Revisiting the MIMO Capacity with Per-antenna Power Constraint: Fixed-point Iteration and Alternating Optimization

Thuy M. Pham, *Student Member, IEEE,* Ronan Farrell, *Member, IEEE* and Le-Nam Tran, *Senior Member, IEEE*

***Abstract*—In this paper, we revisit the fundamental problem of computing MIMO capacity under per-antenna power constraint (PAPC). Unlike the sum power constraint counterpart which likely admits water-filling-like solutions, MIMO capacity with PAPC has been largely studied under the framework of generic convex optimization. The two main shortcomings of these approaches are (i) their complexity scales quickly with the problem size, which is not appealing for large-scale antenna systems, and/or (ii) their convergence properties are sensitive to the problem data. As a starting point, we first consider a single user MIMO scenario and propose two provably-convergent iterative algorithms to find its capacity, the first method based on fixed-point iteration and the other based on alternating optimization and minimax duality. In particular, the two proposed methods can leverage the water-filling algorithm in each iteration and converge faster, compared to current methods. We then extend the proposed solutions to multiuser MIMO systems with dirty paper coding (DPC) based transmission strategies. In this regard, capacity regions of Gaussian broadcast channels with PAPC are also computed using closed-form expressions. Numerical results are provided to demonstrate the outperformance of the proposed solutions over existing approaches.**

***Index Terms*—MIMO, fixed-point iteration, alternating optimization, minimax duality, water-filling, dirty paper coding.**

I. Introduction

Since its invention in the mid-90s [3], [4], multiple-input multiple-output (MIMO) technology has been adopted in all modern mobile wireless networks. From a system design perspective, one of the most fundamental problems is to compute the capacity of the system of interest. For a single user MIMO (SU-MIMO) channel, pioneer studies proved that the capacity can be achieved by Gaussian input signaling [3], [4]. For multiuser MIMO (MU-MIMO) scenarios, the seminal work of [5] showed that dirty-paper coding (DPC) in fact achieves the entire capacity region of Gaussian MIMO broadcast channel (BC). Since finding the capacity of MIMO channels is computationally expensive in general, one is also interested in near-capacity achieving transmission strategies such as successive zero-forcing DPC (SZF-DPC) [6], [7], for which the achievable rate region is much easier to characterize.

The capacity of MIMO systems is investigated along with a certain type of constraint on the input covariance matrices. To this end, a majority of the related literature assumes a sum power constraint (SPC) as it usually leads to efficiently computational algorithms. In particular, under perfect channel state information (CSI) at both transmitter and receiver, the capacity of a SU-MIMO channel is found using the closed-form water-filling (WF) algorithm [3], [4]. In [8], Yu *et al.* presented an iterative WF (IWF) algorithm to compute the sum capacity for a Gaussian vector multiple access channel (MAC). In [9], Jindal *et al.* proposed sum power IWF to determine the sum capacity of Gaussian MIMO BCs by exploiting the MAC-BC duality. The entire capacity region of MIMO-BCs with a SPC was characterized in [10], [11], using conjugate gradient projection (CGP)- and pre-conditioned gradient projection-based approaches, respectively.

In reality, each antenna is associated with a separate power amplifier, each having a different dynamic range. As such, per-antenna power constraint (PAPC) is of more practical importance. If a sum power constraint is considered, some antennas may be allocated a power level that is beyond their dynamic range of the associated power amplifier, depending on fading situations. This will result in nonlinear distortion that has a detrimental impact on the whole system. In [5], it was shown that DPC still achieves the full capacity region of the MIMO BC under PAPC. However, finding the DPC region with PAPC is more numerically difficult than with a SPC. In fact, no closed-form design has been reported for the computation of the capacity region of the MIMO BC subject to PAPC. For this reason, numerous research endeavors have been made to understand performance limits of various sub-optimal transmission strategies such as zero-forcing (ZF) beamforming, minimum mean square error (MMSE), and SZF-DPC [12]–[19].

The capacity of a SU-MIMO channel with PAPC was studied in [20], [21]. In particular, the author in [20], [21] proposed an iterative mode-dropping algorithm based on closed-form expressions to find the optimal input covariance. As shown later, this algorithm still requires high computational complexity and its convergence proof is not complete. Also, the mode-dropping algorithm assumes a full-rank channel which hardly holds true in various realistic conditions.

To the best of our knowledge, the only attempt to characterize the entire capacity region of the MIMO BC subject to PAPC was made in [22]. Specifically, the authors established

This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) and is co-funded under the European Regional Development Fund under Grant Number 13/RC/2077. This publication has also emanated from research supported in part by a Grant from Science Foundation Ireland under Grant number 17/CDA/4786.

Parts of this paper were presented at the IEEE Vehicular Technology Conference (Spring), Sydney, Australia, June 2017 [1], [2].

T. M. Pham and R. Farrell are with the Department of Electronic Engineering, Maynooth University, Co. Kildare, Ireland (email: minhthuy.pham@mu.ie, ronan.farrell@mu.ie).

L.-N. Tran is with the School of Electrical and Electronic Engineering, University College Dublin, Ireland. He was with the Department of Electronic Engineering, Maynooth University, Co. Kildare, Ireland. Email: nam.tran@ucd.ie.

a modified duality between the MAC and BC and transformed the input optimization problem in the BC into a minimax optimization problem in the corresponding MAC. Then the resulting program is solved by a standard barrier interior-point routine. Similarly, Tran *et al.* also proposed customized interior-point methods to study the achievable rate region of SZF-DPC in [18], [19]. However, the complexity of such second order optimization methods increases quadratically with the number of input dimensions, which is not practically appealing for large-scale antenna systems (also known as massive MIMO).

In this paper, we consider the problems of finding the capacity of various MIMO settings subject to PAPC, ranging from the SU-MU to MIMO BC. The goal is to arrive at closed-form design for all considered problems which are particularly useful for analyzing large-scale systems. In particular, our specific contributions include the following.

- For a SU-MIMO channel we proposed two fast-converging low-complexity iterative algorithms to compute the optimal input covariance matrices under PAPC. The first method is based on manipulating the optimality conditions of the considered problem and fixed-point iteration. The second one relies on the well-known MAC-BC duality but the resulting minimax problem is solved by a novel alternating optimization (AO) algorithm. Specifically, we proposed to optimize the upper bound of the objective with respect to a coordinate, eliminating the zigzag effect likely occurring in a pure AO method. Both proposed methods are provably convergent without any specific assumption on the channel matrix. Extensive analytical and numerical results are provided to demonstrate the superior performance of the proposed method, compared to the mode-dropping algorithm in [20], [21].
- We also characterize the entire capacity region of the MIMO BC, which was studied in [22]. For the MIMO BC, the weighted sum capacity is neither a concave nor convex function of the covariance matrices. Thus, the MAC-BC duality is invoked to obtain a convex formulation in the dual MAC, which is given in the form of a minimax optimization problem [22]. Instead of applying a standard interior-point method to find a saddle point of the resulting minimax program, we propose a closed-form design based on AO, similar to the case of SU-MIMO. The idea is to leverage the fact that the weighted sum capacity problem under a SPC can be solved by closed-form expressions in combination with a CGP method [10].

The remainder of the paper is organized as follows. The capacity of SU-MIMO is described in Section II. Section III derives closed-form expressions for the capacity region of a Gaussian MIMO BC while Section IV presents the numerical results. Finally, we conclude the paper in Section V.

Notation: Standard notations are used in this paper. Bold lower and upper case letters represent vectors and matrices, respectively. $\mathbf{I}$ defines an identity matrix, of which the size can be easily inferred from the context; $\mathbb{C}^{M\times N}$ denotes the space of $M \times N$ complex matrices; $\mathbf{H}^\dagger$ and $\mathbf{H}^T$ are Hermitian and normal transpose of $\mathbf{H}$, respectively; $\mathbf{H}_{i,j}$ is the (i,j)th entry of $\mathbf{H}$; $|\mathbf{H}|$ is the determinant of $\mathbf{H}$; $\mathrm{rank}(\mathbf{H})$ and $\mathrm{null}(\mathbf{H})$ stand for the rank of $\mathbf{H}$ and a basis of the null space of $\mathbf{H}$, respectively; $\mathrm{diag}(\mathbf{x})$ denotes the diagonal matrix with diagonal elements being $\mathbf{x}$; $\mathrm{diag}(\mathbf{H})$, where $\mathbf{H}$ is a square matrix, returns the vector of diagonal elements of $\mathbf{H}$. The notation $\mathbf{x}\odot\mathbf{y}$ denotes the Hadamard product (i.e., the entrywise product) of $\mathbf{x}$ and $\mathbf{y}$. The notations $\mathbf{A}\succeq\mathbf{B}$ and $\mathbf{A}\succ\mathbf{B}$ mean that $(\mathbf{A}-\mathbf{B})$ is a positive semidefinite and definite matrix, respectively. Furthermore, we denote $[x]_+ = \max(x,0)$, $\mathbf{0}_n$ and $\mathbf{1}_n$ to be a *row* vector of size n with all zeros and ones, respectively. The Euclidean and Frobenius norms are denoted by $||\cdot||_2$ and $||\cdot||_F$, respectively.

II. Capacity of SU-MIMO

A. System Model

Consider a SU-MIMO channel, where the transmitter is equipped with N antennas and the receiver with M antennas. The channel matrix is represented by $\mathbf{H}\in\mathbb{C}^{M\times N}$, which is assumed to be known perfectly at the transmitter. The received signal is given by

$$\mathbf{y} = \mathbf{H}\mathbf{s} + \mathbf{z} \tag{1}$$

where $\mathbf{s}$ is the vector of transmitted symbols of zero-mean, and $\mathbf{z}\in\mathbb{C}^{M\times 1}$ is the background noise with distribution $\mathcal{CN}(\mathbf{0},\mathbf{I}_M)$. Let $\mathbf{S}=E\{\mathbf{s}\mathbf{s}^\dagger\}$ be the input covariance matrix for the transmitted signal. We are interested in finding the capacity of the above channel with PAPC, which is formulated as

$$\underset{\mathbf{S}\succeq\mathbf{0}}{\text{maximize}} \quad \log|\mathbf{I}+\mathbf{H}\mathbf{S}\mathbf{H}^\dagger| \tag{2a}$$

$$\text{subject to} \quad [\mathbf{S}]_{i,i}\le P_i,\ i=1,2,\ldots,N \tag{2b}$$

where P_i is the maximum power constraint on the ith antenna. The problem (2) is a convex program, and can be solved by general-purpose optimization software.[1] However, the computational complexity of these convex solvers, which are usually based on interior-point methods, increases rapidly with the number of transmit antennas N, thereby not suitable for large-scale MIMO systems. Herein, we propose two efficient iterative algorithms which will be numerically shown to achieve a superlinear convergence rate.

B. Proposed Algorithms

1) Fixed-point Iteration

We first note that the Slater condition is satisfied for (2) and thus strong duality holds. Now, consider the partial Lagrangian function of (2), which is given by

$$\mathcal{L}(\mathbf{S},\mathbf{\Lambda}) = \log|\mathbf{I}+\mathbf{H}\mathbf{S}\mathbf{H}^\dagger| - \mathrm{tr}(\mathbf{\Lambda}(\mathbf{S}-\mathbf{P})) \tag{3}$$

where $\mathbf{P} = \mathrm{diag}(P_1,P_2,\ldots,P_N)$, and $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1,\lambda_2,\ldots,\lambda_N)$ is the diagonal matrix comprising the dual variables for the N power constraints in (2b). The dual objective of (2) is

$$g(\mathbf{\Lambda}) = \max_{\mathbf{S}\succeq\mathbf{0}} \mathcal{L}(\mathbf{S},\mathbf{\Lambda}). \tag{4}$$

[1]More specifically, (2) in the current form is in fact a MAXDET program [23] but can be reformulated as a semidefinite program (SDP) [24, p. 149] for which dedicated solvers are more available.

To find the optimal solution of (2), we only need to consider the case where $\mathbf{\Lambda} \succ \mathbf{0}$, i.e, $\lambda_i > 0$ for all i, otherwise $g(\mathbf{\Lambda})$ is unbounded above, which cannot be the dual optimal of (2). This can be easily seen by contradiction. Suppose $\lambda_i = 0$ for some i. Then create a diagonal matrix $\mathbf{S} = \mathrm{diag}([0, \ldots, 0, \alpha_i, 0, \ldots, 0]^T)$. Accordingly, we can check that $\mathcal{L}(\mathbf{S}, \mathbf{\Lambda}) = \log(1 + \alpha_i \sum_{j=1}^{M} |\mathbf{H}_{j,i}|^2) \to \infty$ if $\alpha_i \to \infty$. Moreover, for a given $\mathbf{\Lambda} \succ \mathbf{0}$, we can solve (4) efficiently as described next. Let us denote $\hat{\mathbf{S}} = \mathbf{\Lambda}^{1/2}\mathbf{S}\mathbf{\Lambda}^{1/2}$. Then finding $\mathbf{S}$ to maximize $\mathcal{L}(\mathbf{S}, \mathbf{\Lambda})$ amounts to solving the following problem

$$\underset{\hat{\mathbf{S}} \succeq \mathbf{0}}{\text{maximize}} \ \log|\mathbf{I} + \mathbf{H}\mathbf{\Lambda}^{-1/2}\hat{\mathbf{S}}\mathbf{\Lambda}^{-1/2}\mathbf{H}^\dagger| - \mathrm{tr}(\hat{\mathbf{S}}) \quad (5)$$

The above problem admits the solution based on water-filling algorithm with *fixed water level* [25]. Explicitly, let $\mathbf{V}\mathbf{\Sigma}\mathbf{V}^\dagger = \mathbf{\Lambda}^{-1/2}\mathbf{H}^\dagger\mathbf{H}\mathbf{\Lambda}^{-1/2}$ be the eigenvalue decomposition (EVD) of $\mathbf{\Lambda}^{-1/2}\mathbf{H}^\dagger\mathbf{H}\mathbf{\Lambda}^{-1/2}$, where $\mathbf{V} \in \mathbb{C}^{N\times N}$ are unitary matrix, and $\mathbf{\Sigma} \in \mathbb{C}^{N\times N}$ is a matrix of *(possibly zero) eigenvalues in decreasing order* of $\mathbf{\Lambda}^{-1/2}\mathbf{H}^\dagger\mathbf{H}\mathbf{\Lambda}^{-1/2}$. Let $r = \mathrm{rank}(\mathbf{H}\mathbf{\Lambda}^{-1/2})$, and ρ_i, $i = 1, \ldots r$, be r positive eigenvalues of $\mathbf{\Lambda}^{-1/2}\mathbf{H}^\dagger\mathbf{H}\mathbf{\Lambda}^{-1/2}$. Then, $\hat{\mathbf{S}}$ can be found as

$$\hat{\mathbf{S}} = \mathbf{V}\,\mathrm{diag}([1 - \frac{1}{\rho_1}]_+, \ldots, [1 - \frac{1}{\rho_r}]_+, \mathbf{0}_{N-r})\mathbf{V}^\dagger. \quad (6)$$

Consequently, $\mathbf{S}$ is given by

$$\mathbf{S} = \mathbf{\Lambda}^{-1/2}\mathbf{V}\big(\mathrm{diag}([1 - \frac{1}{\rho_1}]_+, \ldots, [1 - \frac{1}{\rho_r}]_+, \mathbf{0}_{N-r}\big)\mathbf{V}^\dagger\mathbf{\Lambda}^{-1/2}. \quad (7)$$

As a closer look at (7), let s be the largest number such that $1 - \frac{1}{\rho_s} > 0$. Then, $\mathbf{S}$ is equivalently written as

$$\mathbf{S} = \mathbf{\Lambda}^{-1/2}\mathbf{V}\,\mathrm{diag}(1 - \frac{1}{\rho_1}, \ldots, 1 - \frac{1}{\rho_s}, \mathbf{0}_{N-s})\mathbf{V}^\dagger\mathbf{\Lambda}^{-1/2}. \quad (8)$$

