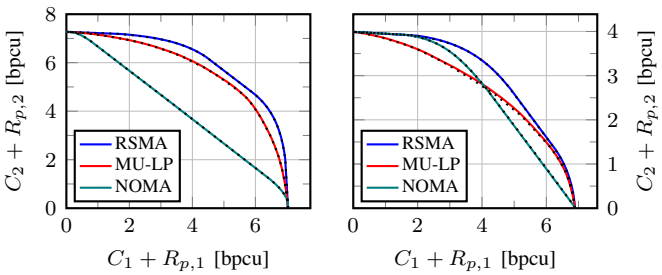
In this section, we employ the developed algorithm to compare RSMA with MU-LP and NOMA in terms of achievable rate region, maximum sum rate and EE. Further, the performance of first-order optimal solution methods for these problems is measured against the global solution and some numerical properties of Algorithm 2 are examined.

### A. System Performance

Consider a BS with $M = 2$ transmit antennas that serves $K = 2$ single antenna users on the same spectrum as described in Section II. We employ Algorithm 2 to solve the beamforming problem (5) for RSMA and its special cases MU-LP and 2-user NOMA. The goal of the experiments in this subsection is to evaluate the performance gap between these schemes based on the strong optimality guarantees provided by Algorithm 2. In addition, we also obtain beamforming solutions using state-of-the-art first-order optimal algorithms and compare them to the results of Algorithm 2. This will give an indication of the usability of those faster algorithms for practical system evaluation.

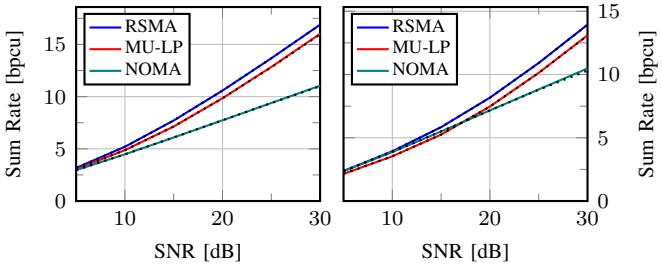
The channels of user $k$ are chosen randomly using circularly symmetric complex Gaussian distribution with zero mean and variance $\sigma_k^2$. A set of 100 independent and identically distributed (i.i.d.) feasible channel realizations is generated for each simulation separately. Computation time and memory consumption per problem instance were limited. This leads to results being averaged over less than 100 samples per simulation. Two different channel statistics are considered: one where both users' channels are generated with equal variances and one with roughly 10 dB disparity in the variances.

*1) Rate Region:* We start with the achievable rate region for RSMA, MU-LP and NOMA. The boundary points for each strategy are calculated by setting $R_k^{th} = 0$ and $u_1 = 1$ in (3). Following [54], we vary the weight $u_2 \in \{10^x | x = -3, -1, -0.95, -0.8, \cdots, 0.95, 1, 3\}$. The resulting rate region is obtained from the convex hull over the computed boundary points. Results for a SNR of 20 dB are displayed in Fig. 2. Figure 2a was averaged over 63 channel realizations, while Fig. 2b was obtained from 61 realizations. It can be observed that the achievable rate region of RSMA is strictly larger than that of MU-LP and NOMA. In case of equal channel statistics, MU-LP also strictly outperforms NOMA, while in the case with disparate statistics neither MU-LP nor NOMA is superior to the other. However, as RSMA includes both strategies as special cases and allows arbitrary combinations of them, its rate region is strictly larger. These observations are in line with previous evaluations [3] but are scientifically more reliable since they rely on proven globally optimal solutions to (5).




(a) $\sigma_1^2 = \sigma_2^2 = 1$ (b) $\sigma_1^2 = 1, \sigma_2^2 = 0.09$

Fig. 2. Achievable rate regions for RSMA, MU-LP and NOMA at an SNR of 20 dB. Colored lines are globally optimal results obtained from Algorithm 2 and dashed lines are the corresponding results from a WMMSE algorithm.




