A Unified MIMO Optimization Framework Relying on the KKT Conditions

Shiqi Gong, Chengwen Xing, *Member, IEEE*, Yindi Jing, *Senior Member, IEEE*, Shuai Wang, Jiaheng Wang, *Senior Member, IEEE*, Sheng Chen, *Fellow, IEEE*, and Lajos Hanzo, *Fellow, IEEE*

***Abstract*— A popular technique of designing multiple-input multiple-output (MIMO) communication systems relies on optimizing the positive semidefinite covariance matrix at the source. In this paper, a unified MIMO optimization framework based on the Karush-Kuhn-Tucker (KKT) conditions is proposed. In this framework, with the aid of matrix optimization theory, Theorem 1 presents a generic optimal transmit covariance matrix for MIMO systems with diverse objective functions subject to various power constraints and different levels of channel state information (CSI). Specifically, Theorem 1 fundamentally reveals that for a diverse family of MIMO systems, the optimal transmit covariance matrices associated with different objective functions under various power constraints can be derived in a unified generic water-filling-like form. When applying Theorem 1 to the case of multiple general power constraints, we firstly equivalently transform multiple power constraints into a single counterpart by introducing multiple weighting factors based on Pareto optimization theory. The optimal weighting factors can be found by the proposed modified subgradient method. On the other hand, for the imperfect MIMO system with statistical CSI errors, we firstly address the non-convexity of the robust optimization problem by following the idea of alternating optimization. Finally, our numerical results verify the optimal solution structure in Theorem 1 and the global optimality of the proposed modified subgradient method, as well as demonstrate the performance advantages of the proposed alternating optimization algorithm.**

***Index Terms*— Convex optimization, MIMO communications, positive semi-definite matrix optimization, Karush-Kuhn-Tucker conditions.**

Manuscript received December 21, 2020; revised May 12, 2021 and July 16, 2021; accepted July 25, 2021. Date of publication August 5, 2021; date of current version November 18, 2021. This work was supported in part by the National Natural Science Foundation of China under Grants 61620106001, U1836201, 62071035, 61971130 and 61720106003. L. Hanzo would like to acknowledge the financial support of the Engineering and Physical Sciences Research Council projects EP/P034284/1 and EP/P003990/1 (COALESCE) as well as of the European Research Council's Advanced Fellow Grant QuantCom (Grant No. 789028). The associate editor coordinating the review of this article and approving it for publication was M. Bhatnagar. *(Corresponding author: Shuai Wang.)*

Shiqi Gong is with the School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China, and also with the State Key Laboratory of Internet of Things for Smart City, University of Macau, Taipa, Macao 999078, China (e-mail: shiqigong@um.edu.mo).

Chengwen Xing and Shuai Wang are with the School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China (e-mail: xingchengwen@gmail.com; swang@bit.edu.cn).

Yindi Jing is with the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB T6G 1H9, Canada (e-mail: yindi@ualberta.ca).

Jiaheng Wang is with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China (e-mail: jhwang@seu.edu.cn).

Sheng Chen is with the School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, U.K., and also with the Faculty of Engineering, King Abdulaziz University, Jeddah 21589, Saudi Arabia (e-mail: sxn@ecs.soton.ac.uk; sqc@ecs.soton.ac.uk).

Lajos Hanzo is with the School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, U.K. (e-mail: lh@ecs.soton.ac.uk).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCOMM.2021.3102641.

Digital Object Identifier 10.1109/TCOMM.2021.3102641

I. Introduction

Wireless technologies are rapidly developing relying on the novel concepts of cloud computing, green communications and so on in support of smart cities [1]–[4], [7]. In order to make these concepts come true, powerful physical layer techniques characterized by high spectrum and energy efficiencies are expected. Multiple-input multiple-output (MIMO) techniques have been widely regarded as one of the most important ingredients for a variety of wireless systems, which have also been investigated for a wide range of wireless applications [1]–[4], [7]–[14]. Without loss of generality, many design problems for MIMO communication systems aim to derive the optimal transceiver architectures [11]–[13]. Interestingly, despite considering different performance metrics or different power constraints, the optimal MIMO transmit precoding matrices usually have water-filling structures [2]–[4], [7]–[15]. Such optimal water-filling structures have the potential of greatly simplifying optimization problems and reducing the dimension of optimization variables. Futhermore, numerous variants of optimal water-filling structures have been discovered, such as general water-filling [16], polite water-filling [13], cluster water-filling [17], matrix-field water-filling [12], and cave water-filling [18]. Throughout this evolution, a large volume of elegant results have been derived for various MIMO scenarios, including point-to-point MIMO systems [3], multiuser MIMO systems [19], distributed MIMO networks [8], [20], and multi-hop amplify-and-forward MIMO relaying networks [21], [23]. It has been shown that the water-filling structures are also available even for some transceiver optimization problems in face of channel state information (CSI) errors [23]–[25].

Given a large body of MIMO literature, a natural question that arises is how these different water-filling structures are derived. Generally speaking, there are three categories of methods. The first category is the Karush-Kuhn-Tucker (KKT)-condition-based method [2], [24], [25]. Based on the fact that the KKT conditions constitute necessary conditions for optimal solutions [6], the common properties derived from KKT conditions are also the properties of optimal solutions. The second category is to apply matrix inequalities [9], [10] or majorization theory [3], [23], to reveal the inequalities between the diagonal elements and eigenvalues of a matrix [26]. Using majorization theory, when the design objectives are Schur-convex or Schur-concave functions of the diagonal elements of the mean-squared-error (MSE) matrix, the optimal matrix variables subject to the total power constraint can be derived in closed form [3], [23]. Unfortunately, this method usually suffers from strict limitations in the practical applications [23]. For example, if the optimization objective is neither Schur-convex nor Schur-concave or is subject to the multiple power constraints, the majorization theory may be inapplicable. The third category is termed as matrix-monotonic optimization, which exploits the properties of positive semi-definite matrices [21], [27]. Similarly, the matrix-monotonic optimization framework also has strict restrictions on the objective functions and constraints of optimization problems [21]. For example, when considering imperfect system CSI, the matrix-monotonic property of MIMO capacity with respect to the transmit covariance matrix may not hold due to the existence of CSI error [24]. Moreover, for the case of multiple conflicting power constraints, the matrix-monotonic optimization theory is not directly applicable, implying its limited application scope compared to the KKT-condition-based method.Generally speaking, the second and third kinds of methods are mostly adopted for the MIMO precoding matrix optimization with perfect CSI. From the information theory perspective, a MIMO precoding matrix in nature determines the covariance matrix of the transmitted signal. When we concentrate our attention on the transmit covariance matrix optimization, some hidden convexity will be revealed. Moreover, many elegant mathematical properties in the positive semi-definite cone can be used. Generally, we apply complex matrix derivatives to firstly derive the KKT conditions for the positive semi-definite matrix, from which the structures of the optimal solutions can be obtained. This kind of KKT-condition-based method has been widely adopted in [2], [11], [15], [24], [25].

Despite that the above three kinds of methods have been widely applied to derive the optimal transmit covariance (precoding) matrix in various MIMO systems, the unified KKT-condition-based framework is still essential due to the following two reasons. Firstly, it is known that the KKT conditions are necessary conditions for the optimal solution of any optimization problem, and even sufficient conditions for convex problems. Therefore, the KKT conditions are able to provide some insights for the optimal transceiver designs under diverse MIMO system setups. Secondly, most current studies aiming to derive the optimal MIMO transmit covariance (precoding) matrix using KKT conditions seem to be largely diverse in the optimization objective and the derivation of optimal solution [2], [24], [29]. For example, in [2], the weighted mean square error (MSE) minimization problem subject to the total transmit power constraint was studied, to which the optimal transmit precoder obtained from KKT conditions is able to diagonalize the MIMO channel into eigen subchannels. The authors of [24] extended the above work to the case of imperfect CSI at both communication ends. In this case, due to the existence of statistical CSI errors, the KKT-condition-based optimal transceiver design cannot realize the MIMO channel diagonalization. In contrast to [2], [24], the authors of [29] studied the capacity maximization problem with per-antenna constraints in the SDP framework and then analyzed its KKT optimality conditions.Generally, the theories and technologies for physical layer designs have common underlying fundamentals, hence, we believe that the above works having seemingly different mathematical derivations can be unified into a framework by revealing the underlying fundamental induced by their KKT-condition-based commonality.

It is known that most of existing MIMO system focus on the sum power constraint or per-antenna power constraints, both of which are suitable for centralized MIMO systems [2], [3], [25]. For further improving MIMO communication quality, there has been an upsurge of interest in distributed MIMO systems, where multiple multi-antenna users form a virtual antenna array to communicate with the multi-antenna base station [5], [12]. Since each user is powered by its own battery, multiple sum power constraints are more appropriate in this scenario. Furthermore, due to the distributed nature and varying circuit characteristics, it is also practical to assign different power weights to multiple antennas. In order to be compatible with these scenarios, our framework developes a general power constraint model.

In this paper, we investigate the widely used KKT-condition-based method for the transmit covariance matrix optimization of MIMO systems, and propose a unified framework for derivations of water-filling-like structured optimal solutions. Different from prior similar frameworks [3], [7], where specific Schur-concave and Schur-convex utility functions are considered under the assumption of perfect CSI, our framework not only proposes the general formulation for the MIMO transmit covariance matrices optimization with diverse utility functions, general power constraints and even different CSI levels, but also explores analytical solutions for MIMO system optimization subject to general power constraints by using Pareto optimization theory and Lagrange duality. Furthermore, armed with the classical successive convex approximation (SCA) technique, our framework also provides new ideas to tackle nonconvex MIMO optimization problems subject to multiple power constraints with low complexity. In particular, we include Table I for demonstratively and intuitively clarifying the generality of our work. The main contributions of our work are then summarized as follows.

- *Firstly*, compared to [2], [24], [25], the proposed framework of KKT conditions based positive semi-definite matrix optimization has a much simpler mathematical formula, which clearly reveals the relationships among various optimal solutions of the positive semi-definite

TABLE I

COMPARISONS BETWEEN OUR WORK AND MOST EXISTING RESEARCHES ON MIMO SYSTEM OPTIMIZATION

MIMO system	Details	Our work	Other works
Design objectives	Total MSE minimization	✓	[2], [3], [5], [9], [21], [28]
	Robust MSE minimization	✓	[17], [24], [30]
	Capacity maximization	✓	[1], [3], [5], [7], [8], [11], [19], [21]
	Robust capacity maximization	✓	[21], [25]
Power constraints	Total power constraint	✓	[1]–[3], [5], [7], [9], [11], [12], [21], [24], [25], [28], [30]
	Per-antenna power constraints	✓	[5], [12], [19]
	General power constraints	✓	[5], [12]
Optimal solutions	General water-filling structure	Unified KKT conditions based water-filling structure	[1]–[3], [5], [7], [8], [11], [19], [21]
	Cluster water-filling structure		[17]
	Matrix-field water-filling structure		[12]

matrix optimization with largely different mathematical formulas.

- *Secondly*, the proposed framework has a wide range of applications. It is applicable to many positive semi-definite optimization problems in MIMO systems relying on different optimization objectives, such as capacity maximization and MSE minimization, under both single and multiple power constraints. Moreover, different levels of CSI are considered.
- *Thirdly*, in view of the difficulty of simultaneously satisfying multiple power constraints considered in this framework, To tackle this issue, +30.2 we jointly apply Pareto optimization theory and Lagrange duality to integrate multiple power constraints into a single one associated with multiple weighting factors whose optimal values can be found by the modified subgradient method.
- *Finally*, due to the non-convexity of robust optimization problems under general transmit and receive spatial correlations, the globally optimal solution cannot be analytically derived. As such, we propose utilizing the SCA technique or the variable substitution to make this framework applicable, based on which the iterative optimization algorithms are proposed for finding the locally optimal solutions. In particular, when considering only transmit or receive spatial correlation, the robust analytical solutions under multiple weighted power constraints are firstly revealed.

Notation: Throughout this paper, $(\cdot)^*$, $(\cdot)^{\mathrm{T}}$ and $(\cdot)^{\mathrm{H}}$ stand for the conjugate, transpose and Hermitian transpose operators, respectively, while $\mathrm{Tr}(\boldsymbol{Z})$ and $|\boldsymbol{Z}|$ are the trace and determinant of the matrix $\boldsymbol{Z}$, respectively. $[\boldsymbol{Z}]_{:,1:N}$ denotes the first N columns of $\boldsymbol{Z}$, while the sub-matrix $[\boldsymbol{Z}]_{1:N,1:N}$ consists of the first N rows and the first N columns of $\boldsymbol{Z}$. $[\boldsymbol{Z}]_{i,i}$ denotes the ith diagonal element of $\boldsymbol{Z}$. $\boldsymbol{Z}^{\frac{1}{2}}$ is the Hermitian square root of the positive semi-definite matrix $\boldsymbol{Z}$. $\lambda_i(\boldsymbol{Z})$ is the ith largest eigenvalue of $\boldsymbol{Z}$, while $\sigma_i(\boldsymbol{Z})$ is the ith largest singular value of $\boldsymbol{Z}$. For two matrices $\boldsymbol{Z}$ and $\boldsymbol{A}$, $\boldsymbol{Z} \succeq \boldsymbol{A}$ means that $\boldsymbol{Z}-\boldsymbol{A}$ is positive semidefinite. $\mathbb{E}\{\cdot\}$ denotes the expectation operation and $(a)^+ = \max\{0, a\}$. To clarify the order of the eigenvalues or singular values, we use $\boldsymbol{\Lambda} \searrow$ to represent a rectangular diagonal matrix with the diagonal elements arranged in decreasing order. The words "with respect to", "independent and identically distributed" and "circularly symmetric complex Gaussian" are abbreviated as "w.r.t.", "i.i.d." and "CSCG", respectively.

II. FUNDAMENTAL RESULTS ON MATRIX OPTIMIZATION

For a convex optimization problem, the KKT conditions are both necessary and sufficient for the optimal solution. When the studied problem is nonconvex, the KKT conditions are only necessary but not sufficient for the optimal solution. However, even in this case, the KKT conditions can still be very useful. This is because if the solution structure is derived from the KKT conditions, then all solutions satisfying the KKT conditions have this structure. Our proposed framework aims to provide a fundamental result for the generic MIMO transmit covariance matrix optimization problem, whose KKT conditions are given by[1]

$$\begin{cases}\boldsymbol{H}^{\mathrm{H}}\boldsymbol{\Pi}^{-\frac{1}{2}}\left(\boldsymbol{I}+\boldsymbol{\Pi}^{-\frac{1}{2}}\boldsymbol{H}\boldsymbol{Q}\boldsymbol{H}^{\mathrm{H}}\boldsymbol{\Pi}^{-\frac{1}{2}}\right)^{-K}\boldsymbol{\Pi}^{-\frac{1}{2}}\boldsymbol{H}=\mu\boldsymbol{\Phi}-\boldsymbol{\Psi},\\ \mathrm{Tr}(\boldsymbol{\Psi}\boldsymbol{Q})=0 \text{ or } \boldsymbol{Q}^{\frac{1}{2}}\boldsymbol{\Psi}\boldsymbol{Q}^{\frac{1}{2}}=\boldsymbol{0}\end{cases} \tag{1}$$

where K can be any positive integer. $\boldsymbol{H}$ and μ are the arbitraty matrix and the positive scalar, respectively. The matrices $\{\boldsymbol{Q}, \boldsymbol{\Psi}, \boldsymbol{\Pi}\}$ are positive semidefinite, while the matrix $\boldsymbol{\Phi}$ is positive definite. Note that by illustrating the physical meanings of all matrices involved in (1) and choosing an appropriate value of K, the KKT conditions in (1) are specified as those for the capacity maximization problem or for the MSE minimization problem. Specifically, we consider that $\boldsymbol{Q} \in \mathbb{C}^{N_t \times N_t}$ is the positive semidefinite transmit covariance matrix to be optimized and $\boldsymbol{H} \in \mathbb{C}^{N_r \times N_t}$ is the MIMO channel matrix. $\boldsymbol{\Phi} \in \mathbb{C}^{N_t \times N_t}$ is the power weighting matrix and $\boldsymbol{\Pi} \in \mathbb{C}^{N_r \times N_r}$ is the interference-plus-noise covariance matrix. The matrix $\boldsymbol{\Psi} \in \mathbb{C}^{N_r \times N_r}$ denotes the Lagrange multiplier associated with $\boldsymbol{Q} \succeq \boldsymbol{0}$. Under this setup, the KKT conditions of the capacity maximization problem and the MSE minimization problem can be obtained from (1) by setting $K=1$ and $K=2$, respectively. Further, based on the formulation (1) and Lemma 1, we present a fundamental result in the following theorem, which is useful for deriving the optimal solution to the generic MIMO optimization problem.

Lemma 1: If all the solutions satisfying the KKT conditions have the same structure, this structure is definitely satisfied by the optimal solutions of the associated optimization problem.