Since $\mathbf{V}\mathbf{V}^\dagger = \mathbf{I}$, $\mathbf{S}$ is further simplified as

$$\mathbf{S} = \mathbf{\Lambda}^{-1} - \mathbf{\Lambda}^{-1/2}\mathbf{V}\,\mathrm{diag}(\frac{1}{\rho_1}, \ldots, \frac{1}{\rho_s}, \mathbf{1}_{N-s})\mathbf{V}^\dagger\mathbf{\Lambda}^{-1/2}. \quad (9)$$

We can prove that at the optimum, $[\mathbf{S}]_{i,i} = P_i$ for all $i = 1, \ldots, N$. Thus, in order to find optimal $\mathbf{S}$, we need to find $\mathbf{\Lambda}$ such that

$$\left[\mathbf{\Lambda}^{-1} - \mathbf{\Lambda}^{-1/2}\mathbf{V}\,\mathrm{diag}(\frac{1}{\rho_1}, \ldots, \frac{1}{\rho_s}, \mathbf{1}_{N-s})\mathbf{V}^\dagger\mathbf{\Lambda}^{-1/2}\right]_{i,i} = P_i. \quad (10)$$

Since $\mathbf{\Lambda}$ is a diagonal matrix, (10) equals to

$$\Big(\mathbf{I} - \big[\mathbf{V}\big(\mathrm{diag}(\frac{1}{\rho_1}, \ldots, \frac{1}{\rho_s}, \mathbf{1}_{N-s}\big)\mathbf{V}^\dagger\big]_{i,i}\Big)\big[\mathbf{\Lambda}^{-1}\big]_{i,i} = P_i. \quad (11)$$

Let $\mathbf{\Psi}(\tilde{\mathbf{\lambda}}) = \big[\mathbf{V}\big(\mathrm{diag}(\frac{1}{\rho_1}, \ldots, \frac{1}{\rho_s}, \mathbf{1}_{N-s}\big)\mathbf{V}^\dagger\big]$. Then, we can rewrite (11) in the form of a nonlinear system as

$$\tilde{\mathbf{\lambda}} - \mathrm{diag}(\mathbf{\Psi}(\tilde{\mathbf{\lambda}})) \odot \tilde{\mathbf{\lambda}} = \mathbf{p} \quad (12)$$

where $\tilde{\mathbf{\lambda}} \triangleq [\lambda_1^{-1}, \lambda_2^{-1}, \ldots, \lambda_N^{-1}]^T$, $\mathbf{p} \triangleq [P_1, P_2, \ldots, P_N]^T$. It is easy to see that

$$\Big[\mathbf{\Psi}(\tilde{\mathbf{\lambda}})\Big]_{i,i} = \sum\nolimits_{j=1}^{N} \tilde{\rho}_j |v_{i,j}|^2 \quad (13)$$

where $\tilde{\rho}_j = \frac{1}{\rho_j} < 1$ for $1 \le j \le s$, and $\tilde{\rho}_j = 1$ for $s < j \le N$. Since $\sum_{j=1}^{N} |v_{i,j}|^2 = 1$, it holds that $\mathbf{\Psi}(\tilde{\mathbf{\lambda}}) \prec \mathbf{I}$ for all $\tilde{\mathbf{\lambda}} \succ \mathbf{0}$, and (12) is thus well defined. Unfortunately, there is no analytical solution to (12), mostly due to the fact that $\mathbf{\Psi}(\tilde{\mathbf{\lambda}})$ is a nonlinear function of $\tilde{\mathbf{\lambda}}$. However, (12) already suggests a way to find $\tilde{\mathbf{\lambda}}$ iteratively as follows

$$\tilde{\mathbf{\lambda}}_{n+1} = \mathbf{p} + \mathrm{diag}(\mathbf{\Psi}(\tilde{\mathbf{\lambda}}_n)) \odot \tilde{\mathbf{\lambda}}_n \triangleq \mathfrak{I}(\tilde{\mathbf{\lambda}}_n). \quad (14)$$

In fact, (14) is written in a fixed-point iteration form and its convergence is stated in the following lemma.

Lemma 1. *The iterations in* (14) *converge to the unique fixed-point of* (12)*, thereby solving* (2)*.*

The proof of Lemma 1 is provided in Appendix A. The key is to show that $\mathfrak{I}(\mathbf{x})$ is a standard interference function.

We can see that the fixed-point algorithm based on (14) requires iteratively performing EVD of $\mathbf{\Lambda}^{-1/2}\mathbf{H}^\dagger\mathbf{H}\mathbf{\Lambda}^{-1/2}$. A simple way is to treat it as a new matrix at each iteration, but this is not computationally efficient. Exploiting the fact that the channel matrix $\mathbf{H}$ remains the same during the whole iterative process, we present a way to compute the EVD of $\mathbf{\Lambda}^{-1/2}\mathbf{H}^\dagger\mathbf{H}\mathbf{\Lambda}^{-1/2}$ more efficiently. To this end, let $\mathbf{H} = \mathbf{G}\mathbf{R}$, where $\mathbf{G}$ is unitary and $\mathbf{R}$ is upper triangular, be a QR factorization of $\mathbf{H}$. Then we can write $\mathbf{H}\mathbf{\Lambda}^{-1/2} = (\mathbf{G}\mathbf{R})\mathbf{\Lambda}^{-1/2} = \mathbf{G}(\mathbf{R}\mathbf{\Lambda}^{-1/2})$. Since $\mathbf{\Lambda}$ is diagonal, $\mathbf{R}\mathbf{\Lambda}^{-1/2}$ is also an upper triangular matrix. Now let $\mathbf{R}\mathbf{\Lambda}^{-1/2} = \mathbf{U}\bar{\mathbf{\Sigma}}\mathbf{V}^\dagger$ be the SVD of $\mathbf{R}\mathbf{\Lambda}^{-1/2}$. Then the EVD of $\mathbf{\Lambda}^{-1/2}\mathbf{H}^\dagger\mathbf{H}\mathbf{\Lambda}^{-1/2}$ is simply given by $\mathbf{V}\bar{\mathbf{\Sigma}}^2\mathbf{V}^\dagger$. We remark that SVD computation for an upper triangular matrix is much cheaper than for a full matrix [26, p. 492], which leads to a huge reduction in the computation cost of the proposed algorithm. The proposed algorithm based on fixed-point iteration is outlined in Algorithm 1.

Algorithm 1 Proposed solution based on fixed-point iteration.

Input: $\mathbf{\Lambda}_0$ diagonal matrix of positive elements, $\epsilon > 0$.

1: Set $n := 0$ and $\tau = 1 + \epsilon$.
2: Perform QR decomposition of $\mathbf{H}$: $\mathbf{H} = \mathbf{G}\mathbf{R}$, where $\mathbf{G}$ is a unitary matrix and $\mathbf{R}$ is an upper triangular matrix.
3: **while** $\tau > \epsilon$ **do**
4: Perform the SVD of $\mathbf{R}{\mathbf{\Lambda}_n}^{-1/2}$: $\mathbf{R}{\mathbf{\Lambda}_n}^{-1/2} = \mathbf{U}_n\bar{\mathbf{\Sigma}}_n\mathbf{V}_n^\dagger$, where $\bar{\mathbf{\Sigma}}_n$ is diagonal. Let $\rho_i = \sigma_i^2$, $i = 1, \ldots, r$, where σ_i is the ith non-zero entry of $\bar{\mathbf{\Sigma}}_n$ and $r = \mathrm{rank}(\mathbf{R}{\mathbf{\Lambda}_n}^{-1/2})$.
5: $\tilde{\mathbf{\Sigma}}_n := \mathrm{diag}([1 - \rho_1^{-1}]_+, ..., [1 - \rho_r^{-1}]_+, \mathbf{0}_{N-r})$.
6: $\mathbf{\Psi}_n := \mathbf{V}_n(\mathbf{I} - \tilde{\mathbf{\Sigma}}_n)\mathbf{V}_n^\dagger$.
7: $\mathbf{S}_n := \mathbf{\Lambda}_n^{-1} - \mathbf{\Lambda}_n^{-1/2}\mathbf{\Psi}_n\mathbf{\Lambda}_n^{-1/2}$.
8: $\tau = \sum_{i=1}^{N}[\mathbf{\Lambda}_n]_{i,i}|[\mathbf{S}_n - \mathbf{P}]|_{i,i}$.
9: $\tilde{\mathbf{\lambda}}_{n+1} = \mathbf{p} + \mathrm{diag}(\mathbf{\Psi}_n) \odot \tilde{\mathbf{\lambda}}_n$.
10: $\mathbf{\Lambda}_{n+1} = (\mathrm{diag}\,\tilde{\mathbf{\lambda}}_{n+1})^{-1}$.
11: $n := n + 1$.
12: **end while**

Output: $\mathbf{S}_n$.

Remark 1. To solve (2), the work of [20], [21] proposed two different algorithms for two corresponding cases: $M \ge N$ and $M < N$. Moreover, these algorithms are dedicated to full-rank channel matrices. In this regard, Algorithm 1 is more universal in the sense that it is applicable to channel matrices of any dimension and rank-deficiency. Another issue

of the methods presented in [20], [21] is that a complete analytical proof of their convergence is sill missing. On the contrary, Algorithm 1 is provably convergent from an arbitrary starting point $\tilde{\boldsymbol{\lambda}}_0 > 0$. Moreover, analytical and numerical results demonstrate Algorithm 1 achieves lower complexity, compared to the ones in [20], [21].

2) Alternating Optimization

The second proposed iterative method exploits an interesting result from the duality between BC and MAC [27], [28]. In fact it is shown that (2) is equivalent to the following minimax optimization problem [22]

$$\begin{array}{ll} \min\limits_{\mathbf{Q}\succeq\mathbf{0}} \max\limits_{\bar{\mathbf{S}}\succeq\mathbf{0}} & \log\frac{|\mathbf{Q}+\mathbf{H}^\dagger\bar{\mathbf{S}}\mathbf{H}|}{|\mathbf{Q}|} \triangleq f(\mathbf{Q},\bar{\mathbf{S}}) \\ \text{subject to} & \mathrm{tr}(\bar{\mathbf{S}}) \le P, \mathrm{tr}(\mathbf{QP}) \le P; \mathbf{Q}: \text{diagonal} \end{array} \tag{15}$$

where $P \triangleq \sum_{i=1}^N P_i$. In the above formulation we define $\log|\mathbf{Q}| = -\infty$ if $\mathbf{Q}$ is singular. For the development of the second proposed method, without loss of optimality, we assume that $||\mathbf{h}_i||_2 > 0$ where $\mathbf{h}_i$ is the ith column of the channel matrix $\mathbf{H}$, which is normally the case in practice. If $\mathbf{h}_i$ happens to be all-zero vector, the ith transmit antenna can be dropped to obtain a reduced channel matrix, to which the following proposed method is applied. As a result of Appendix B, the relationship between (2) and (15) is stated in the following fact.

Fact 1. *There exists a saddle point* $(\bar{\mathbf{S}}^\star, \mathbf{Q}^\star)$ *for* (15) *such that* $\mathbf{Q}^\star \succ \mathbf{0}$. *Denote* $\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\dagger$ *to be an SVD of* $\mathbf{H}(\mathbf{Q}^\star)^{-1/2}$ *where* $\boldsymbol{\Sigma}$ *is square and diagonal. Then, the optimal solution* $\mathbf{S}^\star$ *to* (2) *can be found as*

$$\mathbf{S}^\star = (\mathbf{Q}^\star)^{-1/2}\mathbf{V}\mathbf{U}^\dagger\bar{\mathbf{S}}^\star\mathbf{U}\mathbf{V}^\dagger(\mathbf{Q}^\star)^{-1/2}. \tag{16}$$

The above result is in fact a special case of the MAC-to-BC transformation presented in [28] when applying to a single user system.

It is trivial to see that the optimality of (15) is not affected if the inequalities are made to be equality. To appreciate the idea behind the second proposed method, let us define $\mathcal{Q} \triangleq \{\mathbf{Q}|\mathbf{Q}: \text{diagonal}, \mathbf{Q}\succeq\mathbf{0}, \mathrm{tr}(\mathbf{QP}) = P\}$ and $\mathcal{S} = \{\bar{\mathbf{S}}|\bar{\mathbf{S}}\succeq\mathbf{0}, \mathrm{tr}(\bar{\mathbf{S}}) = P\}$. Now, (15) can be rewritten in an abstract form as

$$\min_{\mathbf{Q}\in\mathcal{Q}} \max_{\bar{\mathbf{S}}\in\mathcal{S}} f(\mathbf{Q},\bar{\mathbf{S}}). \tag{17}$$

We note that $f(\mathbf{Q},\bar{\mathbf{S}})$ is concave with $\bar{\mathbf{S}}$, and convex with $\mathbf{Q}$, and twice differentiable. Thus a saddle point $(\mathbf{Q}^*, \bar{\mathbf{S}}^*)$ exist for (17) and it holds that

$$f(\mathbf{Q}^*,\bar{\mathbf{S}}) \le f(\mathbf{Q}^*,\bar{\mathbf{S}}^*) \le f(\mathbf{Q},\bar{\mathbf{S}}^*). \tag{18}$$

We can see that solving (15) boils down to finding a saddle point for (17). In fact, this interpretation was used in the interior-point method proposed in [22]. The minimax formulation in (17) also suggests a way to find a saddle point by alternatively optimizing $\mathbf{Q}$ and $\bar{\mathbf{S}}$. This method was also mentioned in [22] but note that it is not provably convergent. In fact we have very often observed that this pure method will suffer a ping-pong effect, and thus fail to converge to an optimal solution of (17) (cf. Fig. 2 for an example on this).