(a) $\sigma_1^2 = \sigma_2^2 = 1$ (b) $\sigma_1^2 = 1, \sigma_2^2 = 0.09$

Fig. 3. Maximum achievable sum rate for RSMA, MU-LP and NOMA with increasing QoS constraints (see text for details). Colored lines are globally optimal results obtained from Algorithm 2 and dashed lines are the corresponding results from a WMMSE algorithm.

The first-order optimal solutions are computed using the WMMSE algorithm from [3] for RSMA and NOMA, and MU-LP. For each parameter combination, a single initialization was used. In particular, the common and private stream precoders are initialized using SVD and maximum ratio transmission, respectively, as in [10]. The results are displayed as black dashed lines. While the WMMSE algorithm does not always achieve the optimal solution, as can be seen in the MU-LP performance in Fig. 2b, its solution is sufficiently close to the true solution to allow drawing conclusions based on the results. Moreover, the obtained solution is well suited for practical system design.

*2) Sum Rate Maximization:* We maximize the sum rate under QoS constraints, i.e., we solve (3) with $u_k = 1$, $k = 1, 2$. A SNR range from 5 dB to 30 dB with 5 dB increments is considered. The corresponding QoS constraints $R_k^{th}$ are chosen as 0.1, 0.2, 0.4, 0.6, 0.8, and 1 (all in bpcu), respectively. The results are displayed in Fig. 3. Both plots were obtained by averaging over 90 i.i.d. channel realizations. The same observations as before hold true: RSMA outperforms MU-LP and NOMA in both cases, while there is not a clearly superior strategy when comparing MU-LP and NOMA. Again, the WMMSE algorithm performs quite well compared to the global solution and appears to be a good practical choice.

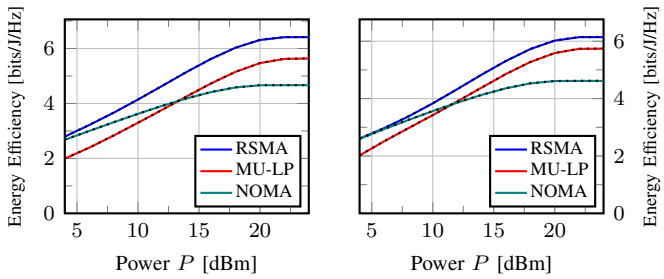



Fig. 4. Energy efficiency of RSMA, MU-LP and NOMA with $R_k^{th} = 1\,\mathrm{bpcu}$. Colored lines are globally optimal results obtained from Algorithm 2 and dashed lines are the corresponding results from an SCA algorithm.

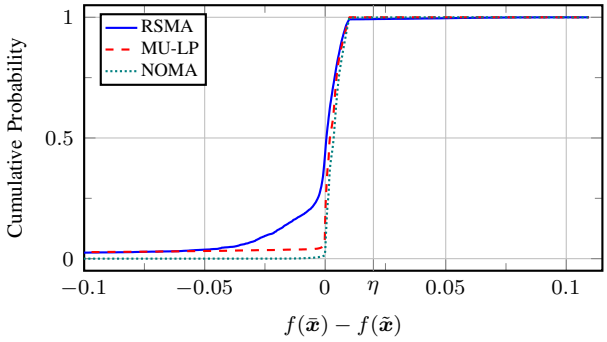



Fig. 5. Empirical CDF of the difference between the optimal values returned by Algorithm 2 and the first-order optimal baseline algorithm, where $f(\bar{\boldsymbol{x}})$ is the optimal value returned by Algorithm 2 and $f(\tilde{\boldsymbol{x}})$ is the corresponding objective value for the solution returned by the first-order optimal solver.