Theorem 1: Using the singular value decomposition (SVD), i.e. $\boldsymbol{\Pi}^{-\frac{1}{2}}\boldsymbol{H}\boldsymbol{\Phi}^{-\frac{1}{2}} = \boldsymbol{U}_{\boldsymbol{\mathcal{H}}}\boldsymbol{\Lambda}_{\boldsymbol{\mathcal{H}}}\boldsymbol{V}_{\boldsymbol{\mathcal{H}}}^{\mathrm{H}}$ *with* $\boldsymbol{\Lambda}_{\boldsymbol{\mathcal{H}}} \searrow$, *the optimal*

[1]Since both $\boldsymbol{Q}$ and $\boldsymbol{\Psi}$ are positive semidefinite, we readily have $\boldsymbol{Q}^{\frac{1}{2}}\boldsymbol{\Psi}\boldsymbol{Q}^{\frac{1}{2}} \succeq \boldsymbol{0}$. It then directly follows from $\mathrm{Tr}(\boldsymbol{\Psi}\boldsymbol{Q}) = \mathrm{Tr}(\boldsymbol{Q}^{\frac{1}{2}}\boldsymbol{\Psi}\boldsymbol{Q}^{\frac{1}{2}}) = 0$ that $\boldsymbol{Q}^{\frac{1}{2}}\boldsymbol{\Psi}\boldsymbol{Q}^{\frac{1}{2}} = \boldsymbol{0}$, and vice verse.

$\boldsymbol{Q}$ satisfying both equations in (1) has the following water-filling structure

$$\boldsymbol{Q} = \boldsymbol{\Phi}^{-\frac{1}{2}} \left[\boldsymbol{V}_{\boldsymbol{\mathcal{H}}}\right]_{:,1:N} \Big(\mu^{-\frac{1}{K}} \left[\boldsymbol{\Lambda}_{\boldsymbol{\mathcal{H}}}\right]_{1:N,1:N}^{\frac{2}{K}-2} - \left[\boldsymbol{\Lambda}_{\boldsymbol{\mathcal{H}}}\right]_{1:N,1:N}^{-2}\Big)^{+} \left[\boldsymbol{V}_{\boldsymbol{\mathcal{H}}}\right]_{:,1:N}^{\mathrm{H}} \boldsymbol{\Phi}^{-\frac{1}{2}}, \quad (2)$$

where $N = \mathrm{Rank}(\boldsymbol{H})$ and $(\boldsymbol{B})^{+}$ denotes the operation $(\cdot)^{+}$ to every element of $\boldsymbol{B}$.

Proof: Upon multiplying both sides of the first equation in (1) by $\boldsymbol{Q}^{\frac{1}{2}}$ and referring to the second equation in (1), we arrive at the following equation.

$$\boldsymbol{Q}^{\frac{1}{2}} \boldsymbol{H}^{\mathrm{H}} \boldsymbol{\Pi}^{-\frac{1}{2}} \big(\boldsymbol{I} + \boldsymbol{\Pi}^{-\frac{1}{2}} \boldsymbol{H} \boldsymbol{Q} \boldsymbol{H}^{\mathrm{H}} \boldsymbol{\Pi}^{-\frac{1}{2}}\big)^{-K} \boldsymbol{\Pi}^{-\frac{1}{2}} \boldsymbol{H} \boldsymbol{Q}^{\frac{1}{2}} = \mu \boldsymbol{Q}^{\frac{1}{2}} \boldsymbol{\Phi} \boldsymbol{Q}^{\frac{1}{2}}. \quad (3)$$

Let us define the new matrix $\boldsymbol{A} = \boldsymbol{\Phi}^{\frac{1}{2}} \boldsymbol{Q}^{\frac{1}{2}}$, then equation (3) can be rewritten as

$$\boldsymbol{A}^{\mathrm{H}} \boldsymbol{\Phi}^{-\frac{1}{2}} \boldsymbol{H}^{\mathrm{H}} \boldsymbol{\Pi}^{-\frac{1}{2}} \big(\boldsymbol{I} + \boldsymbol{\Pi}^{-\frac{1}{2}} \boldsymbol{H} \boldsymbol{\Phi}^{-\frac{1}{2}} \boldsymbol{A} \boldsymbol{A}^{\mathrm{H}} \boldsymbol{\Phi}^{-\frac{1}{2}} \boldsymbol{H}^{\mathrm{H}} \boldsymbol{\Pi}^{-\frac{1}{2}}\big)^{-K} \times \boldsymbol{\Pi}^{-\frac{1}{2}} \boldsymbol{H} \boldsymbol{\Phi}^{-\frac{1}{2}} \boldsymbol{A} = \mu \boldsymbol{A}^{\mathrm{H}} \boldsymbol{A}. \quad (4)$$

According to the SVDs of $\boldsymbol{A}$ and $\boldsymbol{\Pi}^{-\frac{1}{2}} \boldsymbol{H} \boldsymbol{\Phi}^{-\frac{1}{2}} \boldsymbol{A}$, the equation (4) implies that $\boldsymbol{A}$ and $\boldsymbol{\Pi}^{-\frac{1}{2}} \boldsymbol{H} \boldsymbol{\Phi}^{-\frac{1}{2}} \boldsymbol{A}$ have the same right SVD unitary matrix. Based on this fact, it can be concluded that the positive semi-definite matrices $\boldsymbol{A}^{\mathrm{H}} \boldsymbol{A}$ and $\boldsymbol{A}^{\mathrm{H}} \boldsymbol{\Phi}^{-\frac{1}{2}} \boldsymbol{H}^{\mathrm{H}} \boldsymbol{\Pi}^{-1} \boldsymbol{H} \boldsymbol{\Phi}^{-\frac{1}{2}} \boldsymbol{A}$ have the same eigenmatrix. Therefore, the left singular vectors of $\boldsymbol{A}$ corresponding to its nonzero singular values are also the right singular vectors of $\boldsymbol{\Pi}^{-\frac{1}{2}} \boldsymbol{H} \boldsymbol{\Phi}^{-\frac{1}{2}}$. It is worth noting that for zero singular values of $\boldsymbol{A}$, the corresponding right singular vectors can be arbitrary as long as they are orthogonal to each other and orthogonal to the ones corresponding to the nonzero singular values. Furthermore, we assume the following property without loss of optimality.

Property 1: The left singular matrix of $\boldsymbol{A}$ is the right singular matrix of $\boldsymbol{\Pi}^{-\frac{1}{2}} \boldsymbol{H} \boldsymbol{\Phi}^{-\frac{1}{2}}$.

Based on the definition of $\boldsymbol{A}$, the first equation in (1) is equivalent to the following one

$$\boldsymbol{\Phi}^{-\frac{1}{2}} \boldsymbol{H}^{\mathrm{H}} \boldsymbol{\Pi}^{-\frac{1}{2}} \big(\boldsymbol{I} + \boldsymbol{\Pi}^{-\frac{1}{2}} \boldsymbol{H} \boldsymbol{\Phi}^{-\frac{1}{2}} \boldsymbol{A} \boldsymbol{A}^{\mathrm{H}} \boldsymbol{\Phi}^{-\frac{1}{2}} \boldsymbol{H}^{\mathrm{H}} \boldsymbol{\Pi}^{-\frac{1}{2}}\big)^{-K} \times \boldsymbol{\Pi}^{-\frac{1}{2}} \boldsymbol{H} \boldsymbol{\Phi}^{-\frac{1}{2}} = \mu \boldsymbol{I} - \boldsymbol{\Phi}^{-\frac{1}{2}} \boldsymbol{\Psi} \boldsymbol{\Phi}^{-\frac{1}{2}}. \quad (5)$$

Armed with Property 1, it is seen that the right singular matrix of $\boldsymbol{\Pi}^{-\frac{1}{2}} \boldsymbol{H} \boldsymbol{\Phi}^{-\frac{1}{2}}$ is the eigenmatrix associated with the eigenvalue decomposition (EVD) of the left hand side of (5). Based on (5), the eigenmatrix of the left hand side of (5) is exactly the eigenmatrix of the right hand side of (5). In other words, the following property holds.

Property 2: The right singular matrix of $\boldsymbol{\Pi}^{-\frac{1}{2}} \boldsymbol{H} \boldsymbol{\Phi}^{-\frac{1}{2}}$ is the eigenmatrix of $\boldsymbol{\Phi}^{-\frac{1}{2}} \boldsymbol{\Psi} \boldsymbol{\Phi}^{-\frac{1}{2}}$.

Let us now define $a_i = \sigma_i(\boldsymbol{A})$, $h_i = \sigma_i\big(\boldsymbol{\Pi}^{-\frac{1}{2}} \boldsymbol{H} \boldsymbol{\Phi}^{-\frac{1}{2}}\big)$ and $\psi_i = \lambda_i\big(\boldsymbol{\Phi}^{-\frac{1}{2}} \boldsymbol{\Psi} \boldsymbol{\Phi}^{-\frac{1}{2}}\big)$. Then by exploiting Properties 1 and 2, the equation (5) becomes $h_i^2 / \big(1 + a_i^2\, h_i^2\big)^K = \mu - \psi_i$, based on which the parameters a_i^2's can be expressed as $a_i^2 = (h_i^{\frac{2}{K}} / h_i^2) \Big(1/(\mu - \psi_i)^{\frac{1}{K}} - 1/h_i^{\frac{2}{K}}\Big)$. Moreover, because of $\boldsymbol{Q}^{\frac{1}{2}} \boldsymbol{\Psi} \boldsymbol{Q}^{\frac{1}{2}} = \boldsymbol{0}$ and together with the definition $\boldsymbol{A} = \boldsymbol{\Phi}^{\frac{1}{2}} \boldsymbol{Q}^{\frac{1}{2}}$, we may conclude that $\boldsymbol{A}^{\mathrm{H}} \boldsymbol{\Phi}^{-\frac{1}{2}} \boldsymbol{\Psi} \boldsymbol{\Phi}^{-\frac{1}{2}} \boldsymbol{A} = \boldsymbol{0}$, which further implies $a_i^2 \psi_i = 0$ by exploiting Properties 1 and 2. Therefore, ψ_i can be removed from the expression of a_i^2. The reason is as follows. Combining the inequalities $\mu > 0$, $\psi_i \geq 0$ with $a_i^2 \psi_i = 0$, it may be inferred that $\psi_i = 0$ when $a_i^2 > 0$, or $\psi_i > 0$ when $a_i^2 = 0$, based on which we have $a_i^2 = \frac{h_i^{\frac{2}{K}}}{h_i^2} \left(\frac{1}{\mu^{\frac{1}{K}}} - \frac{1}{h_i^{\frac{2}{K}}}\right)^{+}$. Upon recalling the definitions of $\{a_i, h_i, \psi_i\}$, we finally obtain the optimal water-filling structured $\boldsymbol{Q}$ as in (2). □

Remark: It is worth emphasizing that Theorem 1 is very general and independent of the specific MIMO system setups, including the objective functions, power constraints, signal models, and channel assumptions. Moreover, in Theorem 1, the matrices $\boldsymbol{\Pi}$ and $\boldsymbol{\Phi}$ are not restricted to be constant, they can also be functions of $\boldsymbol{Q}$.

A. Differences From the Existing Literature

As water-filling structures have been extensively studied, we would like to discuss the main differences between our derivations and the existing ones. To the best of our knowledge, the relevant literature may be classified into the following two categories.

1) Comparison to the First Category: The first category of the existing water-filling structure derivations is based on matrix inequalities, e.g., [1], [3], [28]. However, it is usually difficult to guarantee that the extreme values of matrix inequalities are actually achieved due to the variations in the objectives or constraints of the associated optimization problems [28]. For example, in Telatar's paper [1], the matrix inequality $\log\left|\boldsymbol{I} + \boldsymbol{H}\boldsymbol{Q}\boldsymbol{H}^{\mathrm{H}}\right| \leq \sum_i \log\big(1 + \lambda_i(\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H})\lambda_i(\boldsymbol{Q})\big)$ is applied to derive the water-filling structure of the optimal solution. If the sum power constraint $\mathrm{Tr}(\boldsymbol{Q}) \leq P$ is replaced by $[\boldsymbol{Q}]_{i,i} \leq P_i$, the equality cannot be achieved [28]. In Section III, we will show that this issue can be overcome by Theorem 1, because $\boldsymbol{\Pi}$ and $\boldsymbol{\Phi}$ in Theorem 1 can both be functions of $\boldsymbol{Q}$. Therefore, compared to [1], [3], [28], our conclusions are more general.

2) Comparison to the Second Category: The second category is purely based on the classic KKT conditions and usually consists of two phases. In the first phase, most KKT conditions based studies have argued that when the product of two matrices, i.e. $\boldsymbol{\Lambda}_1\boldsymbol{\Lambda}_2$, is a Hermitian matrix and $\boldsymbol{\Lambda}_1$ is a diagonal matrix, then the matrix $\boldsymbol{\Lambda}_2$ is also diagonal [2], [24], [29]. However, if some diagonal elements of $\boldsymbol{\Lambda}_1$ are zeros, this claim does not hold. To avoid this issue, these studies also assume that $\boldsymbol{H}$ is of full rank. By contrast, we do not impose the full rank condition on $\boldsymbol{H}$ in Theorem 1. In the second phase, some matrix manipulations are applied for reducing the KKT conditions to some equations that only involve diagonal matrices, and then the optimal covariance matrix can be derived from these equations. Since the diagonal elements must be nonnegative, the operation $(a)^{+} = \max\{a, 0\}$ is introduced. In Theorem 1, this operation appears in the solution via rigorous mathematical derivation, which has clear physical interpretation in practical applications, namely that the transmit power of each eigenchannel is non-negative. By contrast, some existing KKT conditions based derivations directly apply $(a)^+ = \max\{a, 0\}$ to the eigenvalues of the positive semi-definite matrix $\boldsymbol{Q}$, which may be incorrect in some applications [2], [24], [29]. To see this, let us consider the following inequality

$$\begin{aligned}&\boldsymbol{\Phi}^{-\frac{1}{2}}\big[\boldsymbol{V}_{\boldsymbol{\mathcal{H}}}\big]_{:,1:N}\big(\mu^{-\frac{1}{K}}\big[\boldsymbol{\Lambda}_{\boldsymbol{\mathcal{H}}}\big]_{1:N,1:N}^{\frac{2}{K}-2}-\big[\boldsymbol{\Lambda}_{\boldsymbol{\mathcal{H}}}\big]_{1:N,1:N}^{-2}\big)^{+}\\&\quad\times\big[\boldsymbol{V}_{\boldsymbol{\mathcal{H}}}\big]_{:,1:N}^{\mathrm{H}}\times\boldsymbol{\Phi}^{-\frac{1}{2}}\\&\quad\neq\bigg(\boldsymbol{\Phi}^{-\frac{1}{2}}\big[\boldsymbol{V}_{\boldsymbol{\mathcal{H}}}\big]_{:,1:N}\big(\mu^{-\frac{1}{K}}\big[\boldsymbol{\Lambda}_{\boldsymbol{\mathcal{H}}}\big]_{1:N,1:N}^{\frac{2}{K}-2}\\&\qquad-\big[\boldsymbol{\Lambda}_{\boldsymbol{\mathcal{H}}}\big]_{1:N,1:N}^{-2}\big)\big[\boldsymbol{V}_{\boldsymbol{\mathcal{H}}}\big]_{:,1:N}^{\mathrm{H}}\boldsymbol{\Phi}^{-\frac{1}{2}}\bigg)^{+},\end{aligned}\tag{6}$$

where the equality holds only when $\boldsymbol{\Phi} \propto \boldsymbol{I}$. In particular, in some existing treatises, $\boldsymbol{Q}$ is instead replaced by $\boldsymbol{F}\boldsymbol{F}^{\mathrm{H}}$, where $\boldsymbol{F}$ is a tall matrix implying $\boldsymbol{Q}$ is rank-deficient. Then we have the following KKT condition

$$\boldsymbol{H}^{\mathrm{H}}\boldsymbol{\Pi}^{-\frac{1}{2}}\big(\boldsymbol{I}+\boldsymbol{\Pi}^{-\frac{1}{2}}\boldsymbol{H}\boldsymbol{F}\boldsymbol{F}^{\mathrm{H}}\boldsymbol{H}^{\mathrm{H}}\boldsymbol{\Pi}^{-\frac{1}{2}}\big)^{-1}\boldsymbol{\Pi}^{-\frac{1}{2}}\boldsymbol{H}\boldsymbol{F}=\mu\boldsymbol{\Phi}\boldsymbol{F}.\tag{7}$$

In this case, the optimal water-filling structured $\boldsymbol{F}$ cannot be achieved due to the turning-off effect [28] which assumes that any eigenchannel can be turned off (allocated zero power). In this case, although (7) can still be satisfied, it is not equivalent to the equation $\boldsymbol{H}^{\mathrm{H}}\boldsymbol{\Pi}^{-\frac{1}{2}}\big(\boldsymbol{I} + \boldsymbol{\Pi}^{-\frac{1}{2}}\boldsymbol{H}\boldsymbol{F}\boldsymbol{F}^{\mathrm{H}}\boldsymbol{H}^{\mathrm{H}}\boldsymbol{\Pi}^{-\frac{1}{2}}\big)^{-1}\boldsymbol{\Pi}^{-\frac{1}{2}}\boldsymbol{H} = \mu\boldsymbol{\Phi}$, since the right inverse of $\boldsymbol{F}$ may not exist. As such, (7) can be readily seen to be inconsistent with (1), which makes Theorem 1 not workable.