In the second proposed algorithm, we still capitalize on the idea of AO, but do it in a novel way to ensure strict monotonicity. Suppose at the nth iteration, we have obtained $\mathbf{Q}_n$. Then $\bar{\mathbf{S}}_n$ is found as the solution to the following problem

$$\begin{array}{ll} \text{maximize} & \log|\mathbf{Q}_n + \mathbf{H}^\dagger\bar{\mathbf{S}}\mathbf{H}| \\ \text{subject to} & \mathrm{tr}(\bar{\mathbf{S}}) = P; \bar{\mathbf{S}} \succeq \mathbf{0}. \end{array} \tag{19}$$

It is well known that the above problem admits the solution based on water-filling algorithm [3], [4]. More explicitly, let $\mathbf{U}_n\boldsymbol{\Sigma}_n\mathbf{U}_n^\dagger = \mathbf{H}\mathbf{Q}_n^{-1}\mathbf{H}^\dagger$ be the EVD of $\mathbf{H}\mathbf{Q}_n^{-1}\mathbf{H}^\dagger$, where $\boldsymbol{\Sigma}_n = \mathrm{diag}(\rho_1, \ \rho_2, \ldots, \rho_r)$ is a matrix of non-negative eigenvalues of $\mathbf{H}\mathbf{Q}_n^{-1}\mathbf{H}^\dagger$, and $r = \mathrm{rank}(\mathbf{H}\mathbf{Q}_n^{-1/2})$. Then, $\bar{\mathbf{S}}_n$ can be found as

$$\bar{\mathbf{S}}_n = \mathbf{U}_n\hat{\boldsymbol{\Sigma}}_n\mathbf{U}_n^\dagger \tag{20}$$

where $\hat{\boldsymbol{\Sigma}}_n = \mathrm{diag}([\mu - \frac{1}{\rho_1}]_+, [\mu - \frac{1}{\rho_2}]_+, \ldots, [\mu - \frac{1}{\rho_r}]_+)$ and μ is the water-level, which is chosen to satisfy the total power constraint

$$\sum_{i=1}^r [\mu - \frac{1}{\rho_i}]_+ = P. \tag{21}$$

Note that $\bar{\mathbf{S}}_n$ in (20) is the *unique* solution to (19). To find $\mathbf{Q}_{n+1}$, we invoke the following inequality, which results from the concavity of the logdet function,

$$\log|\mathbf{Q} + \mathbf{H}^\dagger\bar{\mathbf{S}}_n\mathbf{H}| \le \log|\boldsymbol{\Phi}_n| + \mathrm{tr}\Big(\boldsymbol{\Phi}_n^{-1}\big(\mathbf{Q} - \mathbf{Q}_n\big)\Big) \tag{22}$$

where $\boldsymbol{\Phi}_n = \mathbf{Q}_n + \mathbf{H}^\dagger\bar{\mathbf{S}}_n\mathbf{H}$. In the second proposed algorithm, $\mathbf{Q}_{n+1}$ is found to optimize the upper bound of (15), i.e., $\mathbf{Q}_{n+1}$ is the solution to the following problem

$$\begin{array}{ll} \underset{\mathbf{Q}\succeq\mathbf{0}}{\text{minimize}} & \mathrm{tr}\Big(\boldsymbol{\Phi}_n^{-1}\mathbf{Q}\Big) - \log|\mathbf{Q}| \\ \text{subject to} & \mathrm{tr}(\mathbf{QP}) = P; \mathbf{Q}: \text{diagonal}. \end{array} \tag{23}$$

We will see shortly that in the second proposed iterative algorithm, $\mathbf{Q}_n \succ \mathbf{0}$ for all iterations n, and thus $\boldsymbol{\Phi}_n^{-1}$ is well defined. Also, it is worth mentioning that the gradient of the objective in (23) with respect to $\mathbf{Q}$ is identical to that of the original objective in (15) when $\mathbf{Q} = \mathbf{Q}_n$. This is essentially to ensure that the first order optimality conditions of the original problem are preserved even with the use of an upper bound. To clarify this point let us write the partial derivative of $f(\mathbf{Q},\bar{\mathbf{S}}_n)$ with respect to q_i as

$$\partial_{\mathbf{q}_i} f(\mathbf{Q},\bar{\mathbf{S}}_n) = [(\mathbf{Q} + \mathbf{H}^\dagger\bar{\mathbf{S}}_n\mathbf{H})^{-1}]_{i,i} - q_i^{-1} \tag{24}$$

The partial derivative of the upper bound with respect to q_i obtained at iteration n of the proposed method is

$$[\boldsymbol{\Phi}_n^{-1}]_{i,i} - q_i^{-1} = [(\mathbf{Q}_n + \mathbf{H}^\dagger\bar{\mathbf{S}}_n\mathbf{H})^{-1}]_{i,i} - q_i^{-1} \tag{25}$$

It is now clear that the partial derivatives of the original objective and the upper bound in (23) with respect to any q_i are the same when $\mathbf{Q} = \mathbf{Q}_n$. Thus when the iterates $\{(\mathbf{Q}_n, \bar{\mathbf{S}}_n)\}$ converge, they will satisfy the KKT conditions of the original problem, though an upper bound of the objective is minimized.

Before proceeding further we provide a remark regarding the use of the upper bound in (23) for updating $\mathbf{Q}$. First it is a well-known fact that if we simply alternate optimization of $\mathbf{Q}$ and $\bar{\mathbf{S}}$ as done in a pure AO method, then convergence to a saddle point is not guaranteed as monotonic convergence of the

objective is not achieved. In the second proposed algorithm, the key point is to make the objective decrease after each cyclic update of $\mathbf{Q}$ and $\bar{\mathbf{S}}$. For this purpose we minimize an upper bound of the objective for updating $\mathbf{Q}$. In fact, this idea is largely inspired by successive convex approximation (SCA) principle for nonconvex optimization problems [29]. Roughly speaking, for SCA-based methods, the nonconvex objective is approximated by a convex upper bound in each iteration, which ensures monotonic decrease of the sequence of the objectives. However, the main challenge is that SCA only concerns minimization (or equivalently maximization) problems, while our considered problem is a minimax program. As such the proof for the convergence of SCA-based algorithms is not applicable to Algorithm 2, as shown in Appendix B.

Since $\mathbf{Q}$ in (23) is in fact diagonal, i.e., $\mathbf{Q} = \mathrm{diag}(\mathbf{q})$, we can rewrite (23) as

$$\begin{array}{ll} \underset{\mathbf{q}\geq 0}{\text{minimize}} & \sum_{i=1}^{N} \phi_{n,i} q_i - \log q_i \\ \text{subject to} & \sum_{i=1}^{N} P_i q_i = P \end{array} \tag{26}$$

where $\phi_{n,i} = \left[\mathbf{\Phi}_n^{-1}\right]_{i,i}$. Interestingly, the above problem also has a water-filling-like solution as

$$q_i = \frac{1}{\phi_{n,i} + \gamma P_i} > 0 \tag{27}$$

where $\gamma \geq 0$ is the solution of the equation

$$\sum_{i=1}^{N} \frac{P_i}{\phi_{n,i} + \gamma P_i} = P. \tag{28}$$

From (27) it is clear that $\mathbf{Q}_n \succ \mathbf{0}$ for all n and thus $\mathbf{\Phi}_n^{-1}$ exists as mentioned below (23). Further, from the definition of $\mathbf{\Phi}_n$, it holds that $\phi_{n,i} = \left[\mathbf{\Phi}_n^{-1}\right]_{i,i} \leq \left[\mathbf{Q}_n^{-1}\right]_{i,i}$. As the result, we obtain $\sum_{i=1}^{N} \frac{P_i}{\phi_{n,i}} \geq \mathrm{tr}(\mathbf{Q}_n \mathbf{P}) = P$, where the equality holds since $\mathbf{Q}_n$ is the solution to (23) in the previous iteration. Note that the left hand side of (28) is decreasing with γ, and thus (28) always has a unique solution, which can be found efficiently, e.g., by the bisection or Newton method. The second proposed algorithm based on AO is summarized in Algorithm 2. The main point of Algorithm 2 is the use of the inequality in (22) to optimize $\mathbf{Q}$ for a given $\bar{\mathbf{S}}$. This step will eliminate the ping-pong effect mentioned above and ensure the objective sequence is strictly decreasing. The convergence proof of Algorithm 2 is provided in Appendix B.

We note that the error tolerance τ in line 4 of Algorithm 2 is only computed for $n \geq 1$. We remark that line 3 in Algorithm 2 involves the EVD of $\mathbf{H}\mathbf{Q}_n^{-1}\mathbf{H}^\dagger$, which can be computed similarly as done in Algorithm 1 to reduce the overall complexity. Specifically let $\mathbf{G}\mathbf{R} = \mathbf{H}$ be the QR decomposition of $\mathbf{H}$. Next we compute the SVD of the upper triangular matrix $\mathbf{R}\mathbf{Q}_n^{-1/2}$ as $\tilde{\mathbf{U}}_n \tilde{\mathbf{\Sigma}}_n \tilde{\mathbf{V}}_n^\dagger = \mathbf{R}\mathbf{Q}_n^{-1/2}$. Then the EVD of $\mathbf{H}\mathbf{Q}_n^{-1}\mathbf{H}^\dagger$ is simply given by $\mathbf{U}_n \mathbf{\Sigma}_n \mathbf{U}_n^\dagger = \mathbf{H}\mathbf{Q}_n^{-1}\mathbf{H}^\dagger$, where $\mathbf{U}_n = \mathbf{G}\tilde{\mathbf{U}}_n$ and $\mathbf{\Sigma}_n = \tilde{\mathbf{\Sigma}}_n^2$. Moreover, we note that $\bar{\mathbf{S}}_n$ needs not be computed explicitly as in (20) for each iteration. The reason is that the diagonal elements of $\mathbf{\Phi}_n^{-1}$ in line 5 can be found efficiently from the SVD of $\mathbf{R}\mathbf{Q}_n^{-1/2}$ as shown in the following.

Algorithm 2 Proposed solution based on alternating optimization.

Input: $\mathbf{Q}_0$ is feasible to $\mathcal{Q}$, and $\epsilon > 0$.

1: Initialize $n := 0$, $\tau = 1 + \epsilon$.
2: **while** $\tau > \epsilon$ **do**
3: Apply water-filling algorithm (i.e., (20) and (21)) to compute $\bar{\mathbf{S}}_n = \underset{\bar{\mathbf{S}}\in\mathcal{S}}{\arg\max}\ \log|\mathbf{Q}_n + \mathbf{H}^\dagger \bar{\mathbf{S}} \mathbf{H}|$.
4: For $n \geq 1$, let $\tau = |f(\mathbf{Q}_n, \bar{\mathbf{S}}_n) - f(\mathbf{Q}_{n-1}, \bar{\mathbf{S}}_{n-1})|$.
5: $\mathbf{\Phi}_n^{-1} := (\mathbf{Q}_n + \mathbf{H}^\dagger \bar{\mathbf{S}}_n \mathbf{H})^{-1}$.
6: Find $\mathbf{Q}_{n+1} = \underset{\mathbf{Q}\in\mathcal{Q}}{\arg\min}\ \mathrm{tr}\big(\mathbf{\Phi}_n^{-1}\mathbf{Q}\big) - \log|\mathbf{Q}|$, using (27) and (28).
7: $n := n + 1$.
8: **end while**

Output: $\bar{\mathbf{S}}_n$ and use (16) to compute optimal $\mathbf{S}$.

Using the matrix-inversion lemma, we can write $\mathbf{\Phi}_n^{-1} = \mathbf{Q}_n^{-1/2}(\mathbf{I} + \mathbf{Q}_n^{-1/2}\mathbf{H}^\dagger \bar{\mathbf{S}}_n \mathbf{H}\mathbf{Q}_n^{-1/2})^{-1}\mathbf{Q}_n^{-1/2} = \mathbf{Q}_n^{-1/2}(\mathbf{I} + \tilde{\mathbf{V}}_n \hat{\mathbf{\Sigma}}_n \tilde{\mathbf{V}}_n^\dagger)^{-1}\mathbf{Q}_n^{-1/2}$, where the latter equality holds due to (20). Now let $\dot{\mathbf{\Sigma}}_n$ be the diagonal matrix containing all *strictly positive* entries of $\hat{\mathbf{\Sigma}}_n$, and $\dot{\mathbf{V}}_n$ be the corresponding singular vectors. Then we can write $(\mathbf{I} + \tilde{\mathbf{V}}_n \hat{\mathbf{\Sigma}}_n \tilde{\mathbf{V}}_n^\dagger)^{-1} = (\mathbf{I} + \dot{\mathbf{V}}_n \dot{\mathbf{\Sigma}}_n \dot{\mathbf{V}}_n^\dagger)^{-1} \overset{(a)}{=} \mathbf{I} - \dot{\mathbf{V}}_n\big(\dot{\mathbf{\Sigma}}_n^{-1} + \dot{\mathbf{V}}_n^\dagger \dot{\mathbf{V}}_n\big)^{-1}\dot{\mathbf{V}}_n^\dagger \overset{(b)}{=} \mathbf{I} - \dot{\mathbf{V}}_n\big(\dot{\mathbf{\Sigma}}_n^{-1} + \mathbf{I}\big)^{-1}\dot{\mathbf{V}}_n^\dagger$, where (a) is due to the matrix inversion lemma, and (b) holds true since $\dot{\mathbf{V}}_n^\dagger \dot{\mathbf{V}}_n = \mathbf{I}$. In summary, we have $\mathbf{\Phi}_n^{-1} = \mathbf{Q}_n^{-1} - \mathbf{Q}_n^{-1/2}\dot{\mathbf{V}}_n\big(\dot{\mathbf{\Sigma}}_n^{-1} + \mathbf{I}\big)^{-1}\dot{\mathbf{V}}_n^\dagger \mathbf{Q}_n^{-1/2}$. Since $\dot{\mathbf{\Sigma}}_n^{-1} + \mathbf{I}$ is diagonal, its inversion can be computed easily. It is also clear that, to compute $\mathbf{\Phi}_n^{-1}$, what we need is only $\tilde{\mathbf{\Sigma}}_n$ and $\tilde{\mathbf{V}}_n$ from the SVD of $\mathbf{R}\mathbf{Q}_n^{-1/2}$.

C. Complexity Analysis

In this section, we analyze the complexity of the proposed algorithms in the preceding section, counted as the number of flops. Although flop counting is a crude way to measure the actual computational complexity, it somewhat captures the order of the computation load. To this end we first assume $M \geq N$ (i.e, more receive than transmit antennas) and summarize the relevant results presented in [26] and [30] as follows. QR decomposition of an $M \times N$ matrix using Householder transformation requires $2N^2(M - N/3)$ flops for only $\mathbf{R}$, and $4M^2N - 2MN^2 + \frac{2}{3}N^3$ flops for both $\mathbf{R}$ and $\mathbf{Q}$. The computation of SVD of a full $M \times N$ matrix needs $4M^2N + 8MN^2 + 9N^3$ flops for $(\mathbf{\Sigma}, \mathbf{V}, \mathbf{U})$, $4MN^2 + 8N^3$ flops for $(\mathbf{\Sigma}, \mathbf{V})$, and $4M^2N - 8MN^2$ flops for $(\mathbf{\Sigma}, \mathbf{U})$ while that of an upper triangular matrix requires $4M^2N + 22N^3, 2M^2N + 11N^3, 4M^2N + 13N^3$, respectively. The number of flops for the water-filling algorithm with N eigenmodes is $2N^2 + 6N$. Inversion of an $N \times N$ symmetric matrix requires N^3 flops. Note that these flop counts are for a real matrix. For complex matrices, we simply treat every operation as a complex multiplication which is equal to 6 real flops [30], [31]. That is, QR decomposition of an $M \times N$ complex matrix requires $4N^2(3M - N)$ flops.

In the complexity analysis presented in the following, we only consider the main operations having the most significant complexity and ignore those contributing negligibly to the

overall complexity (e.g., subtraction or addition).

1) Complexity of Algorithm 1

Algorithm 1 performs a QR decomposition (cf. line 2) at the first iteration and only $\mathbf{R}$ is needed, which requires $4N^2(3M-N)$ flops as explained above. In the subsequent iterations, Algorithm 1 involves an SVD of an upper triangular matrix (line 4), in which only $(\mathbf{\Sigma}, \mathbf{V})$ needs to be computed. This step takes $6(2MN^2+11N^3)$ flops. We note that other operations in Algorithm 1 have minor complexity, compared to QR decomposition and SVD, and thus are neglected.

2) Complexity of Algorithm 2

To reduce the complexity, Algorithm 2 performs a full QR decomposition in the first iteration, which takes $6(4M^2N - 2MN^2 + \frac{2}{3}N^3)$ flops. Then, the complexity incurred in line 3 of Algorithm 2 is due to finding $(\tilde{\mathbf{\Sigma}}_n, \tilde{\mathbf{V}}_n^\dagger)$ in the SVD of the upper triangular matrix $\mathbf{R}\mathbf{Q}_n^{-1/2}$. The flop count of the step is $6(2M^2N+11N^3)$. The water-filling algorithm to find positive eigenmodes that meet the sum power constraint needs $6(2N^2+6N)$ flops. The complexity of line 5 (i.e., computing the diagonal elements of $\mathbf{\Phi}_n^{-1}$) and that of line 6 are lower compared to the remaining steps and thus can be ignored.

3) Complexity of the mode-dropping algorithm in [20], [21]

For comparison purpose we now present the complexity of the so-called mode-dropping algorithm proposed in [20], [21]. Specifically, this method requires an SVD of a *full* $M \times N$ matrix, in the first iteration, for which the flop count is $6(4M^2N+8MN^2+9N^3)$. From the second iteration, the most complex operation of the mode-dropping algorithm is to compute an EVD which requires $6(4MN^2+8N^3)$ flops.