*3) Energy Efficiency:* The third problem type supported by Algorithm 2 is EE maximization. We solve problem (4) for maximum transmit powers ranging from 4 dBm to 30 dBm in steps of 2 dBm, with power amplifier inefficiency $\mu = 0.35$ and static circuit power consumption $P_c = MP_{\mathrm{dyn}} + P_{\mathrm{sta}}$, where $M$ is the number of transmit antennas, $P_{\mathrm{dyn}} = 27\,\mathrm{dBm}$, and $P_{\mathrm{sta}} = 1\,\mathrm{mW}$. Potentially suboptimal solutions are computed with the SCA approach [55] as in [12] for RSMA, NOMA, and MU-LP. Again a single initialization per parameter combination is used, following the same methodology as in [12]. Results are shown in Figs. 4a and 4b and were obtained by averaging over 29 and 32 i.i.d. channel realizations, respectively. The results follow the usual shape of EE maximization, where the EE first increases and then saturates at some point. Interestingly, while RSMA clearly outperforms the other two schemes, MU-LP has always higher EE than NOMA when the transmit power budget is large enough, while for constrained transmit powers, NOMA has slightly higher efficiency than MU-LP. As before, the first-order optimal results, this time obtained with SCA, are very close to the globally optimal solution and we can conclude that in most cases such an algorithm will be sufficient for performance analysis.

### B. Numerical Performance

We have evaluated Algorithm 2 and two state-of-the-art first-order optimal methods in a real world setting. The key observation is that the WMMSE and SCA methods without proven convergence to the global solution perform very well and, on average, are virtually equal to the globally optimal solution. In this subsection, we first take a closer look at the numerical accuracy of the first-order optimal methods and then study the numerical stability of Algorithm 2.

*1) Numerical Accuracy:* The results in Section V-A were obtained from Algorithm 2 with tolerances $\eta = 0.02$ and $\varepsilon = 10^{-7}$. The numerical tolerances of the first-order optimal methods were chosen small enough not be relevant. Figure 5 shows the empirical cumulative distribution function (CDF) of the difference between the globally optimal solution and the first-order optimal solution computed for the analyses in Section V-A with both metrics, WSR and EE. Accordingly, a total of 11 520 computed data points for RSMA, 12 600 points for MU-LP and 24 880 for NOMA form the basis of Fig. 5. A negative value indicates that the solution of Algorithm 2 achieves a larger objective value than that computed by the WMMSE or SCA approach. Instead, a positive value indicates that the first-order optimal solution is better than the one obtained by Algorithm 2. Recalling the definition of $\eta$-optimality in (12), it is apparent that this is not an unexpected outcome.

There exists a small amount of solutions returned by Algorithm 2 that is not within an $\eta$-region around the global optimal solution. This is indicated by a deviation of more than $\eta$ from the first-order optimal solution. In particular, this affects 98 of the RSMA solutions (0.85 %), 40 of the NOMA solutions (0.16 %), and none of the MU-LP solutions. The reason for this is the tightening of nonconvex constraints necessary for the SIT approach. Reducing the size of $\varepsilon$ will resolve this numerical issue but also leads to slower convergence. The fact that the MU-LP solutions are unaffected indicates that the likely reason is in the tightening of (6d)–(6f).

*2) Numerical Stability:* Finally, we evaluate the numerical stability of Algorithm 2 in comparison to conventional BB based methods. We focus on multiple unicast beamforming, i.e., where $\boldsymbol{p}_c = \mathbf{0}$, as this special case of the more general problem is already quite difficult to solve with BB. As baseline comparison, we implemented two methods. The first is a straightforward BB solution of (6) with $\boldsymbol{p}_c = 0$ as published in [39], [40]. We denote this algorithm as "BB" in the results. The bounding problem in this algorithm is often difficult to solve for state-of-the-art convex solvers (e.g., Mosek [49]) since the feasible region can become extremely small. Following [39, §2.2.2], the feasible set can be relaxed such that the bounding problem always has good numerical properties. Interestingly, the resulting problem is similar to the SIT bounding problem in Section IV-A. The downside of this approach is that feasible point acquisition for the BB procedure becomes much harder. We denote this algorithm as "BB2".