3) Summary of Our Derivation: Based on the KKT conditions which are necessary conditions for finding the optimal solution to the corresponding optimization problem, Theorem 1 reveals that all solutions of $\boldsymbol{Q}$ satisfying both equations in (1) have the structure given by (2). The fundamental conclusion in Theorem 1 has the following properties: *1)* It does not require the studied optimization problem to be convex; *2)* It is applicable even when the involved parameters are either constants or functions of optimization variables; *3)* It provides a common structure of the solutions satisfying KKT conditions.

In the following two sections, we use Theorem 1 to derive the water-filling structures of the optimal MIMO transmit covariance matrices considering different objective functions, power constraints, and CSI assumptions. Specifically, for MIMO systems with perfect CSI, we investigate the capacity maximization and MSE minimization problems subject to multiple weighted power constraints. Additionally, the above work is extended to the scenario of realistic MIMO systems with non-negligible CSI errors. Based on Theorem 1, the optimal structures of the robust transmit covariance matrices can also be derived.

III. Transmit Covariance Matrix Optimization Under Perfect CSI

For point-to-point MIMO systems where both the source and destination are equipped with multiple antennas, the received signal vector $\boldsymbol{y}$ is given by

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{s} + \boldsymbol{n},\tag{8}$$

where $\boldsymbol{s}$ is the transmitted signal vector whose covariance matrix is $\boldsymbol{Q} = \mathbb{E}\{\boldsymbol{s}\boldsymbol{s}^{\mathrm{H}}\}$, and $\boldsymbol{H}$ is the channel matrix. $\boldsymbol{n}$ is the additive noise vector at the destination with the covariance matrix $\boldsymbol{R}_{\mathrm{n}} = \mathbb{E}\{\boldsymbol{n}\boldsymbol{n}^{\mathrm{H}}\}$. We intend to apply Theorem 1 to the transmit covariance matrix design relying on perfect CSI, where both single weighted sum power constraint and multiple weighted power constraints are considered. Moreover, two types of widely used design objectives are studied, namely capacity maximization and MSE minimization. We reinvestigate the well-addressed transmit covariance matrix optimization under a single power constraint mainly because the optimal solutions of the relevant problems can be used to validate our derivations in Theorem 1.

A. Single Weighted Sum Power Constraint

When the channel statistics of different antennas in a MIMO system are similar, the sum power constraint is indeed quite practical for transceiver optimization, which is also a special case of the weighted sum power constraint considering different characteristics of realistic transmit RF chains connected to different antennas [12]. We model the weighted sum power constraint as $\mathrm{Tr}(\boldsymbol{W}\boldsymbol{Q}) \leq P$, where P is the maximum transmit power and $\boldsymbol{W}$ is a positive definite weighting matrix. Note that if $\boldsymbol{W}$ has a zero eigenvalue, no power constraint is imposed on the corresponding eigenchannel, which is impractical.

1) Capacity Maximization: The capacity maximization problem under the weighted sum power constraint is formulated as

$$\begin{aligned}\textbf{P1}:\ \min_{\boldsymbol{Q}}\quad & -\log\left|\boldsymbol{I}+\boldsymbol{R}_{\mathrm{n}}^{-1}\boldsymbol{H}\boldsymbol{Q}\boldsymbol{H}^{\mathrm{H}}\right|\\ \mathrm{s.t.}\quad & \mathrm{Tr}(\boldsymbol{W}\boldsymbol{Q})\leq P,\ \boldsymbol{Q}\succeq\boldsymbol{0}.\end{aligned}\tag{9}$$

The full KKT conditions for **P1** are given by

$$\begin{cases}\boldsymbol{H}^{\mathrm{H}}\boldsymbol{R}_{\mathrm{n}}^{-\frac{1}{2}}\big(\boldsymbol{I}+\boldsymbol{R}_{\mathrm{n}}^{-\frac{1}{2}}\boldsymbol{H}\boldsymbol{Q}\boldsymbol{H}^{\mathrm{H}}\boldsymbol{R}_{\mathrm{n}}^{-\frac{1}{2}}\big)^{-1}\boldsymbol{R}_{\mathrm{n}}^{-\frac{1}{2}}\boldsymbol{H}=\mu\boldsymbol{W}-\boldsymbol{\Psi},\\ \mu\geq0,\ \mu\big(\mathrm{Tr}(\boldsymbol{W}\boldsymbol{Q})-P\big)=0,\ \mathrm{Tr}(\boldsymbol{Q}\boldsymbol{\Psi})=0,\\ \boldsymbol{\Psi}\succeq\boldsymbol{0},\quad\mathrm{Tr}(\boldsymbol{W}\boldsymbol{Q})\leq P,\quad\boldsymbol{Q}\succeq\boldsymbol{0},\end{cases}\tag{10}$$

where μ and $\boldsymbol{\Psi}$ are the dual variables associated with the constraints $\mathrm{Tr}(\boldsymbol{W}\boldsymbol{Q}) \leq P$ and $\boldsymbol{Q} \succeq \boldsymbol{0}$, respectively. Based on Theorem 1 together with $\boldsymbol{\Pi} = \boldsymbol{R}_{\mathrm{n}}$ and $\boldsymbol{\Phi} = \boldsymbol{W}$, we have the following conclusion.

*Conclusion 1: The optimal $\boldsymbol{Q}$ for **P1** has the following water-filling structure*

$$\begin{aligned}\boldsymbol{Q}=\boldsymbol{W}^{-\frac{1}{2}}\big[\boldsymbol{V}_{\boldsymbol{\mathcal{H}}}\big]_{:,1:N}\Big(\mu^{-1}\boldsymbol{I}-\big[\boldsymbol{\Lambda}_{\boldsymbol{\mathcal{H}}}\big]_{1:N,1:N}^{-2}\Big)^{+}\\ \times\big[\boldsymbol{V}_{\boldsymbol{\mathcal{H}}}\big]_{:,1:N}^{\mathrm{H}}\boldsymbol{W}^{-\frac{1}{2}},\end{aligned}\tag{11}$$

where the unitary matrix $\boldsymbol{V}_{\boldsymbol{\mathcal{H}}}$ is defined by the SVD

$$\boldsymbol{R}_{\mathrm{n}}^{-\frac{1}{2}}\boldsymbol{H}\boldsymbol{W}^{-\frac{1}{2}}=\boldsymbol{U}_{\boldsymbol{\mathcal{H}}}\boldsymbol{\Lambda}_{\boldsymbol{\mathcal{H}}}\boldsymbol{V}_{\boldsymbol{\mathcal{H}}}^{\mathrm{H}}\ \textit{with}\ \boldsymbol{\Lambda}_{\boldsymbol{\mathcal{H}}}\searrow.\tag{12}$$

Computation of μ: Clearly, the optimal μ in Conclusion 1 should satisfy $\mathrm{Tr}(\boldsymbol{W}\boldsymbol{Q}) = \mathrm{Tr}\big((\mu^{-1}\boldsymbol{I} - [\boldsymbol{\Lambda}_{\boldsymbol{\mathcal{H}}}]_{1:N,1:N}^{-2})^{+}\big) = P$, and thus can be uniquely determined by a standard water-filling procedure. In fact, upon assuming descending eigenvalues in $\boldsymbol{\Lambda}_{\boldsymbol{\mathcal{H}}}$, the optimal μ is readily derived in closed form as $\mu = \frac{L_{\max}}{P+\sum_{l=1}^{L_{\max}}[\boldsymbol{\Lambda}_{\mathcal{H}}]_{l,l}^{-2}}$, where $L_{\max}$ denotes the maximum integer satisfying $\mu^{-1} - [\boldsymbol{\Lambda}_{\mathcal{H}}]_{L_{\max},L_{\max}}^{-2} \geq 0$.

2) MSE Minimization: From the perspective of fully exploiting the spatial multiplexing gain of MIMO systems, we assume the number of transmitted data streams to be $d = \mathrm{rank}(\boldsymbol{H})$. Then the MSE minimization problem is formulated as

$$\begin{aligned} \mathbf{P2}: \; & \min_{\boldsymbol{Q}} \;\; \mathrm{Tr}\left(\left(\boldsymbol{I} + \boldsymbol{R}_{\mathrm{n}}^{-1}\boldsymbol{H}\boldsymbol{Q}\boldsymbol{H}^{\mathrm{H}}\right)^{-1}\right) \\ & \text{s.t.} \;\; \mathrm{Tr}(\boldsymbol{W}\boldsymbol{Q}) \leq P, \; \boldsymbol{Q} \succeq \boldsymbol{0}. \end{aligned} \tag{13}$$

The full KKT conditions of **P2** are given by

$$\begin{cases} \boldsymbol{H}^{\mathrm{H}}\boldsymbol{R}_{\mathrm{n}}^{-\frac{1}{2}}\big(\boldsymbol{I}+\boldsymbol{R}_{\mathrm{n}}^{-\frac{1}{2}}\boldsymbol{H}\boldsymbol{Q}\boldsymbol{H}^{\mathrm{H}}\boldsymbol{R}_{\mathrm{n}}^{-\frac{1}{2}}\big)^{-2}\boldsymbol{R}_{\mathrm{n}}^{-\frac{1}{2}}\boldsymbol{H} = \mu\boldsymbol{W} - \boldsymbol{\Psi}, \\ \mu \geq 0, \; \mu(\mathrm{Tr}(\boldsymbol{W}\boldsymbol{Q}) - P) = 0, \; \mathrm{Tr}(\boldsymbol{Q}\boldsymbol{\Psi}) = 0, \\ \boldsymbol{\Psi} \succeq \boldsymbol{0}, \;\; \mathrm{Tr}(\boldsymbol{W}\boldsymbol{Q}) \leq P, \;\; \boldsymbol{Q} \succeq \boldsymbol{0}. \end{cases} \tag{14}$$

Again, based on Theorem 1 together with $\boldsymbol{\Pi} = \boldsymbol{R}_{\mathrm{n}}$ and $\boldsymbol{\Phi} = \boldsymbol{W}$, the following conclusion holds.

*Conclusion 2: The optimal $\boldsymbol{Q}$ for **P2** has the following water-filling structure*

$$\boldsymbol{Q} = \boldsymbol{W}^{-\frac{1}{2}}\left[\boldsymbol{V}_{\mathcal{H}}\right]_{:,1:N}\left(\mu^{-\frac{1}{2}}\left[\boldsymbol{\Lambda}_{\mathcal{H}}\right]_{1:N,1:N}^{-1} - \left[\boldsymbol{\Lambda}_{\mathcal{H}}\right]_{1:N,1:N}^{-2}\right)^{+} \times \left[\boldsymbol{V}_{\mathcal{H}}\right]_{:,1:N}^{\mathrm{H}}\boldsymbol{W}^{-\frac{1}{2}}. \tag{15}$$

Computation of μ: Similar to Section III-A.1, the optimal μ in Conclusion 2 satisfying $\mathrm{Tr}(\boldsymbol{W}\boldsymbol{Q}) = \mathrm{Tr}\big((\mu^{-\frac{1}{2}}\left[\boldsymbol{\Lambda}_{\mathcal{H}}\right]_{1:N,1:N}^{-1} - \left[\boldsymbol{\Lambda}_{\mathcal{H}}\right]_{1:N,1:N}^{-2})^{+}\big) = P$ can also be numerically computed using the classic water-filling procedure, and its closed-form solution is given by $\mu = \big(\frac{\sum_{l=1}^{L_{\max}}[\boldsymbol{\Lambda}_{\mathcal{H}}]_{l,l}^{-1}}{P+\sum_{l=1}^{L_{\max}}[\boldsymbol{\Lambda}_{\mathcal{H}}]_{l,l}^{-2}}\big)^{2}$, where $L_{\max}$ denotes the maximum integer satisfying $\mu^{-\frac{1}{2}}\left[\boldsymbol{\Lambda}_{\mathcal{H}}\right]_{L_{\max},L_{\max}}^{-1} - \left[\boldsymbol{\Lambda}_{\mathcal{H}}\right]_{L_{\max},L_{\max}}^{-2} \geq 0$.

B. Multiple Weighted Power Constraints

As each antenna in an antenna array has its own amplifier, the individual per-antenna power constraints are more practical than the sum power constraint. We next consider the general multiple weighted power constraints including the case of individual power constraints as a special case. The general multiple power constraints take the following form

$$\mathrm{Tr}(\boldsymbol{\Omega}_i\boldsymbol{Q}) \leq P_i, \quad 1 \leq i \leq I, \tag{16}$$

where the weighting matrices $\boldsymbol{\Omega}_i$'s are positive semidefinite and P_i's denote the corresponding maximum transmit power. In particular, when $\boldsymbol{\Omega}_i = \boldsymbol{b}_i\boldsymbol{b}_i^{\mathrm{H}}$ holds with $\boldsymbol{b}_i$ being the vector whose ith element is one and all other elements are zero, the power constraint of the ith antenna is obtained. Recall Section III-A, it is clear that the positive scalar μ in Theorem 1 corresponds to the weighted sum power constraint $\mathrm{Tr}(\boldsymbol{W}\boldsymbol{Q}) \leq P$. In order to extend Theorem 1 to the case of multiple power constraints, we consider integrating all these power constraints into a single counterpart under some additional requirements, as shown in the following Proposition. Using this operation, Theorem 1 becomes directly applicable to both capacity maximization and MSE minimization problems subject to general multiple power constraints.

Proposition 1: For a matrix-monotone decreasing convex function $f(\boldsymbol{Q})$, the following optimization problem[2]

$$\min_{\boldsymbol{Q}} \; f(\boldsymbol{Q}), \; \text{s.t.} \; \mathrm{Tr}(\boldsymbol{\Omega}_i\boldsymbol{Q}) \leq P_i, \; \boldsymbol{Q} \succeq \boldsymbol{0}, \;\; 1 \leq i \leq I, \tag{17}$$

can be equivalently simplified to

$$\begin{aligned} & \min_{\boldsymbol{Q}} \;\; f(\boldsymbol{Q}), \\ & \text{s.t.} \;\; \mathrm{Tr}(\boldsymbol{\Omega}\boldsymbol{Q}) \leq P, \;\; \boldsymbol{Q} \succeq \boldsymbol{0} \text{ with } P = \sum\nolimits_{i=1}^{I} P_i \text{ and} \\ & \qquad \boldsymbol{\Omega} = \sum\nolimits_{i=1}^{I} \mu_i\boldsymbol{\Omega}_i, \end{aligned} \tag{18}$$

where $\mu_i = (\lambda_i P)/(\sum_{i=1}^{I}\lambda_i P_i)$, $1 \leq i \leq I$. The optimal dual variables λ_i's associated with the general power constraints of problem (17) *satisfy $\lambda_i(\mathrm{Tr}(\boldsymbol{\Omega}_i\boldsymbol{Q}) - P_i) = 0$, $1 \leq i \leq I$. Note that $\boldsymbol{\Omega} \succ \boldsymbol{0}$ must hold for guaranteeing the bounded objective value $f(\boldsymbol{Q})$.*

Proof: Firstly, by referring to [5, Th. 1], the equivalence between problems (17) and (18) has been established for the general matrix-monotone decreasing function $f(\boldsymbol{Q})$ by using Pareto optimization theory. Specifically, it is observed from Appendix of [5] that the optimal $\boldsymbol{Q}$ to problem (17) is actually a Pareto optimal solution of a convex vector optimization problem, where multiple transmit power objectives, i.e. $\mathrm{Tr}(\boldsymbol{\Omega}_i\boldsymbol{Q})$, $1 \leq i \leq I$, are simultaneously minimized. According to [6, Ch. 4.7.4], the scalarization method can be applied to solve this vector optimization problem by defining $\boldsymbol{\Omega} = \sum\limits_{i=1}^{I}\mu_i\boldsymbol{\Omega}_i$ with nonnegative weighting coefficients μ_i's, based on which there exists a Pareto optimal solution $\boldsymbol{Q}$ corresponding to a certain set of μ_i's that achieves the optimal objective value $f(\boldsymbol{Q})$. Note that the corresponding optimal μ_i's are determined according to the following rules

$$\begin{aligned} \mu_i^{(t+1)} &= [\mu_i^{(t)} + a_i^{(t)}\big(\mathrm{Tr}(\boldsymbol{\Omega}_i\boldsymbol{Q}) - P_i\big)]^{+}, \\ & \mu_i^{(t+1)}(\mathrm{Tr}(\boldsymbol{\Omega}_i\boldsymbol{Q}) - P_i) = 0, \quad 1 \leq i \leq I, \end{aligned} \tag{19}$$

where $a_i^{(t)} > 0$ denotes the tth step size of the ith power constraint. Obviously, the weighting coefficients μ_i's are updated in a similar manner to the dual variables λ_i's associated with multiple power constraints. For a special class of convex functions $f(\boldsymbol{Q})$, we further apply the Lagrange dual theory to obtain an accurate mapping between μ_i and λ_i for any $1 \leq i \leq I$, based on which the subgradient method is expected to have faster convergence rate and guaranteed convergence. Specifically, the Lagrangian of the convex problem (17) is given by

$$\begin{aligned} L(\boldsymbol{Q}, \mathbf{Z}, \{\lambda_i\}) &= f(\boldsymbol{Q}) + \sum\nolimits_{i=1}^{N_S}\lambda_i(\mathrm{Tr}(\boldsymbol{\Omega}_i\boldsymbol{Q}) - P_i) - \mathrm{Tr}(\boldsymbol{\Psi}\boldsymbol{Q}) \\ &= f(\boldsymbol{Q}) + \tilde{\mu}\big(\mathrm{Tr}(\boldsymbol{\Omega}\boldsymbol{Q}) - P\big) - \mathrm{Tr}(\boldsymbol{\Psi}\boldsymbol{Q}). \end{aligned} \tag{20}$$

where $\tilde{\mu} = (\sum\limits_{i=1}^{I}\lambda_i P_i)/P$, $P = \sum\limits_{i=1}^{I}P_i$ and $\boldsymbol{\Omega} = \sum\limits_{i=1}^{I}\mu_i\boldsymbol{\Omega}_i$. The one-to-one mapping between μ_i and λ_i is given by

[2]For simplicity, the relevant conclusion in [5] is directly cited in the proof of this proposition. More details are referred to as our latest IRS related work titled "Unified IRS-aided MIMO transceiver designs via majorization theory".