Basically, the complexity of the proposed algorithms for the case $N > M$ can be obtained by simply switching N and M in the above analytical expressions. However, for the mode-dropping algorithm, two additional matrix inversions need to be performed, resulting in an increased complexity. The per-iteration complexity comparison (after the first iteration) is summarized in Table I, where the bold text refers to the algorithm with the lowest complexity, i.e., Algorithm 1 and 2. However, the total complexity of an iterative algorithm heavily depends on the number of iterations required to converge. This issue is evaluated for various numerical experiments in Section IV.

4) Complexity of interior-point methods

As mentioned earlier, problem (2) can be reformulated as an SDP and then solved by general-purpose optimization packages. These optimization tools are normally based on primal-dual path-following methods to solve a convex model. However, the per-iteration complexity of such interior-point solvers is $\mathcal{O}(N^6)$ [15], [32], which is prohibitively high for large-scale MIMO systems. A numerical complexity comparison is shown in Fig. 5 to further demonstrate this point.

III. Capacity Region of a Gaussian MIMO BC

In this scenario we compute the capacity region of a MIMO BC. It was proved in [5] that the capacity region of a Gaussian MIMO BC is achieved by DPC. For a SPC, this problem was addressed in a number of previous studies [9], [10], [37]. The related research for PAPC is quite limited. Specifically, the capacity region can be characterized by solving the following weighted sum rate maximization

$$\begin{array}{ll} \underset{\{\mathbf{S}_k \succeq \mathbf{0}\}}{\text{maximize}} & \sum_{k=1}^K w_k \log \frac{|\mathbf{I}+\mathbf{H}_k \sum_{i=1}^k \mathbf{S}_i \mathbf{H}_k^\dagger|}{|\mathbf{I}+\mathbf{H}_k \sum_{i=1}^{k-1} \mathbf{S}_i \mathbf{H}_k^\dagger|} \\ \text{subject to} & \sum_{k=1}^K [\mathbf{S}_k]_{i,i} \le P_i,\ \forall i \end{array} \tag{29}$$

for different sets of the weights w_i. Without loss of generality, we assume that $0 < w_1 \le w_2 \le ... \le w_K$ and $\sum_{k=1}^K w_k = 1$ in the following. Since (29) is nonconvex at hand, Algorithm 1 cannot be extended to solve it. Fortunately, it can be solved efficiently using the BC-MAC duality and alternating optimization as shown next. First, by the BC-MAC duality [22], (29) is equivalent to

$$\begin{array}{ll} \min\limits_{\mathbf{Q}\succ\mathbf{0}} \max\limits_{\{\bar{\mathbf{S}}_k \succeq \mathbf{0}\}} & \sum_{k=1}^K \Delta_k \log|\mathbf{Q}+\sum_{i=k}^K \mathbf{H}_i^\dagger \bar{\mathbf{S}}_i \mathbf{H}_i| \\ & -w_K \log|\mathbf{Q}| \\ \text{subject to} & \sum_{k=1}^K \text{tr}(\bar{\mathbf{S}}_k) = P, \text{tr}(\mathbf{Q}\mathbf{P}) = P, \mathbf{Q}: \text{diagonal} \end{array} \tag{30}$$

where $\Delta_k = w_k - w_{k-1}$. Before proceeding further we remark for the same problem, an interior-point algorithm was proposed in [22]. The complexity of such a method does not scale favorably with the problem size, compared to our proposed approach presented in what follows, which is based on closed-form expressions.

Let $(\{\bar{\mathbf{S}}_k^n\}, \mathbf{Q}^n)$ denote the value of $(\{\bar{\mathbf{S}}_k\}, \mathbf{Q})$ after n iterations of the proposed method. In view of AO, $\{\bar{\mathbf{S}}_k^n\}$ is the solution to the following problem

$$\begin{array}{ll} \text{maximize} & \sum_{k=1}^K \Delta_k \log|\mathbf{Q}^n+\sum_{i=k}^K \mathbf{H}_i^\dagger \bar{\mathbf{S}}_i \mathbf{H}_i| \\ \text{subject to} & \sum_{k=1}^K \text{tr}(\bar{\mathbf{S}}_k) = P; \{\bar{\mathbf{S}}_k \succeq \mathbf{0}\}. \end{array} \tag{31}$$

Problem (31) can be solved by off-the-shelf convex solvers but it can be solved more efficiently by a CGP method. The motivation is that projection onto the feasible set of (31) can be reduced to projection onto a canonical simplex, as shown in Appendix C. Thus, a CGP method can be derived to find the optimal solution of (31) (The details are skipped due to the space limitation). We note that similar approaches were also presented in [10], [11].

For the important case of the sum capacity of the MIMO BC, (31) becomes

$$\begin{array}{ll} \text{maximize} & \log|\mathbf{Q}^n+\sum_{i=1}^K \mathbf{H}_i^\dagger \bar{\mathbf{S}}_i \mathbf{H}_i| \\ \text{subject to} & \sum_{k=1}^K \text{tr}(\bar{\mathbf{S}}_k) = P; \{\bar{\mathbf{S}}_k \succeq \mathbf{0}\}. \end{array} \tag{32}$$

For the above specific problem, the sum power iterative water-filling algorithm proposed in [9] and dual decomposition based method in [25] are particularly efficient.

TABLE I
Per-iteration complexity comparison

Algorithms	$M \ge N$	$M < N$
Mode-dropping [20], [21]	$6(4MN^2+8N^3)$	$6(4NM^2+8M^3)$ $+12(N-M)^3$
Algorithm 1	$\mathbf{6(2MN^2+11N^3)}$	$\mathbf{6(2NM^2+11M^3)}$
Algorithm 2	$\mathbf{6(2MN^2+11N^3)}$	$\mathbf{6(2NM^2+11M^3)}$

We now turn our attention to finding $\mathbf{Q}^{n+1}$, which can be done exactly the same as for the SU-MIMO case. Specifically, it holds that

$$\log|\mathbf{Q}+\sum_{i=k}^{K}\mathbf{H}_i^{\dagger}\bar{\mathbf{S}}_i^n\mathbf{H}_i| \leq \log|\mathbf{\Phi}_k^n|+\mathrm{tr}\big(\mathbf{\Phi}_k^{-n}\big(\mathbf{Q}-\mathbf{Q}^n\big)\big) \tag{33}$$

where $\mathbf{\Phi}_k^n \triangleq \mathbf{Q}^n+\sum_{i=k}^{K}\mathbf{H}_i^{\dagger}\bar{\mathbf{S}}_i^n\mathbf{H}_i$. Thus, $\mathbf{Q}^{n+1}$ is the solution to the following problem

$$\begin{aligned} \text{minimize} \quad & \sum_{k=1}^{K}\tfrac{\Delta_k}{w_K}\,\mathrm{tr}\Big(\mathbf{\Phi}_k^{-n}\mathbf{Q}\Big)-\log|\mathbf{Q}| \\ \text{subject to} \quad & \mathrm{tr}(\mathbf{QP})=P, \mathbf{Q}:\text{diagonal}; \mathbf{Q}\succeq\mathbf{0}. \end{aligned} \tag{34}$$

We note that the idea of using the upper bound in (34) to optimize $\mathbf{Q}$ follows exactly the same as that of Algorithm 2. The above problem has the same form as (23), and thus closed-form solution using (27) and (28) can be applied. The overall proposed algorithm to solve (29) is summarized in Algorithm 3. We can prove the convergence of Algorithm 3 using the same lines as those for Algorithm 2 and thus the details are omitted due to space limitation. Similar to Algorithm 2, τ is only calculated for $n \geq 1$.

Algorithm 3 Proposed algorithm for the computation of the capacity region of a MIMO BC based on AO.

Input: $\mathbf{Q} := \mathbf{Q}^0$ diagonal matrix of positive elements, $\epsilon > 0$

1: Initialization: Set $n := 0$ and $\tau = 1 + \epsilon$.
2: **while** $(\tau > \epsilon)$ **do**
3: Solve (31) and denote the optimal solution by $\{\bar{\mathbf{S}}_k^n\}$
4: For $n \geq 1$, let $\tau = |f^{\mathrm{DPC}}(\mathbf{Q}^n, \{\bar{\mathbf{S}}^n\}) - f^{\mathrm{DPC}}(\mathbf{Q}^{n-1}, \{\bar{\mathbf{S}}^{n-1}\})|$, where $f^{\mathrm{DPC}}(\cdot)$ denotes the objective in (30).
5: For each k, compute $\mathbf{\Phi}_k^{-n} = (\mathbf{Q}^n + \sum_{i=k}^{K}\mathbf{H}_i^{\dagger}\bar{\mathbf{S}}_i^n\mathbf{H}_i)^{-1}$.
6: Solve (34) to find $\mathbf{Q}^{n+1}$.
7: $n := n + 1$.
8: **end while**

Output: Use the obtained $\{\bar{\mathbf{S}}_k^n\}_{k=1}^K$ and the BC-MAC transformation in [28] to find the optimal solution to (29).

IV. Numerical results

In this section, we numerically evaluate the performance of the proposed algorithms presented in this paper. For all iterative algorithms of comparison, we set an error tolerance of $\epsilon = 10^{-6}$ as the stopping criterion. The condition number κ is defined as the ratio between the largest singular value and the smallest one. The initial values $\mathbf{\Lambda}^0$ and $\mathbf{Q}^0$ in the corresponding proposed algorithms are set to the identity matrix for all simulations, if not mentioned otherwise. Other simulation parameters are specified for each setup. The codes are executed on a 64-bit desktop that supports 8 Gbyte RAM and Intel CORE i7.

A. Single user MIMO

In the first numerical experiment, we demonstrate the convergence rate of Algorithms 1 and 2, and the mode-dropping algorithm in [20], [21]. In particular we consider the same channel matrix as given in [20, Eq. (26)] and a total power of 0 dBW. As can be seen in Fig. 1, monotonic convergence is not always achieved for Algorithm 1, which is expected for an iterative method based on standard interference function. For the considered scenario, Algorithm 2 converges much faster than other methods of comparison. It can be implied from the iteration in (14) that Algorithm 1 will attain a good convergence rate if all the diagonal entries of $\mathbf{\Psi}(\tilde{\boldsymbol{\lambda}}_n)$ are much less than 1 during the whole iterative process, which is likely to occur if the singular values of $\mathbf{H}$ and/or $\mathbf{p}$ are relatively large. The same argument can also be applied to the mode-dropping method. However, this is not the case for the considered scenario, leading to slower convergence rates for Algorithm 1 and the mode-dropping method, compared to Algorithm 2. Further numerical results on this will be provided in Fig. 3.

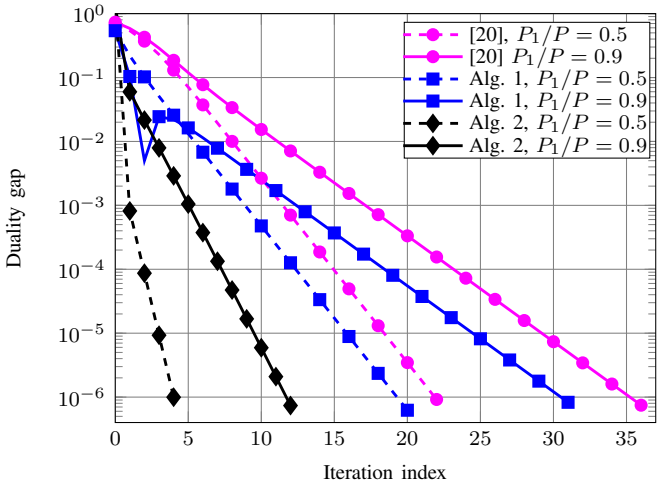


Fig. 1. Convergence comparison of different iterative methods for a point-to-point MIMO system with $N = 2$ and $M = 2$. The channel matrix is taken from [20].

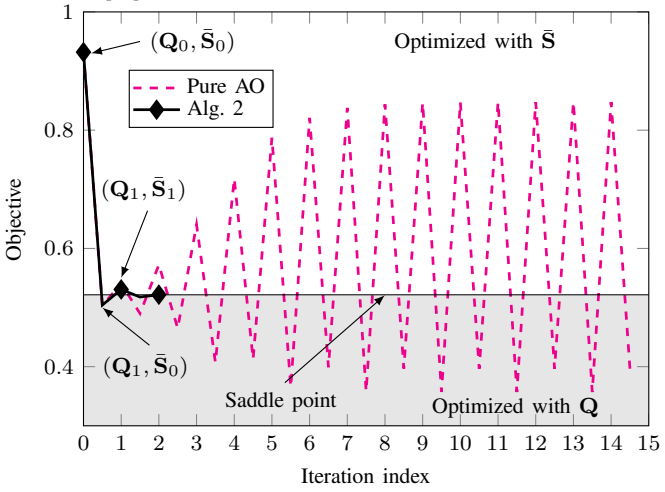


Fig. 2. Illustration of the ping-pong effect of the pure AO method, $P_1/P = 0.5$. The upper part of the figure plots the objective of both methods in comparison when it is optimized with $\bar{\mathbf{S}}$, while the lower part in gray color plots with $\mathbf{Q}$. The border line represents the objective of the saddle point.

In Fig. 2 we provide an example to show that a pure AO approach may fail to yield the optimal solution to (15) as briefly discussed earlier. The channel matrix is

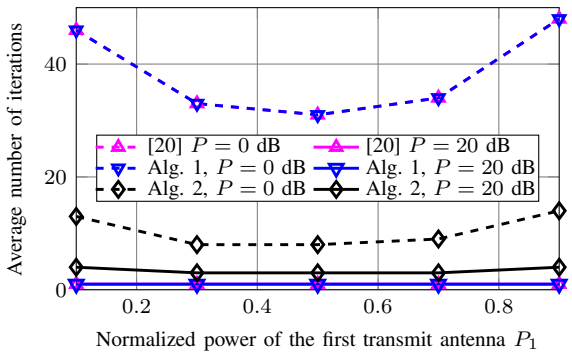
$\mathbf{H} = [-0.0723 - 0.6116i, 0.2257 - 0.1166i; -0.1707 - 0.0212i, 0.2212 + 0.4439i]$, which is generated randomly. The other simulation parameters are the same as those for Fig. 1. The initial value $\mathbf{Q}^0$ is set to identity. We can easily see that the objective returned by the pure AO method is oscillating and not converging to the optimal one. On the contrary, Algorithm 2 always guarantees a monotonically decreasing objective sequence converging to the optimal solution.

In the next set of numerical experiments we further investigate the convergence results of the algorithms in comparison. The numbers of transmit and receive antennas are set to $N = 2$ and $M = 4$, respectively. In particular, $\mathbf{\Lambda}^0$ in Algorithm 1 is generated in the same way as done in the mode-dropping method [20], [21]. Fig. 3 plots the average number of iterations as a function of P_1/P over 100 randomly generated channel realizations, and the total transmit power P is specified in the legends of the figure. In this considered setting, the channel matrix has two singular values. First, entries of $\mathbf{H}$ are generated following the i.i.d. zero mean and unit variance Gaussian, and then the smaller singular value is scaled accordingly to achieve a specific value of κ as given in Figs. 3(a) and 3(b).