The numerical evaluation is based on 100 random i.i.d. channel realizations. We solved (5) for $u_k = 1$, $\boldsymbol{p}_c = \mathbf{0}$, $\mu = 0$, $P_c = 0$, $R_k^{th} = 0$, $\frac{P}{\mathrm{dB}} = -10, -5, \ldots, 20$, and $K = M \in \{2, 3, 4\}$. This results in 700 problem instances per $K$.

For $K = 2$, BB2 stalled in 364 problem instances, while the other algorithms solved all problems. For $K = 3$, BB2 stalled

TABLE II
MEAN / MEDIAN RUN TIMES TO OBTAIN THE OPTIMAL SOLUTION. INSTANCES WHERE NOT ALL ALGORITHMS CONVERGED ARE IGNORED.

| | $K=2$ | $K=3$ | $K=4$ |
|---|---|---|---|
| Alg. 2 | 0.175 s / 0.099 s | 4.579 s / 1.959 s | 334.8 s / 126.3 s |
| BB | 0.173 s / 0.091 s | 7.605 s / 2.606 s | — |
| BB2 | 42.41 s / 2.380 s | 158.5 s / 12.42 s | 704.1 s / 265.8 s |

in 146 instances and BB failed 13 times due to numerical problems of the convex solver. Finally, for $K = 4$, BB did not solve a single problem instance due to numerical issues and BB2 stalled in 27 instances. Moreover, Algorithm 2 and BB2 did not solve the problem within 60 minutes in 4 and 60 instances, respectively. Average computation times on a single core of an Intel Cascade Lake Platinum 9242 CPU are reported in Table II. It can be observed that the proposed Algorithm 2 is more efficient than the two baseline algorithms especially when more users are in the system. Moreover, the joint beamforming problem, i.e., with $\boldsymbol{p}_c \neq \mathbf{0}$, was solved by Algorithm 2 for $K = 2$ with mean and median run times of 942 s and 2786 s. However, 23 instances were not solved within 12 hours and in the simulations presented in Section V-A, we observed some parameter combinations with very slow convergence speed, especially in EE maximization problems.

As a final note, observe from the discussion in Section IV that the complexity scales with $O(\exp(2K))$ in the number of users and polynomially in the number of antennas $M$. Hence, no noticeable changes in the reported run times are to be expected by varying $M$.

## VI. CONCLUSIONS

We have developed a globally optimal beamforming algorithm for WSR and EE maximization in MISO downlink systems with RSMA. The algorithm exhibits finite convergence and is the first method to solve this optimization problem. It is also the first beamforming algorithm based on the SIT-BB approach. Two user NOMA and MU-LP beamforming are incorporated as special cases. We have shown numerically that the proposed algorithm outperforms state-of-the-art globally optimal beamforming algorithms for the MU-LP problem, both in terms of numerical stability and practical convergence speed. Extensive numerical experiments establish that contemporary suboptimal solution methods for RSMA beamforming often obtain a solution very close to the global optimum. In particular, there is virtually no difference between the suboptimal solution and the true optimum when evaluating the average performance over a large number of channel realizations. Hence, this paper establishes that WMMSE and SCA-based methods are suitable choices for such performance comparisons. This effectively strengthens the results of many earlier studies in this area, as it retrospectively validates the numerical approach taken to compare the performance of RSMA against NOMA and MU-LP.

## APPENDIX A
## PROOF OF PROPOSITION 1

Let $(\boldsymbol{x}^\star, \boldsymbol{\gamma}_c^\star)$ be a solution of (5) and set $s^\star = \min_k \gamma_{c,k}^\star$. Constraints (5d) and (5e) are part of both problems. Constraint (5c) is equivalent to

$$\sum_{k\in\mathcal{K}} C_k \leq \log(1 + \min_{k\in\mathcal{K}} \gamma_{c,k}). \tag{56}$$

Since $(\boldsymbol{x}^\star, \boldsymbol{\gamma}_c^\star)$ satisfies (56), $(\boldsymbol{x}^\star, s^\star)$ satisfies (6i). Finally, (6c) is a relaxed version of (5b) and (6d) is equivalent to the definition of $s^\star$.