Algorithm 1 The Modified Subgradient Method for Solving Problem (17)

Initialize: Random dual variables $\lambda_i^{(0)}$'s; iteration index $t = 0$; maximum iteration number $T_{\max}$; sufficiently small threshold $\epsilon > 0$.
1: **repeat**
2: Calculate $\mu_i^{(t)} = \lambda_i^{(t)} P / (\sum_{i=1}^{I} \lambda_i^{(t)} P_i), \quad 1 \le i \le I$.
3: Given $\mathbf{\Omega} = \sum_{i=1}^{I} \mu_i^{(t)} \mathbf{\Omega}_i$, solve problem (18) to obtain $\mathbf{Q}^{(t)}$.
4: Set the step size $a_i^{(t)} = \frac{a}{b \cdot t + c}, \ 1 \le i \le I$, where $\{a, b, c\} > 0$.
5: Update $\lambda_i^{(t+1)} = [\lambda_i^{(t)} + a_i^{(t)} (\mathrm{Tr}(\mathbf{\Omega}_i \mathbf{Q}^{(t)}) - P_i)]^+, \ 1 \le i \le I$.
6: Update $t = t + 1$.
7: **until** $|\mu_i^{(t)} (\mathrm{Tr}(\mathbf{\Omega}_i \mathbf{Q}^{(t)}) - P_i)| \le \epsilon, \ \forall i$, or $t = T_{\max}$.
8: **return** The optimal $\mathbf{Q}^\star = \mathbf{Q}^{(t)}$ to problem (17).

$\mu_i = \lambda_i P / (\sum_{i=1}^{I} \lambda_i P_i), \ 1 \le i \le I$. Then the corresponding KKT optimality conditions are expressed as

$$\nabla^{\mathrm{T}} f(\mathbf{Q}) + \sum_{i=1}^{N_S} \lambda_i \mathbf{\Omega}_i = \nabla^{\mathrm{T}} f(\mathbf{Q}) + \tilde{\mu} \mathbf{\Omega} = \mathbf{\Psi}, \tag{21a}$$

$$\lambda_i (\mathrm{Tr}(\mathbf{\Omega}_i \mathbf{Q}) - P_i) = 0, \ 1 \le i \le I, \ \mathrm{Tr}(\mathbf{\Psi Q}) = 0, \ \mathbf{\Psi} \succeq \mathbf{0}. \tag{21b}$$

It is readily proved that $\tilde{\mu} > 0$ at the optimal solution by contradiction. We then follows from (21b) that $\mathrm{Tr}(\mathbf{\Omega Q}) - P = 0$. The equalities (21a) and $\mathrm{Tr}(\mathbf{\Omega Q}) - P = 0$ are the KKT optimality conditions of the convex problem (18) by considering $\tilde{\mu}$ to be the dual variable associated with the single power constraint $\mathrm{Tr}(\mathbf{\Omega Q}) \le P$. As a result, the equivalence between the two convex problems (17) and (18) holds via dual theory. This completes the whole proof. □

Motivated by Proposition 1, we further develop a novel subgradient method to determine the optimal solution to the original convex problem (17), which is summarized in Algorithm 1. Since the global convergence of the classical subgradient method for solving the convex problem (18) has been well proved in [6], and Proposition 1 establishes the equivalence of the convex problems (18) and (17) by revealing the inherent relationship between their optimal dual variables, it is further concluded that Algorithm 1 is guaranteed to converge to the globally optimal solution of problem (17). The worst-case complexity of Algorithm 1 mainly comes from solving the convex problem (18), which is given by $T_{\max} \mathcal{O}_1$ with $\mathcal{O}_1$ depending on the specific type of problem (18), as elaborated in Section V. In the sequel, we intend to apply Algorithm 1 to both capacity maximization and MSE minimization under multiple weighted power constraints.

1) Capacity Maximization: The capacity maximization of MIMO systems under multiple weighted power constraints is formulated as

$$\textbf{P3}: \ \min_{\mathbf{Q}} \ -\log \left| \mathbf{I} + \mathbf{R}_{\mathrm{n}}^{-1} \mathbf{H Q H}^{\mathrm{H}} \right| \quad \text{s.t.} \ \mathrm{Tr}(\mathbf{\Omega}_i \mathbf{Q}) \le P_i, \ 1 \le i \le I, \ \mathbf{Q} \succeq \mathbf{0}. \tag{22}$$

Based on Proposition 1, the corresponding full KKT conditions are formulated as

$$\begin{cases} \mathbf{H}^{\mathrm{H}} \mathbf{R}_{\mathrm{n}}^{-\frac{1}{2}} \big(\mathbf{I} + \mathbf{R}_{\mathrm{n}}^{-\frac{1}{2}} \mathbf{H Q H}^{\mathrm{H}} \mathbf{R}_{\mathrm{n}}^{-\frac{1}{2}} \big)^{-1} \mathbf{R}_{\mathrm{n}}^{-\frac{1}{2}} \mathbf{H} \\ = \tilde{\mu} \mathbf{\Omega} - \mathbf{\Psi}, \\ \tilde{\mu} > 0, \ \mathrm{Tr}(\mathbf{\Omega Q}) - P = 0, \ \mathbf{\Omega} = \sum_{i=1}^{I} \mu_i \mathbf{\Omega}_i, \\ \mu_i \big(\mathrm{Tr}(\mathbf{\Omega}_i \mathbf{Q}) - P_i \big) = 0, \ \ \mathrm{Tr}(\mathbf{\Omega}_i \mathbf{Q}) \le P_i, \\ \mathrm{Tr}(\mathbf{Q \Psi}) = 0, \ \mathbf{\Psi} \succeq \mathbf{0}, \ \mathbf{Q} \succeq \mathbf{0}, \ 1 \le i \le I. \end{cases} \tag{23}$$

Upon recalling Theorem 1 and setting $\mathbf{\Pi} = \mathbf{R}_{\mathrm{n}}$, we have the following conclusion for **P3**.

*Conclusion 3: The optimal $\mathbf{Q}$ for **P3** has the following water-filling structure*

$$\mathbf{Q} = \mathbf{\Omega}^{-\frac{1}{2}} \big[\mathbf{V}_{\boldsymbol{\mathcal{H}}} \big]_{:,1:N} \Big(\tilde{\mu}^{-1} \mathbf{I} - \big[\mathbf{\Lambda}_{\boldsymbol{\mathcal{H}}} \big]_{1:N,1:N}^{-2} \Big)^{+} \times \big[\mathbf{V}_{\boldsymbol{\mathcal{H}}} \big]_{:,1:N}^{\mathrm{H}} \mathbf{\Omega}^{-\frac{1}{2}}, \tag{24}$$

where the unitary matrix $\mathbf{V}_{\boldsymbol{\mathcal{H}}}$ is defined by the following SVD

$$\mathbf{R}_{\mathrm{n}}^{-\frac{1}{2}} \mathbf{H} \mathbf{\Omega}^{-\frac{1}{2}} = \mathbf{U}_{\boldsymbol{\mathcal{H}}} \mathbf{\Lambda}_{\boldsymbol{\mathcal{H}}} \mathbf{V}_{\boldsymbol{\mathcal{H}}}^{\mathrm{H}} \ \textit{with} \ \mathbf{\Lambda}_{\boldsymbol{\mathcal{H}}} \searrow. \tag{25}$$

In particular, the optimal scalars μ_i's involved in $\mathbf{\Omega}$ can be uniquely found by Algorithm 1.

Computation of μ_i's and $\tilde{\mu}$: Motivated by Proposition 1, the optimal μ_i's enabling the optimal solution of the convex problem **P3** can be uniquely determined by Algorithm 1. Futhermore, according to $\mathrm{Tr}(\mathbf{\Omega Q}) = P$, the optimal closed-form $\tilde{\mu}$ in Conclusion 3 can be calculated as $\tilde{\mu} = \frac{L_{\max}}{P + \sum_{l=1}^{L_{\max}} [\mathbf{\Lambda}_{\boldsymbol{\mathcal{H}}}]_{l,l}^{-2}}$ with $L_{\max}$ being the maximum integer satisfying $\tilde{\mu}^{-1} - [\mathbf{\Lambda}_{\boldsymbol{\mathcal{H}}}]_{L_{\max},L_{\max}}^{-2} \ge 0$.

2) MSE Minimization: The MSE minimization problem under multiple weighted power constraints on the other hand is formulated as

$$\textbf{P4}: \ \min_{\mathbf{Q}} \ \ \mathrm{Tr}\Big(\big(\mathbf{I} + \mathbf{R}_{\mathrm{n}}^{-1} \mathbf{H Q H}^{\mathrm{H}} \big)^{-1} \Big) \quad \text{s.t.} \ \mathrm{Tr}(\mathbf{\Omega}_i \mathbf{Q}) \le P_i, 1 \le i \le I, \ \mathbf{Q} \succeq \mathbf{0}. \tag{26}$$

Similarly, after tedious but straightforward derivations, the corresponding full KKT conditions can be expressed as

$$\begin{cases} \mathbf{H}^{\mathrm{H}} \mathbf{R}_{\mathrm{n}}^{-\frac{1}{2}} \big(\mathbf{I} + \mathbf{R}_{\mathrm{n}}^{-\frac{1}{2}} \mathbf{H Q H}^{\mathrm{H}} \mathbf{R}_{\mathrm{n}}^{-\frac{1}{2}} \big)^{-2} \mathbf{R}_{\mathrm{n}}^{-\frac{1}{2}} \mathbf{H} \\ = \tilde{\mu} \mathbf{\Omega} - \mathbf{\Psi}, \\ \tilde{\mu} > 0, \ \mathrm{Tr}(\mathbf{\Omega Q}) - P = 0, \ \mathbf{\Omega} = \sum_{i=1}^{I} \mu_i \mathbf{\Omega}_i, \\ \mu_i \big(\mathrm{Tr}(\mathbf{\Omega}_i \mathbf{Q}) - P_i \big) = 0, \ \ \mathrm{Tr}(\mathbf{\Omega}_i \mathbf{Q}) \le P_i, \\ \mathrm{Tr}(\mathbf{Q \Psi}) = 0, \ \ \mathbf{\Psi} \succeq \mathbf{0}, \ \ \mathbf{Q} \succeq \mathbf{0}, \ \ 1 \le i \le I, \end{cases} \tag{27}$$

By following the same procedure as that conceived for capacity maximization, the optimal closed-form solution of **P4** can also be derived, as shown in the following conclusion.

*Conclusion 4: The optimal $\mathbf{Q}$ of **P4** has the following closed form*

$$\mathbf{Q} = \mathbf{\Omega}^{-\frac{1}{2}} \big[\mathbf{V}_{\boldsymbol{\mathcal{H}}} \big]_{:,1:N} \Big(\tilde{\mu}^{-1} \big[\mathbf{\Lambda}_{\boldsymbol{\mathcal{H}}} \big]_{1:N,1:N}^{-1} - \big[\mathbf{\Lambda}_{\boldsymbol{\mathcal{H}}} \big]_{1:N,1:N}^{-2} \Big)^{+} \times \big[\mathbf{V}_{\boldsymbol{\mathcal{H}}} \big]_{:,1:N}^{\mathrm{H}} \mathbf{\Omega}^{-\frac{1}{2}}, \tag{28}$$

where $\mathbf{V}_{\boldsymbol{\mathcal{H}}}$ is defined as shown in (25).

Computation of μ_i's **and** $\tilde{\mu}$: Similarly to Section III-B.1, the optimal μ_i's leading to the optimal $\mathbf{Q}$ in (28) can also be computed by the modified subgradient method, and the optimal $\tilde{\mu}$ satisfying $\mathrm{Tr}(\boldsymbol{\Omega Q}) = P$ is calculated as $\tilde{\mu} = \frac{\sum_{l=1}^{L_{\max}}[\boldsymbol{\Lambda_{\mathcal{H}}}]_{l,l}^{-1}}{P+\sum_{l=1}^{L_{\max}}[\boldsymbol{\Lambda_{\mathcal{H}}}]_{l,l}^{-2}}$, where $L_{\max}$ denotes the maximum integer satisfying $\tilde{\mu}^{-1}\big[\boldsymbol{\Lambda_{\mathcal{H}}}\big]_{L_{\max},L_{\max}}^{-1} - \big[\boldsymbol{\Lambda_{\mathcal{H}}}\big]_{L_{\max},L_{\max}}^{-2} \geq 0$.

In a nutshell, it is revealed by Proposition 1 that multiple power constraints of a matrix-monotone convex optimization problem can be integrated into a single one through multiple auxiliary parameters μ_i's whose optimal values can be uniquely determined by the modified subgradient method. Moreover, this artful use of μ_i's can facilitate the derivation of the optimal closed-form solution under multiple power constraints, thereby leading to a lower computational complexity compared to the classical interior point method. Overall, our distinct contributions in terms of solving **P3** and **P4** are summarized as follows.

- Firstly, although **P3** and **P4** have been studied in [16], [27], our proposed modified subgradient method is computationally simpler and overcomes the turning-off effect. Furthermore, the channel matrix is not required to have full-rank column or row.
- Secondly, Proposition 1 reveals that the effect of multiple weighted constraints are in nature equivalent to a weighted total power constraint by comparing the optimal closed-form solutions of **P3** and **P4** to those of **P1** and **P2**.