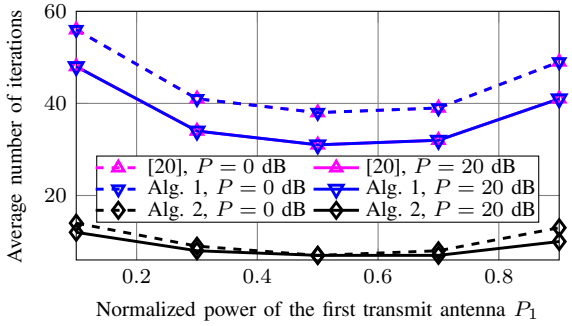
As can be seen clearly in Fig. 3, the convergence behavior of Algorithm 2 is quite consistent for different settings. On the other hand, Algorithm 1 and the mode-dropping scheme obtain the same convergence rate which is sensitive to κ and $\mathbf{p}$. In particular, Algorithm 1 takes more iterations to converge when the channel matrix is ill-conditioned (cf. Fig. 3(b)). However, Algorithm 1 converges faster for well-conditioned channel matrix and large $\mathbf{p}$ (cf. Fig. 3(a)). We can also see that the convergence rate of Algorithm 1 becomes inferior when one of the power limits P_i is small. For such a case, one of the diagonal element of $\mathbf{\Psi}(\tilde{\boldsymbol{\lambda}}_n)$ is very close to 1 for all iterations, making the fixed-point iteration converge slowly. In fact, these observations agree with what has been explained in Fig. 1.

As mentioned earlier, the overall complexity of an iterative algorithm depends on not only the per-iteration complexity but also the number of iterations that it takes to terminate. The overall complexity in terms of flop counts of the iterative methods in case $N \leq M$ is plotted in Fig. 4. As shown in the figure, Algorithm 2 has the lowest overall complexity. The reason is that Algorithm 2 has not only low per-iteration complexity but also (and more importantly) the smallest number of iterations as analyzed in Fig. 3. We also observe that if Algorithm 1 and the mode-dropping method start from the same initial point, the number of iterations to converge is identical. Thus, Algorithm 1 outperforms the mode-dropping method when $N < M$. However, when $N = M$, the total complexity of Algorithm 1 is $6N^3(13n + 4/3)$ while that of mode-dropping is $6N^3(12n + 21)$ where n is the number of iterations to converge. For this special case, it is possible that the total complexity of Algorithm 1 can be higher than the mode-dropping algorithm, depending on the number of iterations ($n \geq 20$) and vice versa.

In Fig. 5 we benchmark the average run time of proposed algorithms against interior-point methods. In particular the commercial interior-point-based solver MOSEK [38] is chosen for this purpose due to its recognized good performance. The



(a) Condition number $\kappa = 10$.



(b) Condition number $\kappa = 100$.

Fig. 3. Average number of iterations required to converge of different iterative algorithms with $N = 2$ and $M = 4$.

results in Fig. 5 are averaged over 1000 channel realizations which are randomly generated using the i.i.d. channel model. It can be seen clearly that the run time of MOSEK increases quickly with the number of transmit antennas. This observation is expected and consistent with the complexity analysis of interior-point methods presented earlier in Section II-C. On the contrary, other algorithms in comparison are more scalable, and Algorithms 1 and 2 still achieve better performance than the method in [21].

B. Multiuser MIMO

In the next part of this section we numerically study the performance of the proposed algorithms for the MIMO systems considered in the paper. In the following simulation, the number of receive antennas M and the number of transmit antennas N are fixed to 1 and 128 antennas, respectively. The number of users K is specified for each setup and the power limit for all antennas is equal to P/N.

Taking the advantage that the proposed algorithms have low complexity, in the last numerical experiment we characterize the capacity region of a massive MIMO system with PAPC. In particular, we also consider achievable rate region of the well-known ZF scheme. The purpose is to understand the performance of ZF in comparison with the capacity achieving coding scheme under some realistic channel models. To this end we consider a simple urban scenario using WINNER II B1 channel model [39], where a base station, equipped with $N = 128$ antennas, is located at the center of the cell and single-antenna receivers are distributed randomly. The total power at the BS is $P = 46$ dBm and each antenna

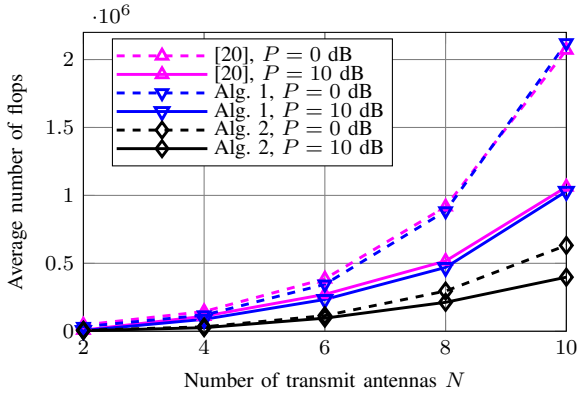


Fig. 4. Total complexity comparison versus the number of transmit antennas N. The number of receive antennas is taken as $M = 10$, $\kappa = 10$.

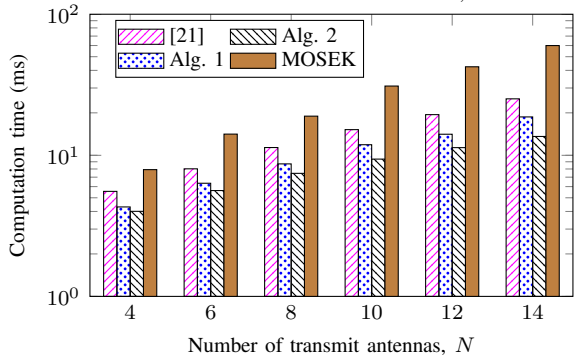
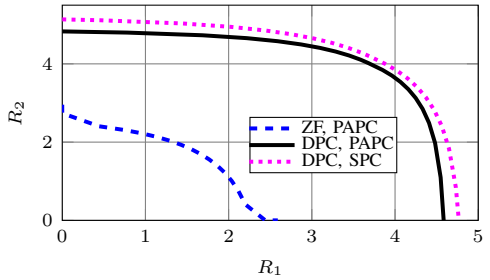


Fig. 5. Run time versus the number of transmit antennas N. Four methods are compared: [21], Algorithm 1, Algorithm 2, and the interior-point-based method implemented in [38]. The number of receive antennas is taken as $M = 2$.

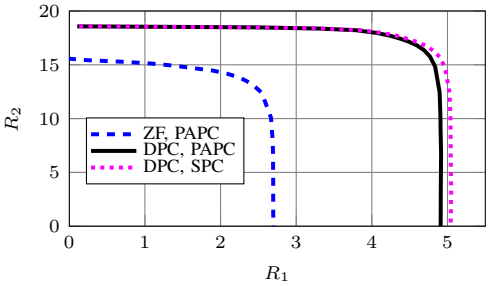
is subject to equal power constraint, i.e., $P_i = P/N$ for $i = 1, 2, \ldots, 128$. As can be seen clearly in Fig. 6, there is a remarkable gap between the achievable rate region of ZF and the capacity region, especially when the number of users increases. This basically implies that ZF is still far from optimal for a practical number of transmit antennas. Our observation opens research opportunities in the future to strike the balance between optimal performance by DPC and low-complexity by ZF.

V. Conclusions

We have solved the problem of computing the capacity of MIMO systems under PAPC. For a SU-MIMO system, two efficient algorithms have been proposed, one based on fixed point iteration and the other based on the MAC-BC duality together with AO. Extensive numerical experiments have been provided to demonstrate the superior performance of the two proposed algorithms over the known methods in [20], [21] in terms of computational complexity. We have also explored the capacity of multiuser MIMO systems subject to PAPC. For DPC which is the capacity-achieving transmission scheme, we have presented a method to compute the full capacity region. Using this low-complexity method, we have also characterized the capacity region of a single cell multiuser massive MIMO system subject to PAPC. The numerical results have demonstrated that the conventional ZF scheme still operates far from the capacity boundary for a practical number of transmit antennas.



(a) Number of users $K = 2$



(b) Number of users $K = 8$

Fig. 6. Comparison of capacity region of a massive MIMO setup with $N = 128$, $M = 1$. For the case $K = 8$ users, the capacity region is projected on the first two users.

Appendix A
Proof of Lemma 1

The key to prove the convergence of the fixed-point iteration in (14) is to show that $\mathfrak{I}(\mathbf{x})$ is a standard interference function. That is, for all $\mathbf{x} \geq 0$ then $\mathfrak{I}(\mathbf{x})$ satisfies the following three properties

- Positivity: $\mathfrak{I}(\mathbf{x}) > 0$.
- Monotonicity: If $\mathbf{x} \geq \mathbf{y}$, then $\mathfrak{I}(\mathbf{x}) \geq \mathfrak{I}(\mathbf{y})$.
- Scalability: For all $\alpha > 1, \alpha\mathfrak{I}(\mathbf{x}) > \mathfrak{I}(\alpha\mathbf{x})$.

According to [40, Theorem 2], if a function satisfies three properties listed above, it will converge to a unique fixed point. The positivity is obvious and the scalability can be easily shown by the following inequalities

$$\begin{aligned} \mathfrak{I}(\alpha\mathbf{x}) &= \mathbf{p} + \alpha\,\mathrm{diag}(\boldsymbol{\Psi}(\alpha\mathbf{x})) \odot \mathbf{x} && (35)\\ &\overset{(a)}{\leq} \mathbf{p} + \alpha\,\mathrm{diag}(\boldsymbol{\Psi}(\mathbf{x})) \odot \mathbf{x} && (36)\\ &\overset{(b)}{<} \alpha(\mathbf{p} + \mathrm{diag}(\boldsymbol{\Psi}(\mathbf{x})) \odot \mathbf{x}) = \alpha\mathfrak{I}(\mathbf{x}) && (37) \end{aligned}$$

where (a) can be proven from the definition of $\boldsymbol{\Psi}(\mathbf{x})$ as follows. Let $\mathbf{X} = \mathrm{diag}(\mathbf{x})$ and $\mathbf{V}\boldsymbol{\Sigma}\mathbf{V}^\dagger = \mathbf{X}\mathbf{H}^\dagger\mathbf{H}\mathbf{X}$ be the EVD of $\mathbf{X}\mathbf{H}^\dagger\mathbf{H}\mathbf{X}$, where $\boldsymbol{\Sigma} = \mathrm{diag}([\rho_1, \rho_2, \ldots, \rho_r, \mathbf{0}_{N-r}])$ and $r = \mathrm{rank}(\mathbf{H}^\dagger\mathbf{H})$. Then it follows immediately that $\mathbf{V}\tilde{\boldsymbol{\Sigma}}\mathbf{V}^\dagger = \tilde{\mathbf{X}}\mathbf{H}^\dagger\mathbf{H}\tilde{\mathbf{X}}$ where $\tilde{\mathbf{X}} = \mathrm{diag}(\alpha\mathbf{x})$, $\tilde{\boldsymbol{\Sigma}} = \mathrm{diag}([\tilde{\rho}_1, \tilde{\rho}_2, \ldots, \tilde{\rho}_r, \mathbf{0}_{N-r}])$, and $\tilde{\rho}_i = \alpha^2\rho_i$ for $i = 1, 2, \ldots, r$. Since $\alpha > 1$, we have $\frac{1}{\tilde{\rho}_i} = \frac{1}{\alpha^2\rho_i} < \frac{1}{\rho_i}$, and thus

$$\begin{aligned} \boldsymbol{\Psi}(\alpha\mathbf{x}) &= \mathbf{V}\,\mathrm{diag}(\frac{1}{\tilde{\rho}_1}, \cdots, \frac{1}{\tilde{\rho}_{s'}}, \mathbf{1}_{N-s'})\mathbf{V}^\dagger \\ &\preceq \boldsymbol{\Psi}(\mathbf{x}) = \mathbf{V}\,\mathrm{diag}(\frac{1}{\rho_1}, \cdots, \frac{1}{\rho_s}, \mathbf{1}_{N-s})\mathbf{V}^\dagger \qquad (38) \end{aligned}$$

where s' and s are the largest number such that $1 - \frac{1}{\rho_{s'}} > 0$ and $1 - \frac{1}{\rho_s} > 0$, respectively. Note that $s' \geq s$ and thus the above

inequality is easily justified. Consequently, $\mathrm{diag}(\mathbf{\Psi}(\alpha\mathbf{x})) \leq \mathrm{diag}(\mathbf{\Psi}(\mathbf{x}))$ which completes (a). The inequality (b) holds since $\mathbf{p} < \alpha\mathbf{p}$ for $\alpha > 1$, which results in $\mathbf{p}+\alpha\,\mathrm{diag}(\mathbf{\Psi}(\mathbf{x}))\odot \mathbf{x} < \alpha(\mathbf{p} + \mathrm{diag}(\mathbf{\Psi}(\mathbf{x}))\odot\mathbf{x}) = \alpha\mathfrak{J}(\mathbf{x})$.

To prove the monotonicity of $\mathfrak{J}(\mathbf{x})$, we need to show that for all $\mathbf{x},\mathbf{y} \geq 0$ then $\mathfrak{J}(\mathbf{x}) \geq \mathfrak{J}(\mathbf{y})$ or equivalently $\mathrm{diag}(\mathbf{\Psi}(\mathbf{x}))\odot\mathbf{x} \geq \mathrm{diag}(\mathbf{\Psi}(\mathbf{y}))\odot\mathbf{y}$. Let $\mathbf{X} = \mathrm{diag}(\mathbf{x}), \mathbf{Y} = \mathrm{diag}(\mathbf{y})$. Then monotonicity proof is equivalent to showing that $\mathrm{diag}(\mathbf{X}^{1/2}\mathbf{\Psi}(\mathbf{x})\mathbf{X}^{1/2}) \geq \mathrm{diag}(\mathbf{Y}^{1/2}\mathbf{\Psi}(\mathbf{y})\mathbf{Y}^{1/2})$ for $\mathbf{X} \succeq \mathbf{Y} \succeq \mathbf{0}$.