For the converse, let $(\boldsymbol{x}^\star, s^\star)$ be a solution of (6) and set $\gamma_{c,k}^\star = \frac{|\boldsymbol{h}_k^H \boldsymbol{p}_c^\star|^2}{\sum_{j\in\mathcal{K}} |\boldsymbol{h}_k^H \boldsymbol{p}_j^\star|^2 + 1}$ for all $k \in \mathcal{K}$. Since the objective is increasing in $\boldsymbol{\gamma}_p$, constraint (8b) is always active in the optimal solution if $u_k > 0$. Otherwise, i.e., for $u_k = 0$, relaxing (5b) does not relax (5d). Hence, (8b) is equivalent to the $\gamma_{p,k}$ part of (5b) (and the $\gamma_{c,k}$ part is satisfied by definition). Constraint (5c) is equivalent to (56). With an auxiliary variable $s = \min_k \gamma_{c,k}$, the right-hand side (RHS) of (56) can be replaced by $\log(1+s)$. Due to monotonicity, the relaxed version $s \leq \min_k \gamma_{c,k}$ is active in the optimal solution. Observing that (8c) is the smooth variant of this constraint completes this part of the proof.

For the last part of the proposition, it suffices to show that every solution of (6) solves (8). Observe that (8b) is equivalent to

$$\sqrt{\gamma_{p,k}} \left( \sum\nolimits_{j\in\mathcal{K}\setminus k} |\boldsymbol{h}_k^H \boldsymbol{p}_j|^2 + 1 \right)^{1/2} \leq |\boldsymbol{h}_k^H \boldsymbol{p}_k| \tag{57}$$

and that the solution is invariant to rotations of $\boldsymbol{p}_k$, $k \in \mathcal{K}$, i.e., if $\boldsymbol{p}_k^\star$ solves (8), then $\boldsymbol{p}_k^\star e^{j\phi}$ also solves (8) for all real-valued $\phi$ [44]. Hence, constraint (6f) can be added to (8) without reducing the optimal value. Then, (57) is equivalent to (6b).

Similarly, (8c) is equivalent to

$$\sqrt{s} \left( \sum\nolimits_{j\in\mathcal{K}} |\boldsymbol{h}_1^H \boldsymbol{p}_j|^2 + 1 \right)^{1/2} \leq |\boldsymbol{h}_k^H \boldsymbol{p}_c| \tag{58}$$

for all $k \in \mathcal{K}$ and the solution is invariant to rotations in $\boldsymbol{p}_c$. However, except for degenerate cases, only one RHS of (58) can be made real-valued. Without loss of generality, this is done for $k = 1$ by adding (6g) to (8). For the remaining $k - 1$ constraints, introduce auxiliary variables $d_k = |\boldsymbol{h}_k^H \boldsymbol{p}_c|$ and observe that relaxing $0 \leq d_k \leq |\boldsymbol{h}_k^H \boldsymbol{p}_c|$ does not decrease the optimal value of (8). However, it also does not increase the optimal value since it is increasing in $s$ and, hence, also increasing in $d_k$. Finally, introducing the constraint $e_k = \boldsymbol{h}_k^H \boldsymbol{p}_c$ results in (6).

## REFERENCES


[1] B. Matthiesen, Y. Mao, P. Popovski, and B. Clerckx, "Globally optimal beamforming for rate splitting multiple access," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Toronto, Canada, Jun. 2021.

[2] B. Clerckx, H. Joudeh, C. Hao, M. Dai, and B. Rassouli, "Rate splitting for MIMO wireless networks: A promising PHY-layer strategy for LTE evolution," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 98–105, May 2016.