IV. Transmit Covariance Matrix Optimization Under Imperfect CSI

In practice, it is unrealistic to assume perfect CSI, since CSI must be estimated via the training process. The limited training length and the ubiquitous noise together with the time varying nature of wireless channels result in non-negligible channel estimation error [21]. By taking the channel estimation errors into account, the CSI can be modeled as [21], [24]:

$$\boldsymbol{H} = \widehat{\boldsymbol{H}} + \Delta\boldsymbol{H} \text{ with } \Delta\boldsymbol{H} = \boldsymbol{R}_{\mathrm{R}}^{\frac{1}{2}}\boldsymbol{H}_{\mathrm{W}}\boldsymbol{R}_{\mathrm{T}}^{\frac{1}{2}}, \tag{29}$$

where $\widehat{\boldsymbol{H}}$ denotes the estimated channel and $\Delta\boldsymbol{H}$ is CSI error. $\boldsymbol{R}_{\mathrm{R}}$ and $\boldsymbol{R}_{\mathrm{T}}$ are the receive and transmit spatial correlation matrices, respectively. Furthermore, $\boldsymbol{H}_{\mathrm{W}}$ is a random matrix whose elements are i.i.d. distributed as $\mathcal{CN}(0,\gamma_e)$, where γ_e denotes the error variance. Hereafter, we assume that both the transmitter and receiver only have access to the imperfect CSI. Under this assumption, the performance metrics of average capacity and average MSE are considered, both of which are taken w.r.t. the random matrix $\boldsymbol{H}_{\mathrm{W}}$.[3]

A. Total Power Constraint With Imperfect CSI

1) Average Capacity Maximization: Using the CSI error model (29) and assuming white noise, i.e., $\boldsymbol{R}_{\mathrm{n}} = \sigma_n^2\boldsymbol{I}$, the (approximate) average capacity maximization problem under the total power constraint is formulated as [25], [30]

$$\begin{aligned}\mathbf{P5}:\ \min_{\boldsymbol{Q}}\quad & -\log\left|\boldsymbol{I}+\boldsymbol{K}_{\mathrm{n}}^{-1}\widehat{\boldsymbol{H}}\boldsymbol{Q}\widehat{\boldsymbol{H}}^{\mathrm{H}}\right| \\ \text{s.t.}\quad & \mathrm{Tr}(\boldsymbol{Q})\leq P,\ \boldsymbol{Q}\succeq\boldsymbol{0},\end{aligned} \tag{30}$$

where the equivalent noise covariance matrix $\boldsymbol{K}_{\mathrm{n}}$ is denoted by

$$\boldsymbol{K}_{\mathrm{n}} = \sigma_{\mathrm{n}}^2\boldsymbol{I} + \mathrm{Tr}(\boldsymbol{R}_{\mathrm{T}}\boldsymbol{Q})\boldsymbol{R}_{\mathrm{R}}, \tag{31}$$

and P is the maximum total transmit power. Furthermore, using the first-order derivative of $\log\left|\boldsymbol{I}+\boldsymbol{K}_{\mathrm{n}}^{-1}\widehat{\boldsymbol{H}}\boldsymbol{Q}\widehat{\boldsymbol{H}}^{\mathrm{H}}\right|$ w.r.t $\boldsymbol{Q}$, which is given by

$$\begin{aligned}&\frac{\partial\log\left|\boldsymbol{I}+\boldsymbol{K}_{\mathrm{n}}^{-1}\widehat{\boldsymbol{H}}\boldsymbol{Q}\widehat{\boldsymbol{H}}^{\mathrm{H}}\right|}{\partial\boldsymbol{Q}} = \frac{\partial\log\left|\boldsymbol{K}_{\mathrm{n}}+\widehat{\boldsymbol{H}}\boldsymbol{Q}\widehat{\boldsymbol{H}}^{\mathrm{H}}\right|}{\partial\boldsymbol{Q}} - \frac{\partial\log|\boldsymbol{K}_{\mathrm{n}}|}{\partial\boldsymbol{Q}} \\ &= \left(\widehat{\boldsymbol{H}}^{\mathrm{H}}\left(\boldsymbol{K}_{\mathrm{n}}+\widehat{\boldsymbol{H}}\boldsymbol{Q}\widehat{\boldsymbol{H}}^{\mathrm{H}}\right)^{-1}\widehat{\boldsymbol{H}}\right)^{T} - \mathrm{Tr}\left(\boldsymbol{K}_{\mathrm{n}}^{-1}\boldsymbol{R}_{\mathrm{R}}\right)\boldsymbol{R}_{\mathrm{T}}^{T} \\ &\quad + \mathrm{Tr}\left(\left(\boldsymbol{K}_{\mathrm{n}}+\widehat{\boldsymbol{H}}\boldsymbol{Q}\widehat{\boldsymbol{H}}^{\mathrm{H}}\right)^{-1}\boldsymbol{R}_{\mathrm{R}}\right)\boldsymbol{R}_{\mathrm{T}}^{T},\end{aligned} \tag{32}$$

the full KKT conditions of **P5** can be expressed as follows

$$\begin{cases}\widehat{\boldsymbol{H}}^{\mathrm{H}}\left(\boldsymbol{K}_{\mathrm{n}}+\widehat{\boldsymbol{H}}\boldsymbol{Q}\widehat{\boldsymbol{H}}^{\mathrm{H}}\right)^{-1}\widehat{\boldsymbol{H}} = \mu\boldsymbol{I} \\ \quad + \mathrm{Tr}\left(\left(\boldsymbol{K}_{\mathrm{n}}^{-1} - \left(\boldsymbol{K}_{\mathrm{n}}+\widehat{\boldsymbol{H}}\boldsymbol{Q}\widehat{\boldsymbol{H}}^{\mathrm{H}}\right)^{-1}\right)\boldsymbol{R}_{\mathrm{R}}\right)\boldsymbol{R}_{\mathrm{T}} - \boldsymbol{\Psi}, \\ \mu\geq 0,\ \mu\left(\mathrm{Tr}(\boldsymbol{Q})-P\right)=0,\ \mathrm{Tr}(\boldsymbol{Q\Psi})=0, \\ \boldsymbol{\Psi}\succeq\boldsymbol{0},\ \mathrm{Tr}(\boldsymbol{Q})\leq P,\ \boldsymbol{Q}\succeq\boldsymbol{0}.\end{cases} \tag{33}$$

Using Theorem 1, we have the following conclusion for the optimal solution of **P5**.

*Conclusion 5: The optimal $\boldsymbol{Q}$ for **P5** has the following water-filling structure*:

$$\begin{aligned}\boldsymbol{Q} = \boldsymbol{\Phi}^{-\frac{1}{2}}\big[\boldsymbol{V_{\mathcal{H}}}\big]_{:,1:N}&\left(\mu^{-1}\boldsymbol{I} - \big[\boldsymbol{\Lambda_{\mathcal{H}}}\big]_{1:N,1:N}^{-2}\right)^{+} \\ &\times\big[\boldsymbol{V_{\mathcal{H}}}\big]_{:,1:N}^{\mathrm{H}}\boldsymbol{\Phi}^{-\frac{1}{2}},\end{aligned} \tag{34}$$

where $\boldsymbol{\Phi} = \boldsymbol{I} + \frac{1}{\mu}\mathrm{Tr}\left(\left(\boldsymbol{K}_{\mathrm{n}}^{-1} - \left(\boldsymbol{K}_{\mathrm{n}}+\widehat{\boldsymbol{H}}\boldsymbol{Q}\widehat{\boldsymbol{H}}^{\mathrm{H}}\right)^{-1}\right)\boldsymbol{R}_{\mathrm{R}}\right)\boldsymbol{R}_{\mathrm{T}}$ *and the unitary matrix* $\boldsymbol{V_{\mathcal{H}}}$ *satisfies* $\boldsymbol{K}_{\mathrm{n}}^{-\frac{1}{2}}\widehat{\boldsymbol{H}}\boldsymbol{\Phi}^{-\frac{1}{2}} = \boldsymbol{U_{\mathcal{H}}\Lambda_{\mathcal{H}}V_{\mathcal{H}}^{\mathrm{H}}}$ *with* $\boldsymbol{\Lambda_{\mathcal{H}}}\searrow$.

As for Conclusion 5, since the optimal $\boldsymbol{Q}$ appears in both sides of (34), its closed-form structure is usually hard to obtain. Armed with convex and non-convex optimization techniques, we next aim for deriving the optimal closed-form solution from the equation (34) for the following three cases: a) **P5.1**: both the transmit and receive antennas have spatial correlation, i.e. $\boldsymbol{R}_{\mathrm{R}} \neq \boldsymbol{I}_{N_R}$ and $\boldsymbol{R}_{\mathrm{T}} \neq \boldsymbol{I}_{N_T}$; b) **P5.2**: the transmit antennas have no spatial correlation, i.e. $\boldsymbol{R}_{\mathrm{T}} = \boldsymbol{I}_{N_T}$; c) **P5.3**: the receive antennas have no spatial correlation, i.e. $\boldsymbol{R}_{\mathrm{R}} = \boldsymbol{I}_{N_R}$.

a) P5.1: For $\boldsymbol{R}_{\mathrm{T}} \neq \mathbf{I}_{N_T}$ and $\boldsymbol{R}_{\mathrm{R}} \neq \mathbf{I}_{N_R}$, it is clear that **P5** is non-convex due to the coupled $\boldsymbol{K}_n$ and $\widehat{\boldsymbol{H}}\boldsymbol{Q}\widehat{\boldsymbol{H}}^{\mathrm{H}}$ in the objective function. Thus it is much challenging to optimally solve **P5**. Fortunately, inspired by the concavity of the log function, we can apply the SCA technique based on the first-order Taylor expansion of $\log\det(\boldsymbol{K}_n)$ around a point

[3]The interested readers can refer to [30] for the detailed derivation of the average MSE metric. Moreover, based on the fundamental relation between the mutual information and MSE, an analytical lower bound of the average mutual information (capacity) can be derived as **P5**, which is adopted due to its good mathematical tractability [25].

$\boldsymbol{Q}^{(t)}$ (where t denotes the iteration index) to obtain a convex upper-bound of the capacity, which is expressed as

$$\log\det(\boldsymbol{K}_n) \le \log\det\big(\boldsymbol{K}_n^{(t)}\big) + 1/(\ln 2) \times \mathrm{Tr}\big((\boldsymbol{K}_n^{(t)})^{-1}\boldsymbol{R}_\mathrm{R}\big)\mathrm{Tr}\big(\boldsymbol{R}_\mathrm{T}(\boldsymbol{Q}-\boldsymbol{Q}^{(t)})\big). \quad (35)$$

Based on (35), an alternative convex upper bound optimization of **P5** is formulated as

$$\min_{\boldsymbol{Q}} \frac{1}{\ln 2}\mathrm{Tr}\big((\boldsymbol{K}_n^{(t)})^{-1}\boldsymbol{R}_\mathrm{R}\big)\mathrm{Tr}\big(\boldsymbol{R}_\mathrm{T}\boldsymbol{Q}\big) - \log\big|\boldsymbol{K}_\mathrm{n}+\widehat{\boldsymbol{H}}\boldsymbol{Q}\widehat{\boldsymbol{H}}^\mathrm{H}\big| \quad \text{s.t. } \mathrm{Tr}(\boldsymbol{Q}) \le P,\ \boldsymbol{Q}\succeq \boldsymbol{0}, \quad (36)$$

Following the design philosophy of SCA, we can find a locally optimal solution of **P5** by iteratively optimizing the convex problem (36). Specifically, the first order derivative of the Lagrangian of problem (36) w.r.t. $\boldsymbol{Q}$ is expressed as

$$\widehat{\boldsymbol{H}}^\mathrm{H}\big(\boldsymbol{K}_\mathrm{n}+\widehat{\boldsymbol{H}}\boldsymbol{Q}\widehat{\boldsymbol{H}}^\mathrm{H}\big)^{-1}\widehat{\boldsymbol{H}} = \mu\boldsymbol{I} + \mathrm{Tr}\Big(\big((\boldsymbol{K}_\mathrm{n}^{(t)})^{-1} - (\boldsymbol{K}_\mathrm{n}+\widehat{\boldsymbol{H}}\boldsymbol{Q}\widehat{\boldsymbol{H}}^\mathrm{H})^{-1}\big)\boldsymbol{R}_\mathrm{R}\Big)\boldsymbol{R}_\mathrm{T} - \boldsymbol{\Psi}, \quad (37)$$

which is the first KKT condition in (33) upon replacing $\boldsymbol{K}_\mathrm{n}^{-1}$ by $(\boldsymbol{K}_\mathrm{n}^{(t)})^{-1}$. However, even with this substitution, the optimal closed-form $\boldsymbol{Q}$ is still difficult to obtain due to the diverse forms of $\boldsymbol{Q}$ involved in $\boldsymbol{K}_\mathrm{n}$ and $\widehat{\boldsymbol{H}}\boldsymbol{Q}\widehat{\boldsymbol{H}}^\mathrm{H}$. To address this issue, we consider introducing an auxillary variable y to rewrite problem (36) as

$$\min_{\boldsymbol{Q},y} 1/(\ln 2)\ \mathrm{Tr}\big((\boldsymbol{K}_n^{(t)})^{-1}\boldsymbol{R}_\mathrm{R}\big)y - \log\big|\sigma_\mathrm{n}^2\boldsymbol{I} + y\boldsymbol{R}_\mathrm{R} + \widehat{\boldsymbol{H}}\boldsymbol{Q}\widehat{\boldsymbol{H}}^\mathrm{H}\big| \quad \text{s.t. } \mathrm{Tr}(\boldsymbol{Q}) \le P,\ y = \mathrm{Tr}\big(\boldsymbol{R}_\mathrm{T}\boldsymbol{Q}\big),\ \boldsymbol{Q}\succeq\boldsymbol{0}. \quad (38)$$

Then the full KKT conditions of the convex problem (38) are given by

$$\begin{cases} \widehat{\boldsymbol{H}}^\mathrm{H}\big(\sigma_\mathrm{n}^2\boldsymbol{I}+y\boldsymbol{R}_\mathrm{R}+\widehat{\boldsymbol{H}}\boldsymbol{Q}\widehat{\boldsymbol{H}}^\mathrm{H}\big)^{-1}\widehat{\boldsymbol{H}} = \mu\boldsymbol{I} - \beta\boldsymbol{R}_T - \boldsymbol{\Psi}, \\ f(y) = \mathrm{Tr}\Big(\big((\sigma_\mathrm{n}^2\boldsymbol{I}+y\boldsymbol{R}_\mathrm{R}+\widehat{\boldsymbol{H}}\boldsymbol{Q}\widehat{\boldsymbol{H}}^\mathrm{H})^{-1} - (\boldsymbol{K}_n^{(t)})^{-1}\big)\boldsymbol{R}_R\Big) = \beta, \\ y - \mathrm{Tr}\big(\boldsymbol{R}_\mathrm{T}\boldsymbol{Q}\big) = 0, \quad \mu\boldsymbol{I} - \beta\boldsymbol{R}_T \succ \boldsymbol{0} \\ \mu \ge 0,\ \mu\,(\mathrm{Tr}(\boldsymbol{Q}) - P) = 0,\ \mathrm{Tr}(\boldsymbol{Q}\boldsymbol{\Psi}) = 0, \\ \boldsymbol{\Psi}\succeq\boldsymbol{0},\ \mathrm{Tr}(\boldsymbol{Q}) \le P,\ \boldsymbol{Q}\succeq\boldsymbol{0}. \end{cases} \quad (39)$$

Based on (39), the optimal $\boldsymbol{Q}$ of problem (38) is similarly derived as that in Conclusion 5 by setting $\boldsymbol{\Phi} = \boldsymbol{I} - (\beta/\mu)\boldsymbol{R}_T$. Moreover, we readily find that $f(y)$ is monotonically decreasing w.r.t. y, implying that the optimal y can be uniquely determined for a given β. As such, we propose iteratively optimizing $\boldsymbol{Q}$ and y for finding the minimum value of the Lagrangian of problem (38) for each given pair of $\{\mu, \beta\}$. Due to the convexity of problem (38), this iterative procedure is guaranteed to converge to the global minimum of the Lagrangian.

Computation of μ and β: Based on Proposition 1 and regarding $y - \mathrm{Tr}\big(\boldsymbol{R}_\mathrm{T}\boldsymbol{Q}\big) = 0$ as a special weighted power constraint, the optimal dual variables μ and β can be numerically computed using the subgradient method.

b) P5.2: For the case of $\boldsymbol{R}_\mathrm{T} = \boldsymbol{I}_{N_T}$, it follows from (33) that $\boldsymbol{\Phi} \propto \mathbf{I}_{N_R}$ and $\boldsymbol{K}_\mathrm{n} = \sigma_\mathrm{n}^2\boldsymbol{I} + \mathrm{Tr}(\boldsymbol{Q})\boldsymbol{R}_\mathrm{R} = \sigma_\mathrm{n}^2\boldsymbol{I} + P\,\boldsymbol{R}_\mathrm{R}$. Consequently, Conclusion 5 can be simplified as follows.

*Conclusion 5.2 When $\boldsymbol{R}_\mathrm{T} = \boldsymbol{I}_{N_T}$, the optimal $\boldsymbol{Q}$ for **P5** is given by*

$$\boldsymbol{Q} = \big[\boldsymbol{V}_{\boldsymbol{\mathcal{H}}}\big]_{:,1:N}\Big(\mu^{-1}\boldsymbol{I} - \big[\boldsymbol{\Lambda}_{\boldsymbol{\mathcal{H}}}\big]_{1:N,1:N}^{-2}\Big)^{+}\big[\boldsymbol{V}_{\boldsymbol{\mathcal{H}}}\big]_{:,1:N}^\mathrm{H}, \quad (40)$$

where the unitary matrix $\boldsymbol{V}_{\boldsymbol{\mathcal{H}}}$ satisfies $\boldsymbol{K}_\mathrm{n}^{-\frac{1}{2}}\widehat{\boldsymbol{H}} = \boldsymbol{U}_{\boldsymbol{\mathcal{H}}}\boldsymbol{\Lambda}_{\boldsymbol{\mathcal{H}}}\boldsymbol{V}_{\boldsymbol{\mathcal{H}}}^\mathrm{H}$ with $\boldsymbol{\Lambda}_{\boldsymbol{\mathcal{H}}} \searrow$.