Let us first consider the case $N \leq M$ and $\mathbf{H}$ is full column rank. Then we can write the EVD of $\mathbf{\Lambda}^{-1/2}\mathbf{H}^\dagger\mathbf{H}\mathbf{\Lambda}^{-1/2}$ as

$$\underbrace{\mathbf{\Lambda}^{-1/2}\mathbf{H}^\dagger\mathbf{H}\mathbf{\Lambda}^{-1/2}}_{\mathbf{B}} = \mathbf{V}\mathbf{\Sigma}\mathbf{V}^\dagger. \tag{39}$$

For notational convenience, let $\mathbf{K} = \mathbf{H}^\dagger\mathbf{H}$. Note that $\mathbf{K}$ is full-rank and thus invertible. Then the above equation can be rewritten as

$$\mathbf{B}^{-1} = \mathbf{\Lambda}^{1/2}\mathbf{K}^{-1}\mathbf{\Lambda}^{1/2} = \mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{V}^\dagger \tag{40}$$

which the results in

$$\mathbf{B}^{-1} - \mathbf{I} = \mathbf{V}\big(\mathbf{\Sigma}^{-1} - \mathbf{I}\big)\mathbf{V}^\dagger. \tag{41}$$

Let $\tilde{\mathbf{\Sigma}}$ be the $(N-k)$ *positive* eigenvalues of $\mathbf{B}^{-1}-\mathbf{I}$ and $\tilde{\mathbf{V}}_k$ consist of the corresponding $N-k$ eigenvectors, $\bar{\mathbf{\Sigma}}$ be the k *non-positive* eigenvalues of $\mathbf{B}^{-1}-\mathbf{I}$, and $\bar{\mathbf{V}}_k$ consist of the corresponding k eigenvectors, and define

$$\mathbf{A}^+ = \tilde{\mathbf{V}}_k\tilde{\mathbf{\Sigma}}\tilde{\mathbf{V}}_k^\dagger \tag{42a}$$

$$\mathbf{A}^- = \bar{\mathbf{V}}_k\bar{\mathbf{\Sigma}}\bar{\mathbf{V}}_k^\dagger. \tag{42b}$$

Then it holds that

$$\mathbf{B}^{-1} - \mathbf{I} = \mathbf{A}^+ + \mathbf{A}^- \tag{43}$$

and that $\mathbf{A}^-\mathbf{A}^+ = \mathbf{0}$. Now we can write $\mathbf{\Psi}(\tilde{\boldsymbol{\lambda}}) = \mathbf{A}^- + \mathbf{I} = \mathbf{B}^{-1} - \mathbf{A}^+$ and thus

$$\begin{aligned}
[\mathbf{\Psi}(\tilde{\boldsymbol{\lambda}})\mathbf{\Lambda}^{-1}]_{i,i} &= [\mathbf{\Lambda}^{-1/2}\mathbf{\Psi}(\tilde{\boldsymbol{\lambda}})\mathbf{\Lambda}^{-1/2}]_{i,i} \\
&= [\mathbf{\Lambda}^{-1/2}\big(\mathbf{B}^{-1} - \mathbf{A}^+\big)\mathbf{\Lambda}^{-1/2}]_{i,i} \\
&= [\mathbf{\Lambda}^{-1/2}\mathbf{B}^{-1}\mathbf{\Lambda}^{-1/2}]_{i,i} - [\mathbf{\Lambda}^{-1/2}\mathbf{A}^+\mathbf{\Lambda}^{-1/2}]_{i,i} \\
&= [\mathbf{K}^{-1}]_{i,i} - [\mathbf{\Lambda}^{-1/2}\mathbf{A}^+\mathbf{\Lambda}^{-1/2}]_{i,i} \\
&= [\mathbf{K}^{-1}]_{i,i} - [\tilde{\mathbf{\Lambda}}^{1/2}\mathbf{A}^+\tilde{\mathbf{\Lambda}}^{1/2}]_{i,i}.
\end{aligned} \tag{44}$$

To proceed further we need to show that if $\mathbf{X} \succeq \mathbf{Y}$ then

$$[\mathbf{X}^{1/2}\mathbf{A}_X^+\mathbf{X}^{1/2}]_{i,i} \leq [\mathbf{Y}^{1/2}\mathbf{A}_Y^+\mathbf{Y}^{1/2}]_{i,i}. \tag{45}$$

Now from (43) we have

$$\mathbf{X}^{-1/2}\mathbf{K}^{-1}\mathbf{X}^{-1/2} - \mathbf{I} = \mathbf{A}_X^+ + \mathbf{A}_X^- \tag{46}$$

which is equivalent to

$$\mathbf{K}^{-1} - \mathbf{X} = \mathbf{X}^{1/2}\mathbf{A}_X^+\mathbf{X}^{1/2} + \mathbf{X}^{1/2}\mathbf{A}_X^-\mathbf{X}^{1/2}. \tag{47}$$

The same result applies to $\mathbf{Y}$, i.e.,

$$\mathbf{K}^{-1} - \mathbf{Y} = \mathbf{Y}^{1/2}\mathbf{A}_Y^+\mathbf{Y}^{1/2} + \mathbf{Y}^{1/2}\mathbf{A}_Y^-\mathbf{Y}^{1/2}. \tag{48}$$

Since $\mathbf{X} \succeq \mathbf{Y} \succ \mathbf{0}$ it holds that

$$\mathbf{K}^{-1} - \mathbf{X} \preceq \mathbf{K}^{-1} - \mathbf{Y}. \tag{49}$$

Substituting (47) and (48) into (49) yields

$$\begin{aligned}
&\mathbf{X}^{1/2}\mathbf{A}_X^+\mathbf{X}^{1/2} + \mathbf{X}^{1/2}\mathbf{A}_X^-\mathbf{X}^{1/2} \\
&\qquad \preceq \mathbf{Y}^{1/2}\mathbf{A}_Y^+\mathbf{Y}^{1/2} + \mathbf{Y}^{1/2}\mathbf{A}_Y^-\mathbf{Y}^{1/2}.
\end{aligned} \tag{50}$$

We now recall the following inequality. For Hermitian matrices $\mathbf{A}$ and $\mathbf{B}$, if $\mathbf{A} \succeq \mathbf{B}$, then $\mathbf{S}\mathbf{A}\mathbf{S}^H \succeq \mathbf{S}\mathbf{B}\mathbf{S}^H$ for $\mathbf{S} \succeq 0$ [41, Observation 7.7.2]. Applying this inequality to (50) leads to

$$\begin{aligned}
&\mathbf{A}_X^+ - \mathbf{X}^{-1/2}\mathbf{Y}^{1/2}\mathbf{A}_Y^+\mathbf{Y}^{1/2}\mathbf{X}^{-1/2} \\
&\qquad \preceq \mathbf{X}^{-1/2}\mathbf{Y}^{1/2}\mathbf{A}_Y^-\mathbf{Y}^{1/2}\mathbf{X}^{-1/2} - \mathbf{A}_X^-
\end{aligned} \tag{51}$$

which is then equivalent to

$$\begin{aligned}
&\big(\mathbf{A}_X^+\big)^{1/2}\big(\mathbf{A}_X^+ - \mathbf{X}^{-1/2}\mathbf{Y}^{1/2}\mathbf{A}_Y^+\mathbf{Y}^{1/2}\mathbf{X}^{-1/2}\big)\big(\mathbf{A}_X^+\big)^{1/2} \\
&\preceq \big(\mathbf{A}_X^+\big)^{1/2}\big(\mathbf{X}^{-1/2}\mathbf{Y}^{1/2}\mathbf{A}_Y^-\mathbf{Y}^{1/2}\mathbf{X}^{-1/2} - \mathbf{A}_X^-\big)\big(\mathbf{A}_X^+\big)^{1/2} \\
&\preceq \mathbf{0}.
\end{aligned} \tag{52}$$

The above inequality holds true since $\big(\mathbf{A}_X^+\big)^{1/2}\mathbf{A}_X^-\big(\mathbf{A}_X^+\big)^{1/2} = \mathbf{0}$. It is easy to see that (52) results in

$$\mathbf{X}^{1/2}\mathbf{A}_X^+\mathbf{X}^{1/2} \preceq \mathbf{Y}^{1/2}\mathbf{A}_Y^+\mathbf{Y}^{1/2} \tag{53}$$

and thus

$$[\mathbf{X}^{1/2}\mathbf{A}_X^+\mathbf{X}^{1/2}]_{i,i} \leq [\mathbf{Y}^{1/2}\mathbf{A}_Y^+\mathbf{Y}^{1/2}]_{i,i} \tag{54}$$

for all i. Here we have used a well-known fact that for $\mathbf{A} \succeq \mathbf{B}$, then $[\mathbf{A}]_{i,i} \geq [\mathbf{B}]_{i,i}$.

We now turn our attention to the general case where $\mathbf{K}$ is singular. This occurs when $N > M$ or $N \leq M$ but $\mathbf{H}$ is not full column rank, i.e. the columns of $\mathbf{H}$ are not linearly independent. First we add a small regularization term to both sides of (39) to obtain

$$\begin{aligned}
\mathbf{B}_\epsilon &= \mathbf{\Lambda}^{-1/2}\mathbf{H}^\dagger\mathbf{H}\mathbf{\Lambda}^{-1/2} + \epsilon\mathbf{\Lambda}^{-1} \\
&= \mathbf{\Lambda}^{-1/2}\left(\mathbf{H}^\dagger\mathbf{H} + \epsilon\mathbf{I}\right)\mathbf{\Lambda}^{-1/2} \\
&= \mathbf{V}\mathbf{\Sigma}\mathbf{V}^\dagger + \epsilon\mathbf{\Lambda}^{-1}.
\end{aligned} \tag{55}$$

We note that $\mathbf{B}_\epsilon$ is invertible for any $\epsilon > 0$. Let $\mathbf{V}_\epsilon\mathbf{\Sigma}_\epsilon\mathbf{V}_\epsilon^\dagger$ be the EVD of $\mathbf{B}_\epsilon$ and thus

$$\mathbf{B}_\epsilon^{-1} = \mathbf{V}_\epsilon\mathbf{\Sigma}_\epsilon^{-1}\mathbf{V}_\epsilon^\dagger. \tag{56}$$

Applying the result for the nonsingular case, we achieve the following inequality

$$\mathbf{\Psi}_\epsilon(\mathbf{x})\mathbf{X} \succeq \mathbf{\Psi}_\epsilon(\mathbf{y})\mathbf{Y} \tag{57}$$

for arbitrarily small ϵ and $\mathbf{X} \succeq \mathbf{Y}$, and $\mathbf{\Psi}_\epsilon(\cdot)$ in constructed from $\mathbf{B}_\epsilon$. To complete the proof we are left to show that $\mathbf{\Psi}_\epsilon(\tilde{\boldsymbol{\lambda}})$ is continuous with ϵ, i.e., $\lim_{\epsilon\to 0^+}\mathbf{\Psi}_\epsilon(\tilde{\boldsymbol{\lambda}}) = \mathbf{\Psi}(\tilde{\boldsymbol{\lambda}}) = \mathbf{A}^- + \mathbf{I}$.

To proceed, we note that (43) is changed into

$$\mathbf{B}_\epsilon^{-1} - \mathbf{I} = \mathbf{A}_\epsilon^+ + \mathbf{A}_\epsilon^- \tag{58}$$

where $\mathbf{A}_\epsilon^+$ and $\mathbf{A}_\epsilon^-$ are defined similarly to (42). We will show that $\lim_{\epsilon\to 0}\mathbf{A}_\epsilon^- \to \mathbf{A}^-$. To this end let $\epsilon_{\min} = \epsilon\times\min_i\{1/\lambda_i\}$ and $\epsilon_{\max} = \epsilon\times\max_i\{1/\lambda_i\}$, where λ_i is the ith diagonal entry of $\mathbf{\Lambda}$. It is clear from from (55) that the following inequality holds

$$\begin{aligned}
&\underbrace{\mathbf{V}\left(\left(\mathbf{\Sigma} + \epsilon_{\min}\mathbf{I}\right)^{-1} - \mathbf{I}\right)\mathbf{V}^\dagger}_{\mathbf{\Xi}_{\epsilon_{\min}}} \succ \mathbf{B}_\epsilon^{-1} - \mathbf{I} \\
&\qquad \succ \underbrace{\mathbf{V}\left(\left(\mathbf{\Sigma} + \epsilon_{\max}\mathbf{I}\right)^{-1} - \mathbf{I}\right)\mathbf{V}^\dagger}_{\mathbf{\Xi}_{\epsilon_{\max}}}.
\end{aligned} \tag{59}$$

Further, the matrix $\Xi_{\epsilon_{\min}}$ can be explicitly written as

$$\begin{aligned}
\mathbf{\Xi}_{\epsilon_{\min}} = \mathbf{V}\,\mathrm{diag}([&\frac{1}{\rho_1 + \epsilon_{\min}} - 1, \ldots, \frac{1}{\rho_r + \epsilon_{\min}} - 1, \\
&\underbrace{\frac{1}{\epsilon_{\min}} - 1, \ldots, \frac{1}{\epsilon_{\min}} - 1}_{(N-r)\text{ terms}}])\mathbf{V}^\dagger
\end{aligned} \tag{60}$$

where $r = \mathrm{rank}(\mathbf{K})$. Following (43), we decompose $\mathbf{\Xi}_{\epsilon_{\min}}$ as

$$\mathbf{\Xi}_{\epsilon_{\min}} = \mathbf{A}^{+}_{\epsilon_{\min}} + \mathbf{A}^{-}_{\epsilon_{\min}} \tag{61}$$

where $\mathbf{A}^{+}_{\epsilon_{\min}}$ and $\mathbf{A}^{-}_{\epsilon_{\min}}$ consists of positive and non-positive eigenvalues, respectively. As $\epsilon \to 0^+$ we have $\frac{1}{\rho_i+\epsilon_{\min}} \to \frac{1}{\rho_i}$ for all $i = 1, 2, ..., r$, and $\frac{1}{\epsilon_{\min}} \gg 1$. Thus , the term $\frac{1}{\epsilon_{\min}} - 1$ in (60) becomes strictly positive and thus is excluded in $\mathbf{A}^{-}_{\epsilon_{\min}}$. As a result, we have $\lim_{\epsilon\to 0^+} \mathbf{A}^{-}_{\epsilon_{\min}} = \mathbf{A}^{-}$. Following the same arguments we can also show that $\lim_{\epsilon\to 0^+} \mathbf{A}^{-}_{\epsilon_{\max}} = \mathbf{A}^{-}$. From (59) it is clear that $\lim\limits_{\epsilon\to 0^+} \mathbf{A}^{-}_{\epsilon} = \mathbf{A}^{-}$ and thus

$$\lim_{\epsilon\to 0^+} \mathbf{\Psi}_{\epsilon}(\tilde{\boldsymbol{\lambda}}) = \lim_{\epsilon\to 0^+} (\mathbf{A}^{-}_{\epsilon} + \mathbf{I}) = \mathbf{A}^{-} + \mathbf{I} = \mathbf{\Psi}(\tilde{\boldsymbol{\lambda}}). \tag{62}$$

By the continuity property shown above, the monotonicity of Algorithm 1 also holds for the singular case, which completes the proof.

Appendix B
Convergence Proof of Algorithm 2

We note that the function $\log|\mathbf{Q}+\mathbf{H}^{\dagger}\bar{\mathbf{S}}\mathbf{H}|$ is *jointly concave* with $\mathbf{Q}$ and $\bar{\mathbf{S}}$. Thus the following inequality holds

$$\log|\mathbf{Q}+\mathbf{H}^{\dagger}\bar{\mathbf{S}}\mathbf{H}| \le \log|\underbrace{\mathbf{Q}_n + \mathbf{H}^{\dagger}\bar{\mathbf{S}}_n\mathbf{H}}_{\mathbf{\Phi}_n}| + \mathrm{tr}(\mathbf{\Phi}_n^{-1}(\mathbf{Q}-\mathbf{Q}_n)) + \mathrm{tr}(\mathbf{H}\mathbf{\Phi}_n^{-1}\mathbf{H}^{\dagger}(\bar{\mathbf{S}}-\bar{\mathbf{S}}_n)) \tag{63}$$

for all $\mathbf{Q} \in \mathcal{Q}$ and $\bar{\mathbf{S}} \in \mathcal{S}$. The above inequality comes from the first order approximation of $\log|\mathbf{Q} + \mathbf{H}^{\dagger}\bar{\mathbf{S}}\mathbf{H}|$ around the point $(\mathbf{Q}_n, \bar{\mathbf{S}}_n)$. Substitute $\mathbf{Q} := \mathbf{Q}_{n+1}$ and $\bar{\mathbf{S}} := \bar{\mathbf{S}}_{n+1}$ into the above equality, we have

$$\log|\mathbf{Q}_{n+1}+\mathbf{H}^{\dagger}\bar{\mathbf{S}}_{n+1}\mathbf{H}| \le \log|\mathbf{\Phi}_n| + \mathrm{tr}(\mathbf{\Phi}_n^{-1}(\mathbf{Q}_{n+1}-\mathbf{Q}_n)) + \mathrm{tr}(\mathbf{H}\mathbf{\Phi}_n^{-1}\mathbf{H}^{\dagger}(\bar{\mathbf{S}}_{n+1}-\bar{\mathbf{S}}_n)). \tag{64}$$