[3] Y. Mao, B. Clerckx, and V. O. K. Li, "Rate-splitting multiple access for downlink communication systems: bridging, generalizing, and outperforming SDMA and NOMA," *EURASIP J. Wireless Commun. Netw.*, vol. 2018, no. 1, p. 133, May 2018.

[4] Y. Mao and B. Clerckx, "Beyond dirty paper coding for multi-antenna broadcast channel with partial CSIT: A rate-splitting approach," *IEEE Trans. Commun.*, vol. 68, no. 11, pp. 6775–6791, Nov. 2020.

[5] Y. Mao *et al.*, "Rate-splitting multiple access: Fundamentals, survey, and future research trends," *arXiv preprint arXiv:2201.03192*, 2022.

[6] B. Clerckx, Y. Mao, R. Schober, and H. V. Poor, "Rate-splitting unifying SDMA, OMA, NOMA, and multicasting in MISO broadcast channel: A simple two-user rate analysis," *IEEE Wireless Commun. Lett.*, vol. 9, pp. 349–353, Mar. 2020.

[7] A. Carleial, "Interference channels," *IEEE Trans. Inf. Theory*, vol. 24, no. 1, pp. 60–70, 1978.

[8] T. Han and K. Kobayashi, "A new achievable rate region for the interference channel," *IEEE Trans. Inf. Theory*, vol. 27, Jan. 1981.

[9] E. Piovano and B. Clerckx, "Optimal DoF region of the K-user MISO BC with partial CSIT," *IEEE Commun. Lett.*, vol. 21, no. 11, pp. 2368–2371, Nov. 2017.

[10] Y. Mao, E. Piovano, and B. Clerckx, "Rate-splitting multiple access for overloaded cellular internet of things," *IEEE Trans. Commun.*, vol. 69, no. 7, pp. 4504–4519, 2021.

[11] H. Joudeh and B. Clerckx, "Sum-rate maximization for linearly precoded downlink multiuser MISO systems with partial CSIT: A rate-splitting approach," *IEEE Trans. Commun.*, vol. 64, no. 11, Nov. 2016.

[12] Y. Mao, B. Clerckx, and V. O. K. Li, "Energy efficiency of rate-splitting multiple access, and performance benefits over SDMA and NOMA," in *Proc. IEEE Int. Symp. Wireless Commun. Syst. (ISWCS)*, Aug. 2018.

[13] Y. Mao, B. Clerckx, and V. O. K. Li, "Rate-splitting for multi-antenna non-orthogonal unicast and multicast transmission: Spectral and energy efficiency analysis," *IEEE Trans. Commun.*, vol. 67, no. 12, pp. 8754–8770, Dec. 2019.

[14] Y. Mao, B. Clerckx, J. Zhang, V. O. K. Li, and M. Arafah, "Max-min fairness of K-user cooperative rate-splitting in MISO broadcast channel with user relaying," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6362–6376, Oct. 2020.

[15] Z. Li, C. Ye, Y. Cui, S. Yang, and S. Shamai, "Rate splitting for multi-antenna downlink: Precoder design and practical implementation," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1910–1924, 2020.

[16] H. Fu, S. Feng, W. Tang, and D. W. K. Ng, "Robust secure beamforming design for two-user downlink MISO rate-splitting systems," *IEEE Trans. Wireless Commun.*, vol. 19, no. 12, pp. 8351–8365, 2020.

[17] C. Hao, Y. Wu, and B. Clerckx, "Rate analysis of two-receiver MISO broadcast channel with finite rate feedback: A rate-splitting approach," *IEEE Trans. Commun.*, vol. 63, no. 9, pp. 3232–3246, Sep. 2015.

[18] O. Dizdar, Y. Mao, and B. Clerckx, "Rate-splitting multiple access to mitigate the curse of mobility in (massive) MIMO networks," *IEEE Trans. Commun.*, vol. 69, no. 10, pp. 6765–6780, 2021.

[19] G. Lu, L. Li, H. Tian, and F. Qian, "MMSE-based precoding for rate splitting systems with finite feedback," *IEEE Commun. Lett.*, vol. 22, no. 3, pp. 642–645, Mar. 2018.