Computation of μ: The dual variable μ can be computed based on the equality $\mathrm{Tr}(\boldsymbol{Q}) = \mathrm{Tr}\big((\mu^{-1}\boldsymbol{I} - \big[\boldsymbol{\Lambda}_{\boldsymbol{\mathcal{H}}}\big]_{1:N,1:N}^{-2})^{+}\big) = P$, and its optimal closed-form solution is readily obtained as $\mu = \frac{L_\mathrm{max}}{P+\sum_{l=1}^{L_\mathrm{max}}[\boldsymbol{\Lambda}_{\boldsymbol{\mathcal{H}}}]_{l,l}^{-2}}$ with L_max being the maximum integer satisfying $\mu^{-1} - [\boldsymbol{\Lambda}_{\boldsymbol{\mathcal{H}}}]_{L_\mathrm{max},L_\mathrm{max}}^{-2} \ge 0$.

c) P5.3: For the case of $\boldsymbol{R}_\mathrm{R} = \boldsymbol{I}$, we firstly rewrite $\boldsymbol{K}_\mathrm{n}$ as $\boldsymbol{K}_\mathrm{n} = k_\mathrm{n}\boldsymbol{I}$ with $k_\mathrm{n} = \sigma_\mathrm{n}^2 + \mathrm{Tr}(\boldsymbol{Q}\boldsymbol{R}_\mathrm{T})$. Using the fact $\mathrm{Tr}(\boldsymbol{Q}) = P$ at the optimal $\boldsymbol{Q}$ yields $k_\mathrm{n} = \mathrm{Tr}\big(\boldsymbol{B}\boldsymbol{Q}\big)/P$ with $\boldsymbol{B} = \sigma_\mathrm{n}^2\boldsymbol{I} + P\boldsymbol{R}_\mathrm{T}$. Furthermore, by defining $\widetilde{\boldsymbol{Q}} = \boldsymbol{B}^{\frac{1}{2}}\boldsymbol{Q}\boldsymbol{B}^{\frac{1}{2}}/k_\mathrm{n}$ with $\mathrm{Tr}(\widetilde{\boldsymbol{Q}}) = P$, the first KKT condition in (33) can be rewritten as

$$\widehat{\boldsymbol{H}}^\mathrm{H}\big(\boldsymbol{I} + \widehat{\boldsymbol{H}}\boldsymbol{B}^{-\frac{1}{2}}\widetilde{\boldsymbol{Q}}\boldsymbol{B}^{-\frac{1}{2}}\widehat{\boldsymbol{H}}^\mathrm{H}\big)^{-1}\widehat{\boldsymbol{H}} = \mu\boldsymbol{I} + \mathrm{Tr}\Big(\widehat{\boldsymbol{H}}^\mathrm{H}\big(\boldsymbol{I}+\widehat{\boldsymbol{H}}\boldsymbol{B}^{-\frac{1}{2}}\widetilde{\boldsymbol{Q}}\boldsymbol{B}^{-\frac{1}{2}}\widehat{\boldsymbol{H}}^\mathrm{H}\big)^{-1}\widehat{\boldsymbol{H}}\boldsymbol{B}^{-\frac{1}{2}}\widetilde{\boldsymbol{Q}}\boldsymbol{B}^{-\frac{1}{2}}\Big) \times \boldsymbol{R}_\mathrm{T} - \boldsymbol{\Psi}. \quad (41)$$

By right multiplying $\widetilde{\boldsymbol{Q}}$ and applying the trace operation to both sides of (41) yields

$$\mathrm{Tr}\Big(\widehat{\boldsymbol{H}}^\mathrm{H}\big(\boldsymbol{I}+\widehat{\boldsymbol{H}}\boldsymbol{B}^{-\frac{1}{2}}\widetilde{\boldsymbol{Q}}\boldsymbol{B}^{-\frac{1}{2}}\widehat{\boldsymbol{H}}^\mathrm{H}\big)^{-1}\widehat{\boldsymbol{H}}\boldsymbol{B}^{-\frac{1}{2}}\widetilde{\boldsymbol{Q}}\boldsymbol{B}^{-\frac{1}{2}}\Big) = \frac{\mu\mathrm{Tr}(\boldsymbol{B}^{-\frac{1}{2}}\widetilde{\boldsymbol{Q}}\boldsymbol{B}^{-\frac{1}{2}})}{1 - \mathrm{Tr}(\boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{R}_T\boldsymbol{B}^{-\frac{1}{2}}\widetilde{\boldsymbol{Q}})} = \frac{\mu\mathrm{Tr}(\boldsymbol{Q})}{k_\mathrm{n} - \mathrm{Tr}(\boldsymbol{R}_\mathrm{T}\boldsymbol{Q})} = \frac{\mu P}{\sigma_\mathrm{n}^2}. \quad (42)$$

Furthermore, we substitute (42) into (41) to obtain $\widehat{\boldsymbol{H}}^\mathrm{H}(\boldsymbol{I} + \widehat{\boldsymbol{H}}\boldsymbol{B}^{-\frac{1}{2}}\widetilde{\boldsymbol{Q}}\boldsymbol{B}^{-\frac{1}{2}}\widehat{\boldsymbol{H}}^\mathrm{H})^{-1}\widehat{\boldsymbol{H}} = \frac{\mu}{\sigma_n^2}\boldsymbol{B} - \boldsymbol{\Psi}$. Based on Theorem 1 and $\boldsymbol{\Phi} = \boldsymbol{I}$, the optimal $\widetilde{\boldsymbol{Q}}$ is obtained as

$$\widetilde{\boldsymbol{Q}} = \big[\boldsymbol{V}_{\boldsymbol{\mathcal{H}}}\big]_{:,1:N}\big(\mu^{-1}\boldsymbol{I} - \big[\boldsymbol{\Lambda}_{\boldsymbol{\mathcal{H}}}\big]_{1:N,1:N}^{-2}\big)^{+}\big[\boldsymbol{V}_{\boldsymbol{\mathcal{H}}}\big]_{:,1:N}^\mathrm{H}, \quad (43)$$

where the unitary matrix $\boldsymbol{V}_{\boldsymbol{\mathcal{H}}}$ is defined by the SVD: $\widehat{\boldsymbol{H}}\boldsymbol{\Phi}^{-\frac{1}{2}} = \boldsymbol{U}_{\boldsymbol{\mathcal{H}}}\boldsymbol{\Lambda}_{\boldsymbol{\mathcal{H}}}\boldsymbol{V}_{\boldsymbol{\mathcal{H}}}^\mathrm{H}$ with $\boldsymbol{\Lambda}_{\boldsymbol{\mathcal{H}}} \searrow$. Based on the water-filling structure of (43), we arrive at the following conclusion.

*Conclusion 5.3 When $\boldsymbol{R}_\mathrm{R} = \boldsymbol{I}$, the optimal $\boldsymbol{Q}$ of **P5** can be simplified to*

$$\boldsymbol{Q} = \frac{P}{\mathrm{Tr}(\boldsymbol{B}^{-\frac{1}{2}}\widetilde{\boldsymbol{Q}}\boldsymbol{B}^{-\frac{1}{2}})}\boldsymbol{B}^{-\frac{1}{2}}\widetilde{\boldsymbol{Q}}\boldsymbol{B}^{-\frac{1}{2}}. \quad (44)$$

Computation of μ: Note that the optimal $\widetilde{\boldsymbol{Q}}$ in (43) has a similar expression to the optimal $\boldsymbol{Q}$ in (40), hence the associated optimal scalar μ satisfying $\mathrm{Tr}(\widetilde{\boldsymbol{Q}}) = P$ can be obtained in the same closed form as that in **P5.2**.

2) Average MSE Minimization: Proceeding in a similar manner to Section III-A, by assuming the number of transmitted data streams to be $d = \mathrm{rank}(\widehat{\boldsymbol{H}})$, the average MSE minimization problem can be formulated in the following form [23], [30].

$$\begin{aligned}\mathbf{P6}:\ \min_{\boldsymbol{Q}}\ & \mathrm{Tr}\Big(\big(\boldsymbol{I}+\boldsymbol{K}_{\mathrm{n}}^{-1}\widehat{\boldsymbol{H}}\boldsymbol{Q}\widehat{\boldsymbol{H}}^{\mathrm{H}}\big)^{-1}\Big)\\ \text{s.t.}\ & \mathrm{Tr}(\boldsymbol{Q})\le P,\ \boldsymbol{Q}\succeq\boldsymbol{0},\end{aligned} \tag{45}$$

where $\boldsymbol{K}_{\mathrm{n}}$ is given in (31). Based on the following matrix derivative equality

$$\begin{aligned}&\frac{\partial\mathrm{Tr}\Big(\big(\boldsymbol{I}_{N_R}+\boldsymbol{K}_{\mathrm{n}}^{-1}\widehat{\boldsymbol{H}}\boldsymbol{Q}\widehat{\boldsymbol{H}}^{\mathrm{H}}\big)^{-1}\Big)}{\partial\boldsymbol{Q}}\\ &=-\Big(\widehat{\boldsymbol{H}}^{\mathrm{H}}\boldsymbol{K}_{\mathrm{n}}^{-\frac{1}{2}}\Big(\boldsymbol{I}+\boldsymbol{K}_{\mathrm{n}}^{-\frac{1}{2}}\widehat{\boldsymbol{H}}\boldsymbol{Q}\widehat{\boldsymbol{H}}^{\mathrm{H}}\boldsymbol{K}_{\mathrm{n}}^{-\frac{1}{2}}\Big)^{-2}\boldsymbol{K}_{\mathrm{n}}^{-\frac{1}{2}}\widehat{\boldsymbol{H}}\Big)^{T}\\ &\quad+\mathrm{Tr}\Big(\boldsymbol{K}_{\mathrm{n}}^{-1}\widehat{\boldsymbol{H}}\boldsymbol{Q}^{\frac{1}{2}}\big(\boldsymbol{I}+\boldsymbol{Q}^{\frac{1}{2}}\widehat{\boldsymbol{H}}^{\mathrm{H}}\boldsymbol{K}_{\mathrm{n}}^{-1}\widehat{\boldsymbol{H}}\boldsymbol{Q}^{\frac{1}{2}}\big)^{-2}\\ &\quad\times\boldsymbol{Q}^{\frac{1}{2}}\widehat{\boldsymbol{H}}^{\mathrm{H}}\boldsymbol{K}_{\mathrm{n}}^{-1}\boldsymbol{R}_{\mathrm{R}}\Big)\boldsymbol{R}_{\mathrm{T}}^{T},\end{aligned} \tag{46}$$

the full KKT conditions of **P6** are then given by

$$\begin{cases}\widehat{\boldsymbol{H}}^{\mathrm{H}}\boldsymbol{K}_{\mathrm{n}}^{-\frac{1}{2}}\big(\boldsymbol{I}+\boldsymbol{K}_{\mathrm{n}}^{-\frac{1}{2}}\widehat{\boldsymbol{H}}\boldsymbol{Q}\widehat{\boldsymbol{H}}^{\mathrm{H}}\boldsymbol{K}_{\mathrm{n}}^{-\frac{1}{2}}\big)^{-2}\boldsymbol{K}_{\mathrm{n}}^{-\frac{1}{2}}\widehat{\boldsymbol{H}}=\mu\boldsymbol{I}-\boldsymbol{\Psi}\\ +\mathrm{Tr}\Big(\boldsymbol{K}_{\mathrm{n}}^{-1}\widehat{\boldsymbol{H}}\boldsymbol{Q}^{\frac{1}{2}}\big(\boldsymbol{I}+\boldsymbol{Q}^{\frac{1}{2}}\widehat{\boldsymbol{H}}^{\mathrm{H}}\boldsymbol{K}_{\mathrm{n}}^{-1}\widehat{\boldsymbol{H}}\boldsymbol{Q}^{\frac{1}{2}}\big)^{-2}\\ \times\boldsymbol{Q}^{\frac{1}{2}}\widehat{\boldsymbol{H}}^{\mathrm{H}}\boldsymbol{K}_{\mathrm{n}}^{-1}\boldsymbol{R}_{\mathrm{R}}\Big)\boldsymbol{R}_{\mathrm{T}}\\ \mu\ge 0,\ \mu(\mathrm{Tr}(\boldsymbol{Q})-P)=0,\ \mathrm{Tr}(\boldsymbol{Q}\boldsymbol{\Psi})=0,\\ \boldsymbol{\Psi}\succeq\boldsymbol{0},\ \mathrm{Tr}(\boldsymbol{Q})\le P,\ \boldsymbol{Q}\succeq\boldsymbol{0}.\end{cases} \tag{47}$$

By recalling Theorem 1, we further have the following conclusion.

*Conclusion 6: The optimal $\boldsymbol{Q}$ of **P6** has the following closed form*

$$\begin{aligned}\boldsymbol{Q}=\boldsymbol{\Phi}^{-\frac{1}{2}}\big[\boldsymbol{V}_{\boldsymbol{\mathcal{H}}}\big]_{:,1:N}\Big(\mu^{-\frac{1}{2}}\big[\boldsymbol{\Lambda}_{\boldsymbol{\mathcal{H}}}\big]_{1:N,1:N}^{-1}-\big[\boldsymbol{\Lambda}_{\boldsymbol{\mathcal{H}}}\big]_{1:N,1:N}^{-2}\Big)^{+}\\ \times\big[\boldsymbol{V}_{\boldsymbol{\mathcal{H}}}\big]_{:,1:N}^{\mathrm{H}}\boldsymbol{\Phi}^{-\frac{1}{2}},\end{aligned} \tag{48}$$

where $\boldsymbol{\Phi}$ is given by $\boldsymbol{\Phi}=\boldsymbol{I}+\frac{1}{\mu}\mathrm{Tr}\Big(\boldsymbol{K}_{\mathrm{n}}^{-1}\widehat{\boldsymbol{H}}\boldsymbol{Q}^{\frac{1}{2}}\big(\boldsymbol{I}+\boldsymbol{Q}^{\frac{1}{2}}\widehat{\boldsymbol{H}}^{\mathrm{H}}\boldsymbol{K}_{\mathrm{n}}^{-1}\widehat{\boldsymbol{H}}\boldsymbol{Q}^{\frac{1}{2}}\big)^{-2}\boldsymbol{Q}^{\frac{1}{2}}\widehat{\boldsymbol{H}}^{\mathrm{H}}\boldsymbol{K}_{\mathrm{n}}^{-1}\boldsymbol{R}_{\mathrm{R}}\Big)\boldsymbol{R}_{\mathrm{T}}$, and the unitary matrix $\boldsymbol{V}_{\boldsymbol{\mathcal{H}}}$ satisfies $\boldsymbol{K}_{\mathrm{n}}^{-\frac{1}{2}}\widehat{\boldsymbol{H}}\boldsymbol{\Phi}^{-\frac{1}{2}}=\boldsymbol{U}_{\boldsymbol{\mathcal{H}}}\boldsymbol{\Lambda}_{\boldsymbol{\mathcal{H}}}\boldsymbol{V}_{\boldsymbol{\mathcal{H}}}^{\mathrm{H}}$ with $\boldsymbol{\Lambda}_{\boldsymbol{\mathcal{H}}}\searrow$.