Since $\bar{\mathbf{S}}_n = \arg\max\limits_{\bar{\mathbf{S}}\in\mathcal{S}} \log|\mathbf{Q}_n + \mathbf{H}^{\dagger}\bar{\mathbf{S}}\mathbf{H}|$, the optimality condition results in

$$\mathrm{tr}(\mathbf{H}\mathbf{\Phi}_n^{-1}\mathbf{H}^{\dagger}(\bar{\mathbf{S}}-\bar{\mathbf{S}}_n)) \le 0 \tag{65}$$

for all $\bar{\mathbf{S}} \in \mathcal{S}$. For $\bar{\mathbf{S}} = \bar{\mathbf{S}}_{n+1}$ the above inequality means

$$\mathrm{tr}(\mathbf{H}\mathbf{\Phi}_n^{-1}\mathbf{H}^{\dagger}(\bar{\mathbf{S}}_{n+1}-\bar{\mathbf{S}}_n)) \le 0 \tag{66}$$

which leads to

$$\log|\mathbf{Q}_{n+1}+\mathbf{H}^{\dagger}\bar{\mathbf{S}}_{n+1}\mathbf{H}| \le \log|\mathbf{\Phi}_n| + \mathrm{tr}(\mathbf{\Phi}_n^{-1}(\mathbf{Q}_{n+1}-\mathbf{Q}_n)). \tag{67}$$

Subtract both sides of the above inequality by $\log|\mathbf{Q}_{n+1}|$ results in

$$\begin{aligned} f(\mathbf{Q}_{n+1}, \bar{\mathbf{S}}_{n+1}) &= \log|\mathbf{Q}_{n+1} + \mathbf{H}^{\dagger}\bar{\mathbf{S}}_{n+1}\mathbf{H}| - \log|\mathbf{Q}_{n+1}| \\ &\le \log|\mathbf{\Phi}_n| + \mathrm{tr}(\mathbf{\Phi}_n^{-1}(\mathbf{Q}_{n+1}-\mathbf{Q}_n)) - \log|\mathbf{Q}_{n+1}|. \end{aligned} \tag{68}$$

Since $\mathbf{Q}_{n+1}$ solves (23) it holds that

$$\begin{aligned} &\log|\mathbf{\Phi}_n| + \mathrm{tr}\Big(\mathbf{\Phi}_n^{-1}\big(\mathbf{Q}_{n+1}-\mathbf{Q}_n\big)\Big) - \log|\mathbf{Q}_{n+1}| \\ &\qquad \le \log|\mathbf{\Phi}_n| + \mathrm{tr}\Big(\mathbf{\Phi}_n^{-1}\big(\mathbf{Q}-\mathbf{Q}_n\big)\Big) - \log|\mathbf{Q}| \end{aligned} \tag{69}$$

for all $\mathbf{Q} \in \mathcal{Q}$. For the special case $\mathbf{Q} := \mathbf{Q}_n$, the above inequality is reduced to

$$\log|\mathbf{\Phi}_n| + \mathrm{tr}\Big(\mathbf{\Phi}_n^{-1}\big(\mathbf{Q}_{n+1}-\mathbf{Q}_n\big)\Big) - \log|\mathbf{Q}_{n+1}| \le \underbrace{\log|\mathbf{\Phi}_n| - \log|\mathbf{Q}_n|}_{f(\mathbf{Q}_n,\bar{\mathbf{S}}_n)}. \tag{70}$$

Combining (68) and (70) results in $f(\mathbf{Q}_n, \bar{\mathbf{S}}_n) \ge f(\mathbf{Q}_{n+1}, \bar{\mathbf{S}}_{n+1})$.

It is easy to see that $\{f(\mathbf{Q}_n, \bar{\mathbf{S}}_n)\}$ is bounded below, and thus $\{f(\mathbf{Q}_n, \bar{\mathbf{S}}_n)\}$ is convergent. We also note that (22) is strict if $\mathbf{Q} \neq \mathbf{Q}_n$. Consequently, the sequence $\{f(\mathbf{Q}_n, \bar{\mathbf{S}}_n)\}$ is *strictly decreasing* unless it is convergent.

Let us consider the set $\mathcal{Q}_+ \triangleq \{\mathrm{tr}(\mathbf{QP}) \le P; \mathbf{Q} \succ \mathbf{0}\}$. Note that $\mathcal{Q}_+$ is open. As mentioned previously, $\mathbf{Q}_n \in \mathcal{Q}_+$ for all n. We will prove the two following properties regarding the convergence of Algorithm 2:

- Algorithm 2 generates at least a convergent subsequence.
- Let $\mathbf{Q}^*$ be the limit point of $\{\mathbf{Q}_n\}$. Then $\mathbf{Q}^*$ is nonsingular, i.e. $\mathbf{Q}^* \in \mathcal{Q}_+$.

The first property is relatively trivial. It is easy to see that the set $\mathcal{Q}_+$ is bounded (though it is open). As $\mathcal{Q}_+$ and $\mathcal{S}$ are both bounded, Algorithm 2 must produce at least a convergent subsequence, due to the Bolzano-Weierstrass theorem [42], [43]. The proof for the second property is quite involved, which is done by contraction as follows.

Suppose the contrary that $\mathbf{Q}^*$ is singular, i.e., there exists $\{q_{n,i}\} \to 0$ for some i. Recall that $\bar{\mathbf{S}}_n = \arg\max \log|\mathbf{Q}_n + \mathbf{H}^{\dagger}\bar{\mathbf{S}}\mathbf{H}|$, and thus replacing $\bar{\mathbf{S}}_n = \frac{P}{N}\mathbf{I}$ which is a feasible point to the maximization problem results in

$$\log|\mathbf{Q}_n + \mathbf{H}^{\dagger}\bar{\mathbf{S}}_n\mathbf{H}| \ge \log|\mathbf{Q}_n + \tfrac{P}{N}\mathbf{H}^{\dagger}\mathbf{H}|. \tag{71}$$

Consequently we have

$$\begin{aligned} f(\mathbf{Q}_n, \bar{\mathbf{S}}_n) &\ge \log|\mathbf{Q}_n + \tfrac{P}{N}\mathbf{H}^{\dagger}\mathbf{H}| - \log|\mathbf{Q}_n| \\ &= \log|\mathbf{I} + \tfrac{P}{N}\mathbf{Q}_n^{-1/2}\mathbf{H}^{\dagger}\mathbf{H}\mathbf{Q}_n^{-1/2}| \\ &= \log|\mathbf{I} + \tfrac{P}{N}\mathbf{H}\mathbf{Q}_n^{-1}\mathbf{H}^{\dagger}| \\ &= \log|\mathbf{I} + \tfrac{P}{N}\sum_{l=1}^{N} q_{n,l}^{-1}\mathbf{h}_l\mathbf{h}_l^{\dagger}|. \end{aligned} \tag{72}$$

Note that $\mathbf{h}_l$ is the lth column of $\mathbf{H}$. Let $\mathbf{A}_{n,i} = \mathbf{I} + \frac{P}{N}\sum_{l\neq i}^{N} q_{n,l}^{-1}\mathbf{h}_l\mathbf{h}_l^{\dagger}$. Then we can write

$$\begin{aligned} f(\mathbf{Q}_n, \bar{\mathbf{S}}_n) &\ge \log|\mathbf{A}_{n,i} + \tfrac{P}{N}q_{n,i}^{-1}\mathbf{h}_i\mathbf{h}_i^{\dagger}| \\ &= \log|\mathbf{A}_{n,i}| + \log|\mathbf{I} + \tfrac{P}{N}q_{n,i}^{-1}\mathbf{A}_{n,i}^{-1/2}\mathbf{h}_i\mathbf{h}_i^{\dagger}\mathbf{A}_{n,i}^{-1/2}| \\ &= \log|\mathbf{A}_{n,i}| + \log(1 + \tfrac{P}{N}q_{n,i}^{-1}\mathbf{h}_i^{\dagger}\mathbf{A}_{n,i}^{-1}\mathbf{h}_i). \end{aligned} \tag{73}$$

Let $\upsilon_{n,i}^{\max}$ be the maximum eigenvalue of $\mathbf{A}_{n,i}$, and thus $\frac{1}{\upsilon_{n,i}^{\max}}$ is the minimum eigenvalue of $\mathbf{A}_{n,i}^{-1}$. Then we have

$$f(\mathbf{Q}_n, \bar{\mathbf{S}}_n) \ge \log(\upsilon_{n,i}^{\max}) + \log(1 + \tfrac{P}{Nq_{n,i}\upsilon_{n,i}^{\max}}||\mathbf{h}_i||_2^2) \tag{74}$$

where we have used the fact that all eigenvalues of $\mathbf{A}_{n,i}$ are no less than 1, and that $\mathbf{x}^{\dagger}\mathbf{B}\mathbf{x} \ge \lambda_{\min}||\mathbf{x}||_2^2$, where $\lambda_{\min}$ is the minimum eigenvalue of $\mathbf{B}$.

To proceed further we consider two cases. Specifically, if $\lim_{n\to\infty} \upsilon_{n,i}^{\max} = \infty$, then it immediately holds that $\lim_{n\to\infty} f(\mathbf{Q}_n, \bar{\mathbf{S}}_n) = \infty$. Now suppose that there exists $c < \infty$ such that $1 \le \upsilon_{n,i}^{\max} \le c$ for all n. In this case we obtain

$$f(\mathbf{Q}_n, \bar{\mathbf{S}}_n) \ge \log(1 + \tfrac{P}{Nc}\tfrac{1}{q_{n,i}}||\mathbf{h}_i||_2^2). \tag{75}$$

It is straightforward to see that $f(\mathbf{Q}_n, \bar{\mathbf{S}}_n) \to \infty$ as $q_{n,i} \to 0$, due to the fact that $||\mathbf{h}_i||_2 > 0$.

In summary we have proved that if there exists $\{q_{n,i}\} \to 0$ for some i, then $\{f(\mathbf{Q}_n, \bar{\mathbf{S}}_n)\} \to \infty$. This contradicts the fact that $\infty > f(\mathbf{Q}_0, \bar{\mathbf{S}}_0) \ge f(\mathbf{Q}_n, \bar{\mathbf{S}}_n)$ for all n as proved earlier. Thus it is concluded that the limit point of Algorithm 2 $\mathbf{Q}^*$ is non-singular. By the continuity of $f(\cdot)$ over $\mathcal{S}$ and $\mathcal{Q}_+$, we have $\lim\limits_{n\to\infty} f(\mathbf{Q}_n, \bar{\mathbf{S}}_n) = f(\mathbf{Q}^*, \bar{\mathbf{S}}^*)$.

Now let $\{(\mathbf{Q}_{n_k}, \bar{\mathbf{S}}_{n_k})\}$ be the subsequence converging to the limit point. Next we shall show that $\{(\mathbf{Q}_{n_k+1}, \bar{\mathbf{S}}_{n_k+1})\} \to (\mathbf{Q}^*, \bar{\mathbf{S}}^*)$. In fact, it is sufficient to prove that $\mathbf{Q}_{n_k+1} \to \mathbf{Q}^*$ which can be done by contradiction. Assume the contrary that $\mathbf{Q}_{n_k+1}$ does not converge to $\mathbf{Q}^*$. Consequently, there exists a $d > 0$ such that

$$d \leq d_{n_k} = ||\mathbf{Q}_{n_k+1} - \mathbf{Q}_{n_k}||, \forall k \tag{76}$$

where $||\cdot||$ stands for arbitrary norm. We have

$$\begin{aligned} f(\mathbf{Q}_{n_k+1}, \bar{\mathbf{S}}_{n_k+1}) &\leq F(\mathbf{Q}_{n_k+1}; \mathbf{Q}_{n_k}, \bar{\mathbf{S}}_{n_k}) && (77)\\ &= F(\mathbf{Q}_{n_k} + d_{n_k}\mathbf{\Gamma}_{n_k}; \mathbf{Q}_{n_k}, \bar{\mathbf{S}}_{n_k}) && (78)\\ &\leq F(\mathbf{Q}_{n_k} + \delta d\mathbf{\Gamma}_{n_k}; \mathbf{Q}_{n_k}, \bar{\mathbf{S}}_{n_k}), \forall \delta \in [0,1] \\ &\leq F(\mathbf{Q}_{n_k}; \mathbf{Q}_{n_k}, \bar{\mathbf{S}}_{n_k}) && (79)\\ &= f(\mathbf{Q}_{n_k}, \bar{\mathbf{S}}_{n_k}) && (80) \end{aligned}$$

where $\mathbf{\Gamma}_{n_k} \triangleq (\mathbf{Q}_{n_k+1} - \mathbf{Q}_{n_k})/d_{n_k}$ is the normalized distance between $\mathbf{Q}_{n_k+1}$ and $\mathbf{Q}_{n_k}$, $F(\mathbf{Q}_{n_k+1}; \mathbf{Q}_{n_k}, \bar{\mathbf{S}}_{n_k}) = \log|\mathbf{\Phi}_{n_k}| + \mathrm{tr}\Big(\mathbf{\Phi}_{n_k}^{-1}(\mathbf{Q}_{n_k+1} - \mathbf{Q}_{n_k})\Big) - \log|\mathbf{Q}_{n_k+1}|$. Note that $||\mathbf{\Gamma}_{n_k}|| = 1$ and thus $\mathbf{\Gamma}_{n_k}$ lies in a compact set and has a limit point $\mathbf{\Gamma}^*$. Letting $k \to \infty$ (by further restricting to a subsequence converging to $\mathbf{\Gamma}^*$) leads to

$$f(\mathbf{Q}^*, \bar{\mathbf{S}}^*) \leq F(\mathbf{Q}^* + \delta d\mathbf{\Gamma}^*; \mathbf{Q}^*, \bar{\mathbf{S}}^*) \leq f(\mathbf{Q}^*, \bar{\mathbf{S}}^*) \tag{81}$$

or equivalently

$$f(\mathbf{Q}^*, \bar{\mathbf{S}}^*) = F(\mathbf{Q}^* + \delta d\mathbf{\Gamma}^*; \mathbf{Q}^*, \bar{\mathbf{S}}^*), \forall \delta \in [0,1]. \tag{82}$$

Furthermore

$$\begin{aligned} F(\mathbf{Q}_{n_{k+1}}; \mathbf{Q}_{n_{k+1}}, \bar{\mathbf{S}}_{n_{k+1}}) &= f(\mathbf{Q}_{n_{k+1}}, \bar{\mathbf{S}}_{n_{k+1}}) \\ &\leq f(\mathbf{Q}_{n_k+1}, \bar{\mathbf{S}}_{n_k+1}) \leq F(\mathbf{Q}_{n_k+1}, \mathbf{Q}_{n_k}, \bar{\mathbf{S}}_{n_k}) \\ &\leq F(\mathbf{Q}; \mathbf{Q}_{n_k}, \bar{\mathbf{S}}_{n_k}), \forall \mathbf{Q} \in \mathcal{Q}_+. \end{aligned} \tag{83}$$

Letting $k \to \infty$ we obtain

$$F(\mathbf{Q}^*; \mathbf{Q}^*, \bar{\mathbf{S}}^*) \leq F(\mathbf{Q}; \mathbf{Q}^*, \bar{\mathbf{S}}^*), \forall \mathbf{Q} \in \mathcal{Q}_+ \tag{84}$$

which further implies that $\mathbf{Q}^*$ is the minimizer of $F(\cdot; \mathbf{Q}^*, \bar{\mathbf{S}}^*)$. Since $\mathbf{Q}_{n_k+1} = \underset{\mathbf{Q}\in\mathcal{Q}_+}{\arg\min} F(\mathbf{Q}; \mathbf{Q}_{n_k}, \bar{\mathbf{S}}_{n_k})$ it follows that

$$F(\mathbf{Q}_{n_k+1}; \mathbf{Q}_{n_k}, \bar{\mathbf{S}}_{n_k}) \leq F(\mathbf{Q}; \mathbf{Q}_{n_k}, \bar{\mathbf{S}}_{n_k}), \forall \mathbf{Q} \in \mathcal{Q}_+. \tag{85}$$

Letting $k \to \infty$ implies

$$F(\mathbf{Q}^*; \mathbf{Q}^*, \bar{\mathbf{S}}^*) \leq F(\mathbf{Q}; \mathbf{Q}^*, \bar{\mathbf{S}}^*), \forall \mathbf{Q} \in \mathcal{Q}_+. \tag{86}$$

That is

$$\langle \nabla_{\mathbf{Q}} F(\mathbf{Q}; \mathbf{Q}^*, \bar{\mathbf{S}}^*)|_{\mathbf{Q}=\mathbf{Q}^*}, \mathbf{Z} - \mathbf{Q}^* \rangle \geq 0, \forall \mathbf{Z} \in \mathcal{Q}_+ \tag{87}$$

where $\langle . \rangle$ denotes the inner product. Recall that $F(\cdot; \mathbf{Q}, \bar{\mathbf{S}})$ is the first order of $f(\mathbf{Q}, \bar{\mathbf{S}})$. Thus it is easy to see that

$$\nabla_{\mathbf{Q}} F(\mathbf{Q}; \mathbf{Q}^*, \bar{\mathbf{S}}^*)|_{\mathbf{Q}=\mathbf{Q}^*} = \nabla f(\mathbf{Q}^*, \bar{\mathbf{S}}^*) \tag{88}$$

and thus (87) is equivalent to

$$\langle \nabla_{\mathbf{Q}} f(\mathbf{Q}^*, \bar{\mathbf{S}}^*), \mathbf{Z} - \mathbf{Q}^* \rangle \geq 0, \forall \mathbf{Z} \in \mathcal{Q}_+. \tag{89}$$

In the same way we can show that

$$\langle \nabla_{\bar{\mathbf{S}}} f(\mathbf{Q}^*, \bar{\mathbf{S}}^*), \mathbf{W} - \bar{\mathbf{S}}^* \rangle \leq 0, \forall \mathbf{W} \in \mathcal{S}. \tag{90}$$

Two above inequalities imply that $(\mathbf{Q}^*, \bar{\mathbf{S}}^*)$ is a saddle point of (15), which completes the proof.