[20] G. Zhou, Y. Mao, and B. Clerckx, "Rate-splitting multiple access for multi-antenna downlink communication systems: Spectral and energy efficiency tradeoff," *IEEE Trans. Wireless Commun.*, 2021.

[21] M. Dai, B. Clerckx, D. Gesbert, and G. Caire, "A rate splitting strategy for massive MIMO with imperfect CSIT," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 4611–4624, Jul. 2016.

[22] A. Papazafeiropoulos and T. Ratnarajah, "Rate-splitting robustness in multi-pair massive MIMO relay systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 8, pp. 5623–5636, Aug. 2018.

[23] A. R. Flores, R. C. De Lamare, and B. Clerckx, "Linear precoding and stream combining for rate splitting in multiuser MIMO systems," *IEEE Commun. Lett.*, vol. 24, no. 4, pp. 890–894, 2020.

[24] A. A. Ahmad, Y. Mao, A. Sezgin, and B. Clerckx, "Rate splitting multiple access in C-RAN: A scalable and robust design," *IEEE Trans. Commun.*, pp. 1–1, 2021.

[25] A. A. Ahmad, B. Matthiesen, A. Sezgin, and E. Jorswieck, "Energy efficiency in C-RAN using rate splitting and common message decoding," in *Proc. IEEE Int. Conf. Commun. (ICC) Workshop*, 2020, pp. 1–6.

[26] O. Dizdar, A. Kaushik, B. Clerckx, and C. Masouros, "Rate-splitting multiple access for joint radar-communications with low-resolution DACs," in *Proc. IEEE Int. Conf. Commun. (ICC) Workshop*, 2021.

[27] X. Su, L. Li, H. Yin, and P. Zhang, "Robust power- and rate-splitting-based transceiver design in $k$ -user MISO SWIPT interference channel under imperfect CSIT," *IEEE Commun. Lett.*, vol. 23, no. 3, pp. 514–517, Mar. 2019.

[28] M. R. Camana Acosta, C. E. G. Moreta, and I. Koo, "Joint power allocation and power splitting for MISO-RSMA cognitive radio systems with SWIPT and information decoder users," *IEEE Syst. J.*, 2020.

[29] J. An, O. Dizdar, B. Clerckx, and W. Shin, "Rate-splitting multiple access for multi-antenna broadcast channel with imperfect CSIT and CSIR," in *Proc. IEEE Annu. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, 2020.

[30] E. Piovano, H. Joudeh, and B. Clerckx, "Overloaded multiuser MISO transmission with imperfect CSIT," in *Proc. 50th Asilomar Conf. Signals, Syst. Comput.*, Nov. 2016, pp. 34–38.

[31] A. Papazafeiropoulos, B. Clerckx, and T. Ratnarajah, "Rate-splitting to mitigate residual transceiver hardware impairments in massive MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 66, no. 9, Sep. 2017.

[32] J. Zhang *et al.*, "Energy and spectral efficiency tradeoff via rate splitting and common beamforming coordination in multicell networks," *IEEE Trans. Commun.*, vol. 68, no. 12, pp. 7719–7731, 2020.

[33] C. Xu, B. Clerckx, S. Chen, Y. Mao, and J. Zhang, "Rate-splitting multiple access for multi-antenna joint radar and communications," *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 6, pp. 1332–1347, 2021.

[34] R. Cerna-Loli, O. Dizdar, and B. Clerckx, "A rate-splitting strategy to enable joint radar sensing and communication with partial CSIT," in *Proc. IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, 2021, pp. 491–495.

[35] M. Razaviyayn, "Successive convex approximation: Analysis and applications," Ph.D. dissertation, University of Minnesota, 2014.

[36] M. Medra and T. N. Davidson, "Robust downlink transmission: An offset-based single-rate-splitting approach," in *Proc. IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, June 2018, pp. 1–5.