Like solving **P5**, the analytical optimal $\boldsymbol{Q}$ given by Conclusion 6 is usually not available, since the unknown $\boldsymbol{Q}$ still exists in both sides of (48). As a result, we also focus on the same three cases of $\boldsymbol{R}_{\mathrm{T}}$ and $\boldsymbol{R}_{\mathrm{R}}$ as in Section IV-A.1 to explore the optimal closed-from $\boldsymbol{Q}$ from the equation (48), as elaborated below.

a) P6.1: Like **P5.1**, for $\boldsymbol{R}_{\mathrm{T}}\neq\boldsymbol{I}$ and $\boldsymbol{R}_{\mathrm{R}}\neq\boldsymbol{I}$, we apply an auxiliary variable $\boldsymbol{G}\in\mathbb{C}^{N\times N}$ to solve the non-convex problem **P6**. Motivated by the fact that the average MSE metric in (45) is achieved by the Wiener filter, we can rewrite the average MSE objective as [24]

$$\begin{aligned}&\mathrm{Tr}\big(\big(\boldsymbol{I}+\boldsymbol{K}_{\mathrm{n}}^{-1}\widehat{\boldsymbol{H}}\boldsymbol{Q}\widehat{\boldsymbol{H}}^{\mathrm{H}}\big)^{-1}\big)\\ &=\min_{\boldsymbol{G}}\ \mathrm{Tr}\big(\boldsymbol{I}-\boldsymbol{Q}^{\frac{1}{2}}\widehat{\boldsymbol{H}}^{\mathrm{H}}\boldsymbol{G}^{\mathrm{H}}-\boldsymbol{G}\widehat{\boldsymbol{H}}(\boldsymbol{Q}^{\frac{1}{2}})^{\mathrm{H}}+\boldsymbol{G}(\boldsymbol{K}_{n}\\ &\qquad+\widehat{\boldsymbol{H}}\boldsymbol{Q}\widehat{\boldsymbol{H}}^{\mathrm{H}})\boldsymbol{G}^{\mathrm{H}}\big),\end{aligned} \tag{49}$$

where $\boldsymbol{Q}=(\boldsymbol{Q}^{\frac{1}{2}})^{\mathrm{H}}\boldsymbol{Q}^{\frac{1}{2}}$. Note that $\boldsymbol{Q}^{\frac{1}{2}}$ is not required to be positive semidefinite. The optimal $\boldsymbol{G}$ to the inner minimization problem (49) is readily derived as

$$\boldsymbol{G}=\boldsymbol{Q}^{\frac{1}{2}}\widehat{\boldsymbol{H}}^{\mathrm{H}}(\boldsymbol{K}_{n}+\widehat{\boldsymbol{H}}\boldsymbol{Q}\widehat{\boldsymbol{H}}^{\mathrm{H}})^{-1}, \tag{50}$$

which is exactly the Wiener filter. As such, the equality in (49) holds. Further, for addressing the nonconvexity of **P6**, we propose an iterative algorithm that alternates between optimizing $\boldsymbol{G}$ and $\boldsymbol{Q}$. Firstly, for any given $\boldsymbol{Q}$, the optimal $\boldsymbol{G}$ to the inner minimization problem (49) is given by (50). Secondly, for a fixed $\boldsymbol{G}$, we apply the simple matrix manipulation to design $\boldsymbol{Q}=\boldsymbol{Q}^{\frac{1}{2}}(\boldsymbol{Q}^{\frac{1}{2}})^{H}$ such that

$$\begin{aligned}\min_{\boldsymbol{Q}}\ & \mathrm{Tr}\big(\boldsymbol{Q}^{\frac{1}{2}}\boldsymbol{D}(\boldsymbol{Q}^{\frac{1}{2}})^{\mathrm{H}}-\boldsymbol{Q}^{\frac{1}{2}}\widehat{\boldsymbol{H}}^{\mathrm{H}}\boldsymbol{G}^{\mathrm{H}}-\boldsymbol{G}\widehat{\boldsymbol{H}}(\boldsymbol{Q}^{\frac{1}{2}})^{\mathrm{H}}\big)+C,\\ \text{s.t.}\ & \|\boldsymbol{Q}^{\frac{1}{2}}\|_F^2\le P,\end{aligned} \tag{51}$$

where $\boldsymbol{D}=\mathrm{Tr}(\boldsymbol{G}\boldsymbol{R}_R\boldsymbol{G}^{\mathrm{H}})\boldsymbol{R}_T+\widehat{\boldsymbol{H}}^{\mathrm{H}}\boldsymbol{G}^{\mathrm{H}}\boldsymbol{G}\widehat{\boldsymbol{H}}$ and $C=\mathrm{Tr}(\boldsymbol{I}+\sigma_n^2\boldsymbol{G}\boldsymbol{G}^{\mathrm{H}})$. Since problem (51) is convex w.r.t. $\boldsymbol{Q}^{\frac{1}{2}}$, its full KKT conditions are given by

$$\begin{cases}\boldsymbol{Q}^{\frac{1}{2}}(\boldsymbol{D}+\mu\boldsymbol{I})=\boldsymbol{G}\widehat{\boldsymbol{H}}\\ \mu\ge 0,\ \mu(\|\boldsymbol{Q}^{\frac{1}{2}}\|_F^2-P)=0.\end{cases} \tag{52}$$

where the parameter μ can be uniquely found using the bisection method [22]. In particular, for $\mu=0$, the optimal $\boldsymbol{Q}$ is directly obtained as $\boldsymbol{Q}^{\frac{1}{2}}=\boldsymbol{G}\widehat{\boldsymbol{H}}\boldsymbol{D}^{\dagger}$. It is worth noting that the KKT conditions in (52) are quite different from those in (47) derived for the positive semidefinite matrix $\boldsymbol{Q}$ and have much simplified mathematical forms. In fact, it can be regarded as a special case of (47) by setting $\boldsymbol{K}_n=\boldsymbol{G}$, $\boldsymbol{Q}=\boldsymbol{0}$ and $\boldsymbol{\Psi}=\mu\boldsymbol{I}-(\boldsymbol{D}+\mu\boldsymbol{I})\boldsymbol{Q}(\boldsymbol{D}+\mu\boldsymbol{I})$ in the involved first KKT condition, respectively.

b) P6.2: For $\boldsymbol{R}_{\mathrm{T}}=\boldsymbol{I}$, by following similar derivations to that of **P5.2** and observing the KKT conditions in (47), we arrive at the following conclusion.

*Conclusion 6.2 When $\boldsymbol{R}_{\mathrm{T}}=\boldsymbol{I}$, the optimal $\boldsymbol{Q}$ of **P6** is given by*

$$\begin{aligned}\boldsymbol{Q}=\big[\boldsymbol{V}_{\boldsymbol{\mathcal{H}}}\big]_{:,1:N}\Big(\mu^{-1}\big[\boldsymbol{\Lambda}_{\boldsymbol{\mathcal{H}}}\big]_{1:N,1:N}^{-1}\\ -\big[\boldsymbol{\Lambda}_{\boldsymbol{\mathcal{H}}}\big]_{1:N,1:N}^{-2}\Big)^{+}\big[\boldsymbol{V}_{\boldsymbol{\mathcal{H}}}\big]_{:,1:N}^{\mathrm{H}},\end{aligned} \tag{53}$$

Computation of μ: Similar to Conclusion 4, the optimal scalar μ satisfying $\mathrm{Tr}(\boldsymbol{Q})=P$ can be calculated as $\mu=\frac{\sum_{l=1}^{L_{\max}}[\boldsymbol{\Lambda}_{\boldsymbol{\mathcal{H}}}]_{l,l}^{-1}}{P+\sum_{l=1}^{L_{\max}}[\boldsymbol{\Lambda}_{\boldsymbol{\mathcal{H}}}]_{l,l}^{-2}}$, where $L_{\max}$ denotes the maximum integer satisfying $\mu^{-1}\big[\boldsymbol{\Lambda}_{\boldsymbol{\mathcal{H}}}\big]_{L_{\max},L_{\max}}^{-1}-\big[\boldsymbol{\Lambda}_{\boldsymbol{\mathcal{H}}}\big]_{L_{\max},L_{\max}}^{-2}\ge 0$.

c) P6.3: For $\boldsymbol{R}_{\mathrm{R}}=\boldsymbol{I}$, we recall **P5.3** and rewrite the first KKT condition in (47) as

$$\begin{aligned}\boldsymbol{B}^{-\frac{1}{2}}\widehat{\boldsymbol{H}}^{\mathrm{H}}\Big(\boldsymbol{I}+\widehat{\boldsymbol{H}}\boldsymbol{B}^{-\frac{1}{2}}\widetilde{\boldsymbol{Q}}\boldsymbol{B}^{-\frac{1}{2}}\widehat{\boldsymbol{H}}^{\mathrm{H}}\Big)^{-2}\widehat{\boldsymbol{H}}\boldsymbol{B}^{-\frac{1}{2}}\\ =(\mu k_{\mathrm{n}})/\sigma_{\mathrm{n}}^2\boldsymbol{I}-k_{\mathrm{n}}\boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{\Psi}\boldsymbol{B}^{-\frac{1}{2}}.\end{aligned} \tag{54}$$

According to Theorem 1, we have the following optimal $\widetilde{\boldsymbol{Q}}$

$$\begin{aligned}\widetilde{\boldsymbol{Q}}=\big[\boldsymbol{V}_{\boldsymbol{\mathcal{H}}}\big]_{:,1:N}\Big(\mu^{-1}\big[\boldsymbol{\Lambda}_{\boldsymbol{\mathcal{H}}}\big]_{1:N,1:N}^{-1}-\big[\boldsymbol{\Lambda}_{\boldsymbol{\mathcal{H}}}\big]_{1:N,1:N}^{-2}\Big)^{+}\\ \times\big[\boldsymbol{V}_{\boldsymbol{\mathcal{H}}}\big]_{:,1:N}^{\mathrm{H}},\end{aligned} \tag{55}$$

where the unitary matrix $\boldsymbol{V}_{\boldsymbol{\mathcal{H}}}$ is similarly specified by the SVD: $\widehat{\boldsymbol{H}}\boldsymbol{\Phi}^{-\frac{1}{2}} = \boldsymbol{U}_{\boldsymbol{\mathcal{H}}}\boldsymbol{\Lambda}_{\boldsymbol{\mathcal{H}}}\boldsymbol{V}_{\boldsymbol{\mathcal{H}}}^{\mathrm{H}}$ with $\boldsymbol{\Lambda}_{\boldsymbol{\mathcal{H}}} \searrow$. Based on the optimal $\widetilde{\boldsymbol{Q}}$, the optimal $\boldsymbol{Q}$ of **P6** has the same closed form as that in Conclusion 5.3.

Computation of μ: Since the optimal $\widetilde{\boldsymbol{Q}}$ in (55) has a similar expression to that in (53), the corresponding optimal scalar μ satisfying $\mathrm{Tr}(\widetilde{\boldsymbol{Q}}) = P$ can be derived similarly to that in **P6.2**.

In a nutshell, the optimal $\boldsymbol{Q}$'s in **P5.2**, **P5.3**, **P6.2** and **P6.3** with either $\boldsymbol{R}_T = \boldsymbol{I}$ or $\boldsymbol{R}_R = \boldsymbol{I}$ are all directly obtained in closed forms based on the original objective function, while the optimal solutions to **P5.1** and **P6.1** focusing on the general $\boldsymbol{R}_R \neq \boldsymbol{I}$ and $\boldsymbol{R}_T \neq \boldsymbol{I}$ are derived through applying the iterative optimization procedure to the modified objective function.

B. Multiple Weighted Power Constraints With Imperfect CSI

Recently, there have been a few studies on MIMO transceiver designs considering different power allocation schemes under imperfect CSI. For example, the authors of [31] proposed various error-tolerant diagonal precoding schemes with different power allocation strategies to overcome the effect of imperfect MIMO channel feedback. However, to the best of our knowledge, for more general imperfect MIMO system setup, the globally optimal solutions to the robust MIMO transceiver designs under general power constraints have not been investigated. In the sequel, we mainly investigate the robust transmit covariance matrix optimization from the perspectives of average capacity maximization and average MSE minimization.

1) Average Capacity Maximization: The average capacity maximization problem of MIMO systems under multiple weighted power constraints is formulated as

$$\begin{aligned}\textbf{P7}:\ \min_{\boldsymbol{Q}}\quad & -\log\left|\boldsymbol{I}+\boldsymbol{K}_{\mathrm{n}}^{-1}\widehat{\boldsymbol{H}}\boldsymbol{Q}\widehat{\boldsymbol{H}}^{\mathrm{H}}\right| \\ \mathrm{s.t.}\ & \mathrm{Tr}(\boldsymbol{\Omega}_i\boldsymbol{Q}) \le P_i,\quad 1\le i\le I,\ \boldsymbol{Q}\succeq\boldsymbol{0}.\end{aligned} \tag{56}$$

Similar to **P5.1**, an alternative convex upper-bound optimization of **P7** is given by

$$\begin{aligned}\min_{\boldsymbol{Q},y}\quad & 1/(\ln 2)\,\mathrm{Tr}\big((\boldsymbol{K}_n^{(t)})^{-1}\boldsymbol{R}_{\mathrm{R}}\big)y \\ & -\log\left|\sigma_{\mathrm{n}}^2\boldsymbol{I}+y\boldsymbol{R}_{\mathrm{R}}+\widehat{\boldsymbol{H}}\boldsymbol{Q}\widehat{\boldsymbol{H}}^{\mathrm{H}}\right|, \\ \mathrm{s.t.}\quad & \mathrm{Tr}(\boldsymbol{\Omega}_i\boldsymbol{Q}) \le P_i,\ y=\mathrm{Tr}\big(\boldsymbol{R}_{\mathrm{T}}\boldsymbol{Q}\big),\quad \boldsymbol{Q}\succeq\boldsymbol{0},\ 1\le i\le I.\end{aligned} \tag{57}$$

Accordingly, KKT conditions of problem (57) are given by

$$\begin{cases}\widehat{\boldsymbol{H}}^{\mathrm{H}}\big(\sigma_{\mathrm{n}}^2\boldsymbol{I}+y\boldsymbol{R}_{\mathrm{R}}+\widehat{\boldsymbol{H}}\boldsymbol{Q}\widehat{\boldsymbol{H}}^{\mathrm{H}}\big)^{-1}\widehat{\boldsymbol{H}}=\mu\boldsymbol{\Omega}-\beta\boldsymbol{R}_T-\boldsymbol{\Psi}, \\ f(y)=\mathrm{Tr}\Big(\Big((\sigma_{\mathrm{n}}^2\boldsymbol{I}+y\boldsymbol{R}_{\mathrm{R}}+\widehat{\boldsymbol{H}}\boldsymbol{Q}\widehat{\boldsymbol{H}}^{\mathrm{H}})^{-1}-(\boldsymbol{K}_n^{(t)})^{-1}\Big)\boldsymbol{R}_R\Big) \\ \quad =\beta, \\ y-\mathrm{Tr}\big(\boldsymbol{R}_{\mathrm{T}}\boldsymbol{Q}\big)=0,\quad \mu\boldsymbol{I}-\beta\boldsymbol{R}_T\succ\boldsymbol{0} \\ \tilde{\mu}>0,\ \mathrm{Tr}(\boldsymbol{\Omega}\boldsymbol{Q})-P=0,\ \boldsymbol{\Omega}=\sum_{i=1}^{I}\mu_i\boldsymbol{\Omega}_i, \\ \mu_i\big(\mathrm{Tr}(\boldsymbol{\Omega}_i\boldsymbol{Q})-P_i\big)=0,\quad \mathrm{Tr}(\boldsymbol{\Omega}_i\boldsymbol{Q})\le P_i, \\ \mathrm{Tr}(\boldsymbol{Q}\boldsymbol{\Psi})=0,\ \boldsymbol{\Psi}\succeq\boldsymbol{0},\ \boldsymbol{Q}\succeq\boldsymbol{0},\ 1\le i\le I.\end{cases} \tag{58}$$

It is readily shown that the above KKT conditions have similar formulation to (39). Therefore, the corresponding optimal $\boldsymbol{Q}$ of problem (57) has an almost identical closed-form solution to that of problem (38) by slightly modifying $\boldsymbol{\Phi}$ as $\boldsymbol{\Phi}=\boldsymbol{\Omega}-(\beta/\mu)\boldsymbol{R}_T$ with $\boldsymbol{\Omega}=\sum_{i=1}^{I}\mu_i\boldsymbol{\Omega}_i$, where the optimal μ_i's are uniquely determined by Algorithm 1. In particular, for the special case of $\boldsymbol{R}_{\mathrm{R}}=\boldsymbol{I}$, we replace $\boldsymbol{Q}$ with $\widetilde{\boldsymbol{Q}}=\big(\sigma_n^2+\mathrm{Tr}(\boldsymbol{R}_{\mathrm{T}}\boldsymbol{Q})\big)^{-1}\boldsymbol{Q}$. Then problem (56) is rewritten as

$$\begin{aligned}\min_{\boldsymbol{Q}}\quad & -\log\left|\boldsymbol{I}+\widehat{\boldsymbol{H}}\widetilde{\boldsymbol{Q}}\widehat{\boldsymbol{H}}^{\mathrm{H}}\right| \\ \mathrm{s.t.}\quad & \mathrm{Tr}(\widetilde{\boldsymbol{\Omega}}_i\widetilde{\boldsymbol{Q}})\le P_i,\ 1\le i\le I,\ \widetilde{\boldsymbol{Q}}\succeq\boldsymbol{0},\end{aligned} \tag{59}$$

where $\widetilde{\boldsymbol{\Omega}}_i = \sigma_n^2\boldsymbol{\Omega}_i + P_i\boldsymbol{R}_{\mathrm{T}},\ \forall i$. Note that problem (59) has a similar structure to **P3**, and thereby can be globally solved following the same methodology as that of solving **P3**. Nevertheless, for another special case of $\boldsymbol{R}_T=\boldsymbol{I}$, the optimal closed-from $\boldsymbol{Q}$ to **P7** is hard to obtain directly, since $\mathrm{Tr}(\boldsymbol{Q})$ involved in $\boldsymbol{K}_n$ cannot be determined as a constant like in **P5.2** focusing on the single total power constraint.