APPENDIX C
PROJECTION ONTO THE FEASIBLE SET OF (31)

The projection of $\{\tilde{\mathbf{S}}_k\}$ onto the feasible set of (31) is formulated as

$$\begin{aligned} \text{minimize} \quad & \textstyle\sum_{k=1}^K ||\bar{\mathbf{S}}_k - \tilde{\mathbf{S}}_k||_F^2 \\ \text{subject to} \quad & \textstyle\sum_{k=1}^K \mathrm{tr}(\bar{\mathbf{S}}_k) = P; \{\bar{\mathbf{S}}_k \succeq \mathbf{0}\}. \end{aligned} \tag{91}$$

Let $\mathbf{U}_k\tilde{\mathbf{D}}_k\mathbf{U}_k^\dagger = \tilde{\mathbf{S}}_k$ be the EVD of $\tilde{\mathbf{S}}_k$, where $\mathbf{U}_k$ is unitary and $\tilde{\mathbf{D}}_k$ is diagonal. Then we can write $\bar{\mathbf{S}}_k = \mathbf{U}_k\bar{\mathbf{D}}_k\mathbf{U}_k^\dagger$ for some $\bar{\mathbf{D}}_k \succeq \mathbf{0}$. Since $\mathbf{U}_k$ is unitary, it holds that $\mathrm{tr}(\bar{\mathbf{S}}_k) = \mathrm{tr}(\bar{\mathbf{D}}_k)$ and that $||\bar{\mathbf{S}}_k - \tilde{\mathbf{S}}_k||_F = ||\bar{\mathbf{D}}_k - \tilde{\mathbf{D}}_k||_F$. That is to say, (91) is equivalent to

$$\begin{aligned} \text{minimize} \quad & \textstyle\sum_{k=1}^K ||\bar{\mathbf{D}}_k - \tilde{\mathbf{D}}_k||_F^2 \\ \text{subject to} \quad & \textstyle\sum_{k=1}^K \mathrm{tr}(\bar{\mathbf{D}}_k) = P; \{\bar{\mathbf{D}}_k \succeq \mathbf{0}\}. \end{aligned} \tag{92}$$

It is easy to see that $\bar{\mathbf{D}}_k$ must be diagonal to minimize the objective of (92). Next let $\bar{\mathbf{d}}_k = \mathrm{diag}(\bar{\mathbf{D}}_k)$, $\tilde{\mathbf{d}}_k = \mathrm{diag}(\tilde{\mathbf{D}}_k)$, $\bar{\mathbf{d}} = [\bar{\mathbf{d}}_1^T, \bar{\mathbf{d}}_2^T, \ldots, \bar{\mathbf{d}}_K^T]^T$, and $\tilde{\mathbf{d}} = [\tilde{\mathbf{d}}_1^T, \tilde{\mathbf{d}}_2^T, \ldots, \tilde{\mathbf{d}}_K^T]^T$. Then (92) can be reduced to

$$\begin{aligned} \text{minimize} \quad & \tfrac{1}{2}||\bar{\mathbf{d}} - \tilde{\mathbf{d}}||_2^2 \\ \text{subject to} \quad & \mathbf{1}_{\tilde{M}}\bar{\mathbf{d}} = P; \bar{\mathbf{d}} \geq 0 \end{aligned} \tag{93}$$

where $\tilde{M} = \sum_1^K M_k$. It is now clear that (93) is the projection onto a canonical simplex and efficient algorithms can be found in [35].

REFERENCES

[1] T. M. Pham, R. Farrell, and L.-N. Tran, "Alternating optimization for capacity region of Gaussian MIMO broadcast channels with per-antenna power constraint," in *Proc. IEEE VTC-Spring*, Jun. 2017, pp. 1–6.

[2] ——, "Low-complexity approaches for MIMO capacity with per-antenna power constraint," in *Proc. IEEE VTC-Spring*, Jun. 2017, pp. 1–7.

[3] E. Telatar, "Capacity of multi-antenna Gaussian channels," *Eur. Trans. Telecommun*, vol. 10, pp. 585–598, Nov. 1999.

[4] G. J. Foschini and M. J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless Pers.Commun*, vol. 6, pp. 311–335, Mar. 1998.

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

[6] G. Caire and S. Shamai, "On the achievable throughput of a multiantenna Gaussian broadcast channel," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1691–1706, Jul. 2003.

[7] A. Dabbagh and D. Love, "Precoding for multiple antenna Gaussian broadcast channels with successive zero-forcing," *IEEE Trans. Signal Process.*, vol. 55, no. 7, pp. 3837–3850, Jul. 2007.

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

[9] N. Jindal, W. Rhee, S. Vishwanath, S. Jafar, and A. Goldsmith, "Sum power iterative water-filling for multi-antenna Gaussian broadcast channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1570–1580, Apr. 2005.

[10] J. Liu, Y. T. Hou, and H. D. Sherali, "On the maximum weighted sum-rate of MIMO Gaussian broadcast channels," in *Proc. IEEE ICC*, May 2008, pp. 3664 – 3668.

[11] R. Hunger, D. A. Schmidt, M. Joham, and W. Utschick, "A general covariance-based optimization framework using orthogonal projections," in *Proc. IEEE SPAWC*, Jul. 2008, pp. 76 – 80.

[12] T. M. Pham, and R. J. Farrell, and J. Dooley, and E. Dutkiewicz, and D. N. Nguyen, and L.-N. Tran, "Efficient zero-forcing precoder design for weighted sum-rate maximization with per-antenna power constraint," *IEEE Trans. Veh. Technol.*, vol. PP, no. 99, pp. 1–1, 2017.

[13] T. E. Bogale and L. Vandendorpe, "Weighted sum rate optimization for downlink multiuser MIMO systems with per antenna power constraint: Downlink-uplink duality approach," in *Proc. IEEE ICASSP*, Mar. 2012, pp. 3245 – 3248.

[14] Q. Spencer, A. Swindlehurst, and M. Haardt, "Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels," *IEEE Trans. Signal Process.*, vol. 52, no. 2, pp. 461–471, Feb. 2004.

[15] B. Li, C. Z. Wu, H. H. Dam, A. Cantoni, and K. L. Teo, "A parallel low complexity zero-forcing beamformer design for multiuser MIMO

systems via a regularized dual decomposition method," *IEEE Trans. Signal Process.*, vol. 63, no. 16, pp. 4179 – 4190, Aug. 2015.
[16] S. Shi, M. Schubert, and H. Boche, "Per-antenna power constrained rate optimization for multiuser MIMO systems," in *Proc. WSA*, Feb. 2008, pp. 270 – 277.
[17] A. Wiesel, Y. C. Eldar, and S. Shamai, "Zero-forcing precoding and generalized inverses," *IEEE Trans. Signal Process.*, vol. 56, no. 9, pp. 4409 – 4418, Sep. 2008.
[18] L.-N. Tran, M. Juntti, M. Bengtsson, and B. Ottersten, "Weighted sum rate maximization for MIMO broadcast channels using dirty paper coding and zero-forcing methods," *IEEE Trans. Commun.*, vol. 61, no. 6, pp. 2362–2373, Jun. 2013.
[19] ——, "Beamformer designs for MISO broadcast channels with zero-forcing dirty paper coding," *IEEE Trans. Wireless Commun.*, vol. 12, no. 3, pp. 1173–1185, Mar. 2013.
[20] M. Vu, "MIMO capacity with per-antenna power constraint," in *Proc. IEEE GLOBECOM*, Dec. 2011, pp. 1 – 5.
[21] ——, "The capacity of MIMO channels with per-antenna power constraint," *CoRR*, vol. abs - 1106 - 5039, 2011.
[22] W. Yu and T. Lan, "Transmitter optimization for the multi-antenna downlink with per-antenna power constraints," *IEEE Trans. Signal Process.*, vol. 55, no. 6, pp. 2646–2660, Jun. 2007.
[23] L. Vandenberghe, S. Boyd, and S.-P. Wu, "Determinant maximization with linear matrix inequality constraints," *SIAM J. on Matrix Anal. and Appl.*, vol. 19, no. 2, pp. 499–533, 1998.
[24] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization*. Soc. Ind. Appl. Math. (SIAM), 2001.
[25] W. Yu, "Sum-capacity computation for the Gaussian vector broadcast channel via dual decomposition," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 754 –759, Feb. 2006.
[26] G. H. Golub and C. F. V. Loan, *Matrix Computations*, 4th ed. The John Hopkins Univ. Press, 2013.
[27] W. Yu, "Uplink-downlink duality via minimax duality," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 361–374, Feb. 2006.
[28] S. Vishwanath, N. Jindal, and A. Goldsmith, "Duality, achievable rates and sum-rate capacity of Gaussian MIMO broadcast channels," *IEEE Trans. Inf. Theory*, vol. 49, pp. 2658–2668, Oct. 2003.
[29] A. Beck, A. Ben-Tal, and L. Tetruashvili, "A sequential parametric convex approximation method with applications to nonconvex truss topology design problems," *Journal of Global Optimization*, vol. 47, no. 1, pp. 29–51, 2010.
[30] Z. Shen, R. Chen, J. Andrews, J. Heath, R.W., and B. Evans, "Low complexity user selection algorithms for multiuser MIMO systems with block diagonalization," *IEEE Trans. Signal Process.*, vol. 54, no. 9, pp. 3658–3663, Sep. 2006.
[31] L.-N. Tran, M. Bengtsson, and B. Ottersten, "Iterative precoder design and user scheduling for block-diagonalized systems," *IEEE Trans. Signal Process.*, vol. 60, no. 7, pp. 3726–3739, Jul. 2012.
[32] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.
[33] L.-N. Tran and E.-K. Hong, "Multiuser diversity for successive zero-forcing dirty paper coding: Greedy scheduling algorithms and asymptotic performance analysis," *IEEE Trans. Signal Process.*, vol. 58, no. 6, pp. 3411–3416, Jun. 2010.
[34] R. Fletcher, *Practical Methods of Optimization*, 2nd ed. New York, NY, USA: Wiley-Interscience, 1987.
[35] L. Condat, "Fast projection onto the simplex and the ℓ_1 ball," *Mathematical Programming, Series A*, vol. 158, no. 1, pp. 575 – 585, Jul. 2016.
[36] L. Armijo, "Minimization of functions having Lipschitz continuous first partial derivatives." *Pacific J. Math.*, vol. 16, no. 1, pp. 1 – 3, 1966.
[37] J. Liu, Y. T. Hou, S. Kompella, and H. D. Sherali, "Conjugate gradient projection approach for MIMO Gaussian broadcast channels," in *Proc. IEEE ISIT*, Jun. 2007, pp. 781 – 785.
[38] M. ApS, *The MOSEK optimization toolbox for MATLAB manual. Version 7.1 (Revision 28).*, 2015.
[39] P. Kyösti, J. Meinilä, L. Hentilä, X. Zhao, T. Jämsä, C. Schneider, M. Narandzić, M. Milojević, A. Hong, J. Ylitalo, V.-M. Holappa, M. Alatossava, R. Bultitude, Y. de Jong, and T. Rautiainen, "Winner II channel models," *tech. rep. D1.1.2 V1.2, IST-4-027756 WINNER II*, 2007.
[40] R. D. Yates, "A framework for uplink power control in cellular radio systems," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 7, pp. 1341–1347, Sep. 1995.
[41] R. A. Horn and C. R. Johnson, *Matrix Analysis*. New York, NY, USA: Cambridge University Press, 1986.
[42] R. G. Bartle and D. R. Sherbert, *Introduction to real analysis*, 4th ed. New York, NY, USA: John Wiley & Sons, Inc., 2011.
[43] P. M. Fitzpatrick, *Advanced calculus*, 2th ed. USA: Thomson Brooks/Cole, 2006.

PLACE PHOTO HERE

Thuy M. Pham (S'18) received the M.S. degree in information technology from University of Ulsan, South Korea in 2012. He is currently pursuing a Ph.D. in wireless communications at Maynooth University, Ireland. Between March and August 2018, he worked for Airrays GmbH., Germany, as a part-time research engineer on an LTE project. His research interests include ad hoc wireless routing protocols and wireless communications.

PLACE PHOTO HERE

Ronan Farrell (S'89–M'93) received the Ph.D. degree in electronic engineering from the University College Dublin, Dublin, Ireland, in 1998. Since 2000, he has been with Maynooth University, Maynooth, Ireland, where he is currently a Professor with the Department of Electronic Engineering. He is a Co-Principal Investigator with the Science Foundation Ireland CONNECT Centre for the Internet of Things. He leads a research team in the area of high-frequency radio systems with a particular interest in high-performance systems, power amplifiers, and applications of millimeter wave communications. He has co-authored over 200 journal and conference papers.

PLACE PHOTO HERE

Le-Nam Tran (M'10–SM'17) received the B.S. degree in electrical engineering from Ho Chi Minh City University of Technology, Ho Chi Minh City, Vietnam, in 2003, and the M.S. and Ph.D. degrees in radio engineering from Kyung Hee University, Seoul, Korea, in 2006 and 2009, respectively. He is currently a Lecturer/Assistant Professor at the School of Electrical and Electronic Engineering, University College Dublin, Ireland. Prior to this, he was a Lecturer at the Department of Electronic Engineering, Maynooth University, Co. Kildare, Ireland. From 2010 to 2014, he had held postdoc positions at the Signal Processing Laboratory, ACCESS Linnaeus Centre, KTH Royal Institute of Technology, Stockholm, Sweden (2010-2011), and at Centre for Wireless Communications, University of Oulu, Finland (2011-2014). His research interests are mainly on applications of optimization techniques on wireless communications design. Some recent topics include energy-efficient communications, physical layer security, cloud radio access networks, massive MIMO, and full-duplex transmission. He has authored or co-authored in some 80 papers published in international journals and conference proceedings.

Dr. Tran is an Associate Editor of EURASIP Journal on Wireless Communications and Networking. He was Symposium Co-Chair of Cognitive Computing and Networking Symposium of International Conference on Computing, Networking and Communication (ICNC 2016).