[37] N. D. Sidiropoulos, T. N. Davidson, and Z.-Q. Luo, "Transmit beamforming for physical-layer multicasting," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 2239–2251, Jun. 2006.

[38] Z.-Q. Luo and S. Zhang, "Dynamic spectrum management: Complexity and duality," *IEEE J. Sel. Areas Commun.*, vol. 2, no. 1, Feb. 2008.

[39] E. Björnson and E. A. Jorswieck, *Optimal Resource Allocation in Coordinated Multi-Cell Systems*, ser. Found. Trends Commun. Inf. Theory. Boston, MA, USA: Now, 2013, vol. 9, no. 2–3.

[40] O. Tervo, L.-N. Tran, and M. Juntti, "Optimal energy-efficient transmit beamforming for multi-user MISO downlink," *IEEE Trans. Signal Process.*, vol. 63, no. 20, pp. 5574–5588, Oct. 2015.

[41] C. Lu and Y.-F. Liu, "An efficient global algorithm for single-group multicast beamforming," *IEEE Trans. Signal Process.*, vol. 65, no. 14, pp. 3761–3774, Jul. 2017.

[42] Y.-F. Liu, C. Lu, M. Tao, and J. Wu, "Joint multicast and unicast beamforming for the MISO downlink interference channel," in *Proc. IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*. IEEE, Jul. 2017.

[43] E. Chen, M. Tao, and Y.-F. Liu, "Joint base station clustering and beamforming for non-orthogonal multicast and unicast transmission with backhaul constraints," *IEEE Trans. Wireless Commun.*, vol. 17, no. 9, pp. 6265–6279, Sep. 2018.

[44] M. Bengtsson and B. Ottersten, "Optimal downlink beamforming using semidefinite optimization," in *Proc. 37th Annual Allerton Conf. Communication, Control, and Computing*, 1999, pp. 987–996.

[45] D. E. Knuth, *The Art of Computer Programming: Fundamental Algorithms*, 3rd ed. Reading, MA, USA: Addison-Wesley, 1997, vol. 1.

[46] B. Matthiesen, C. Hellings, E. A. Jorswieck, and W. Utschick, "Mixed monotonic programming for fast global optimization," *IEEE Trans. Signal Process.*, vol. 68, pp. 2529–2544, Mar. 2020.

[47] B. Matthiesen and E. A. Jorswieck, "Efficient global optimal resource allocation in non-orthogonal interference networks," *IEEE Trans. Signal Process.*, vol. 67, no. 21, pp. 5612–5627, Nov. 2019.

[48] B. Matthiesen, "Efficient globally optimal resource allocation in wireless interference networks," Ph.D. Thesis, Technische Universität Dresden, Dresden, Germany, Nov. 2019. [Online]. Available: https://nbn-resolving.org/urn:nbn:de:bsz:14-qucosa2-362878

[49] MOSEK ApS. (2020) MOSEK optimizer 9.2. [Online]. Available: http://mosek.com

[50] H. Tuy, "Robust solution of nonconvex global optimization problems," *J. Global Optim.*, vol. 32, no. 2, pp. 307–323, 6 2005.

[51] ——, "$\mathcal{D}(\mathcal{C})$-optimization and robust global optimization," *J. Global Optim.*, vol. 47, no. 3, pp. 485–501, Oct. 2009.

[52] ——, *Convex Analysis and Global Optimization*, 2nd ed., ser. Springer Optim. Appl. New York; Berlin, Germany; Vienna, Austria: Springer-Verlag, 2016, vol. 110.

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

[54] S. S. Christensen, R. Agarwal, E. De 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.

[55] B. Matthiesen and E. A. Jorswieck, "Optimization techniques for energy efficiency," in *Green Communications for Energy-Efficient Wireless Systems and Networks*, H. A. Suraweera, J. Yang, A. Zappone, and J. S. Thompson, Eds. Hertfordshire, UK: IET, Nov. 2020, ch. 5.