2) Average MSE Minimization: The average MSE minimization problem of MIMO systems subject to multiple weighted power constraints can be formulated as

$$\begin{aligned}\textbf{P8}:\ \min_{\boldsymbol{Q}}\quad & \mathrm{Tr}\big((\boldsymbol{I}+\boldsymbol{K}_{\mathrm{n}}^{-1}\widehat{\boldsymbol{H}}\boldsymbol{Q}\widehat{\boldsymbol{H}}^{\mathrm{H}})^{-1}\big) \\ \mathrm{s.t.}\ & \mathrm{Tr}(\boldsymbol{\Omega}_i\boldsymbol{Q})\le P_i,\quad 1\le i\le I,\ \boldsymbol{Q}\succeq\boldsymbol{0}.\end{aligned} \tag{60}$$

For tackling the non-convexity of **P8**, the proposed alternating optimization algorithm in **P6.1** is still applicable. To be specific, by recalling the equivalent MSE objective in (49), the optimal $\boldsymbol{G}$ to **P8** with any given $\boldsymbol{Q}$ has the same closed form as (51), while the optimal $\boldsymbol{Q}$ for any given $\boldsymbol{G}$ should be designed such that

$$\begin{aligned}\min_{\boldsymbol{Q}^{\frac{1}{2}}}\quad & \mathrm{Tr}\Big(\boldsymbol{Q}^{\frac{1}{2}}\boldsymbol{D}(\boldsymbol{Q}^{\frac{1}{2}})^{\mathrm{H}}-\boldsymbol{Q}^{\frac{1}{2}}\widehat{\boldsymbol{H}}^{\mathrm{H}}\boldsymbol{G}^{\mathrm{H}}-\boldsymbol{G}\widehat{\boldsymbol{H}}(\boldsymbol{Q}^{\frac{1}{2}})^{\mathrm{H}}\Big) \\ \mathrm{s.t.}\quad & \|\boldsymbol{Q}^{\frac{1}{2}}\|_F^2\le P.\end{aligned} \tag{61}$$

The full KKT conditions of problem (61) are then given by

$$\begin{cases}\boldsymbol{Q}^{\frac{1}{2}}(\boldsymbol{D}+\tilde{\mu}\boldsymbol{\Omega})=\boldsymbol{G}\widehat{\boldsymbol{H}}, \\ \tilde{\mu}>0,\ \mathrm{Tr}(\boldsymbol{Q}^{\frac{1}{2}}\boldsymbol{\Omega}(\boldsymbol{Q}^{\frac{1}{2}})^{\mathrm{H}})-P=0,\ \boldsymbol{\Omega}=\sum_{i=1}^{I}\mu_i\boldsymbol{\Omega}_i, \\ \mu_i\ge 0,\ \mu_i\Big(\mathrm{Tr}\big(\boldsymbol{Q}^{\frac{1}{2}}\boldsymbol{\Omega}_i(\boldsymbol{Q}^{\frac{1}{2}})^{\mathrm{H}}\big)-P_i\Big)=0, \\ \mathrm{Tr}\big(\boldsymbol{Q}^{\frac{1}{2}}\boldsymbol{\Omega}_i(\boldsymbol{Q}^{\frac{1}{2}})^{\mathrm{H}}\big)\le P_i,\end{cases} \tag{62}$$

Observe from (62) that for any given $\boldsymbol{\Omega}$, the optimal $\tilde{\mu}$ leading to $\mathrm{Tr}(\boldsymbol{Q}^{\frac{1}{2}}\boldsymbol{\Omega}(\boldsymbol{Q}^{\frac{1}{2}})^{\mathrm{H}}) = P$ can also be found via bisection search, while the optimal μ_i's are computed from Algorithm 1. Furthermore, for the special case of $\boldsymbol{R}_{\mathrm{R}}=\boldsymbol{I}$, **P8** can be rewritten as

$$\begin{aligned}\min_{\widetilde{\boldsymbol{Q}}}\quad & \mathrm{Tr}\big((\boldsymbol{I}+\widehat{\boldsymbol{H}}\widetilde{\boldsymbol{Q}}\widehat{\boldsymbol{H}}^{\mathrm{H}})^{-1}\big) \\ \mathrm{s.t.}\quad & \mathrm{Tr}(\widetilde{\boldsymbol{\Omega}}_i\widetilde{\boldsymbol{Q}})\le P_i,\ \widetilde{\boldsymbol{Q}}\succeq\boldsymbol{0},\quad 1\le i\le I.\end{aligned} \tag{63}$$

Since problem (63) has a similar structure to problem (26), the optimal $\widetilde{\boldsymbol{Q}}$ can be calculated as that in Conclusion 4. However, like the average capacity maximization, for $\boldsymbol{R}_{\mathrm{T}}=\boldsymbol{I}$, the optimal closed-form $\boldsymbol{Q}$ to **P8** is hard to obtain due to the unknown $\mathrm{Tr}(\boldsymbol{Q})$ under multiple power constraints.

Fortunately, the proposed alternating optimization algorithm is still applicable.

V. Summary and Discussion

A. Convergence and Complexity Analysis

In summary, when considering spatial correlations at both transmitter and receiver antenna arrays, i.e. $\mathbf{R}_T \neq \mathbf{I}$ and $\mathbf{R}_R \neq \mathbf{I}$, we apply the classical SCA technique to tackle the non-convex average capacity maximization **P5.1** and **P7**. As for the non-convex average MSE minimization, we firstly introduce an auxiliary variable $\mathbf{G}$, and then apply the two-block alternating optimization algorithm to solve **P6.1** and **P8**. Actually, both the proposed SCA-based algorithm and two-block alternating optimization algorithm belong to the block successive upper-bound minimization (BSUM) framework [32]. For each above-mentioned problem, the unique optimal solutions to its involved convex subproblems are all available, and the corresponding objective value is upper-bounded due to the closed and bounded feasible region, i.e. $\mathrm{Tr}(\mathbf{\Omega}_i\mathbf{Q}) \leq P_i,\ 1 \leq i \leq I$. Taking the average capacity maximization in **P5.1** as an example, we must have $f(\boldsymbol{Q}^{(t)}) = g(\boldsymbol{Q}^{(t)}|\boldsymbol{Q}^{(t)}) \overset{(a)}{\geq} g(\boldsymbol{Q}^{(t+1)}|\boldsymbol{Q}^{(t)}) \geq f(\boldsymbol{Q}^{(t+1)})$ during the iteration process, where $f(\boldsymbol{Q})$ denotes the optimal objective value of **P5**, and $g(\boldsymbol{Q}|\boldsymbol{Q}^{(t)})$ dentoes a locally tight upper-bound of $f(\boldsymbol{Q})$ based on (35). The inequality (a) holds due to the minimization in **P5**. The similar property can also observed for **P7**, **P6.1** and **P8**. Motivated by the above facts, the proposed SCA-based algorithm and two-block alternating optimization algorithm are both guaranteed to monotonically converge to a locally optimal solution of the corresponding optimization problem [32, Th. 2]. Furthermore, for the special case of $\mathbf{R}_T = \mathbf{I}$ or $\mathbf{R}_R = \mathbf{I}$, the corresponding optimization problems **P7** and **P8** under multiple power constraints are both solved by Algorithm 1, whose global convergence has been illustrated in Section III-B.

In addition, within each iteration of the proposed SCA-based algorithm for average capacity maximization, the computational complexity mainly comes from the update of dual variables using the classical subgradient method (for **P5.1**) or Algorithm 1 (for **P7**). As a result, the worst-case complexities of solving **P5.1** and **P7** in each SCA iteration are given by $I_{\mathrm{sub}}\mathcal{O}_1$ and $T_{\max}\mathcal{O}_1$ with $\mathcal{O}_1 = \mathcal{O}\Big(I_{\mathrm{inn}}\big(N_T^3 + N_R^3 + \min(N_R^2\ N_T, N_R N_T^2) + \log(P\lambda_{\max}(\boldsymbol{R}_T)/\epsilon)\big)\Big)$, respectively, where I_{sub} and I_{inn} denote the numbers of iterations for finding the optimal dual variables $\{\mu, \beta\}$ and for alternately optimizing $\boldsymbol{Q}$ and y until convergence, respectively. The involved complexity $\mathcal{O}\Big(I_{\mathrm{inn}}\big(N_T^3 + N_R^3 + \min(N_R^2\ N_T, N_R N_T^2)\big)\Big)$ is mainly due to the matrix inversion and SVD of size $N_T \times N_T$, $N_R \times N_R$ and $N_T \times N_R$, while $\mathcal{O}\big(I_{\mathrm{inn}}\log(P\lambda_{\max}(\boldsymbol{R}_T)/\epsilon)\big)$ denotes the worst-case complexity of the bisection search for finding the optimal y [5].

Furthermore, the complexities of solving **P6.1** and **P8** are both dominated by the second block optimization of $\boldsymbol{Q}$, where the dual variables are updated by either the bisection search or Algorithm 1. Specifically, **P6.1** and **P8** have the worst-case complexities of $I_A(\mathcal{O}_2 + \mathcal{O}_3)$ and $I_A(\mathcal{O}_2 + T_{\max}\mathcal{O}_3)$, respectively, where I_A denotes the number of iterations required for alternately optimizing $\boldsymbol{G}$ and $\boldsymbol{Q}$ until convergence. Similarly, the involved complexity $\mathcal{O}_2 = \mathcal{O}(N_R^3 + N_T^2 N_R + N_R^2 N_T)$ comes from the matrix inversion and multiplication of size $N_R \times N_R$, $N_R \times N_T$, while $\mathcal{O}_3 = \mathcal{O}\Big(\big(N_T^3 + \log(\frac{\|\boldsymbol{G}\widehat{\boldsymbol{H}}\|}{\sqrt{P_\epsilon}})\big)\Big)$ is due to the matrix inversion of size $N_T \times N_T$ and the bisection search on the optimal μ. Finally, for two convex problems **P7** and **P8** with $\boldsymbol{R}_R = \boldsymbol{I}$, whose semi-closed-form solutions are available based on Proposition 1, their worst-case complexities are mainly from Algorithm 1, and both are calculated as $\mathcal{O}_4 = \mathcal{O}\big(T_{\max}\big(N_T^3 + \min(N_T^2 N_R, N_R^2 N_T)\big)\big)$.

B. Extension to Multiuser Scenarios

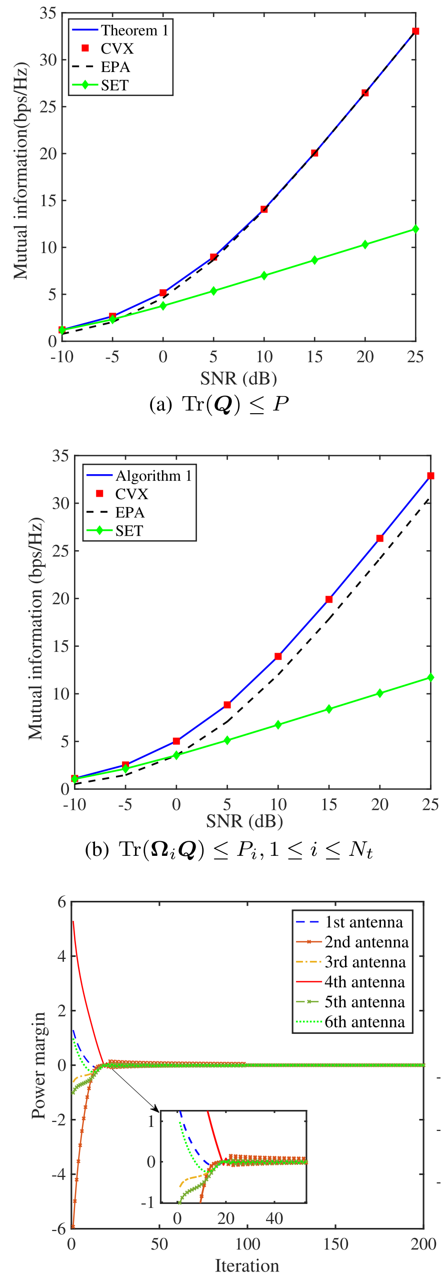
Our proposed Theorem 1 can also be extended to uplink multiuser MIMO systems with successive interference cancellation (SIC) at the transmitter. Specifically, motivated by the iterative water-filling strategy, where the single-user transmit covariance matrix optimization is studied by regarding all other users' signals as noise, the KKT conditions of both capacity maximization and MSE minimization subject to general power constraints under perfect CSI have the similar structure to (1), and thus Theorem 1 is applicable. In contrast, both average capacity maximization and MSE minimization for the imperfect CSI case become more complex over the MIMO scenarios, since channel error exists in each user's channel and thus render Theorem 1 inapplicable. Furthermore, in downlink multi-user MIMO systems, the existence of multi-user interference usually leads to highly coupled optimization variables and complicated non-convex objectives. Similarly, in MIMO multi-hop relaying systems, the relay forwarding matrices are also coupled in the optimization objective and multiple power constraints. For these two scenarios, it is quite difficult to summarize a general structure of KKT conditions of the general utility optimization problem as that in Theorem 1, not to mention the derivation of the unified analytical solution. A potential alternative scheme is to firstly apply the majorization-minimization technique to find an easier-to-handle approximation of the original intractable problem, and then utilize Theorem 1 to solve this approximate problem. Since this topic is beyond the scope of our work, we consider it as our future research direction.

VI. Numerical Results

In this section, the global optimality of the KKT conditions based solution drawn from Theorem 1 and Algorithm 1 is demonstrated by numerical simulations. Unless otherwise stated, we consider a $N_t \times N_r = 6 \times 4$ MIMO system with $d = 4$ data streams for evaluating the performance of the proposed perfect transceiver designs of Section III concerning a pair of classical performance metrics, namely achievable information rate and MSE, as well as their robust counterparts in Section IV. Moreover, for each type of transceiver design, both the total power constraint and individual per-antenna power constraints are considered. To be specific, under the assumption of the unit noise variance and the CSCG distributed channel, i.e. $\boldsymbol{H} \sim \mathcal{CN}(\mathbf{0}, \boldsymbol{I}_{N_t N_r})$, the transmit signal-to-noise ratio (SNR) is defined as $\mathrm{SNR} = 10\log(P)$dB. For simplicity, the maximum per-antenna power is assumed to be $P_1 = \cdots = P_{N_t} = P/N_t$. Correspondingly, the power weighting matrices are set as $\boldsymbol{\Omega}_i = \mathrm{diag}[\mathbf{0}_{1\times(i-1)}, 1, \mathbf{0}_{1\times(N_t-i)}]$, $\forall i = 1,\cdots,N_t$.

Furthermore, the element-wise variance of the random CSI error component $\boldsymbol{H}_W$ is chosen as $\gamma_e = 0.02$. Accordingly, the estimated channel $\widehat{\boldsymbol{H}}$ is generated according to $\mathcal{CN}(0,(1-\gamma_e)\boldsymbol{R}_T^T\otimes\boldsymbol{R}_R)$, implying that the random component of channel $\boldsymbol{H}$ follows the standard CSCG distribution $\mathcal{CN}(\mathbf{0},\boldsymbol{I}_{N_rN_t})$. Armed with the widely adopted exponential model [33], the transmit and receive channel correlation matrices (i.e. $\boldsymbol{R}_T$ and $\boldsymbol{R}_R$) of the robust transceiver designs in Section IV are modeled as $[\boldsymbol{R}_T]_{i,j} = p_t^{|i-j|}$, $[\boldsymbol{R}_R]_{m,n} = p_r^{|m-n|}$, $\forall i,j = 1,\cdots,N_t$; $\forall m,n = 1,\cdots,N_r$, where $p_t, p_r \in [0,1]$ whose values are specified later.[4] For the proposed alternating optimization (Prop-Alt) algorithms in **P5.1** and **P6.1** of Section IV, the convergence threshold characterized by the relative increment/decrement in the objective value is set to $\epsilon = 10^{-4}$. All simulations are carried out using MATLAB R2019b and Intel(R) Core i3 Processor, and the points are obtained by averaging over 200 channel realizations.

We compare the performance of our proposed perfect and robust transceiver designs to the following benchmark schemes: **1) CVX solution:** The optimal transmit covariance matrix $\boldsymbol{Q}$ is derived using the well-known CVX tool [34]. **2) Equal power allocation (EPA):** For the total power constraint, we design the optimal $\boldsymbol{Q}$ by allocating all transmit power equally among all eigenchannels of $\boldsymbol{H}$. In particular, this transmission strategy is asymptotically optimal for MIMO systems in the high-SNR regime. For the individual per-antenna power constraints, the optimal $\boldsymbol{Q} = P/N_t\boldsymbol{I}_{N_t}$ is simply chosen. **3) Strongest eigenchannel transmission (SET):** For the total power constraint, the optimal $\boldsymbol{Q}$ is designed for ensuring all transmit power is allocated to the strongest eigenchannel of $\boldsymbol{H}$. Note that this scheme is asymptotically optimal for MIMO systems in the low-SNR regime. Furthermore, for satisfying the per-antenna power constraints, we simply modify the optimal $\boldsymbol{Q}$ by adjusting the transmit power allocated to each element of the strongest eigenchannel. **4) Nonrobust design:** In this scheme, the optimal $\boldsymbol{Q}$ is obtained only based on the estimated channels, and then substituted into the expressions of the average information rate and MSE.



(a) $\mathrm{Tr}(\boldsymbol{Q}) \le P$

(b) $\mathrm{Tr}(\boldsymbol{\Omega}_i\boldsymbol{Q}) \le P_i, 1 \le i \le N_t$

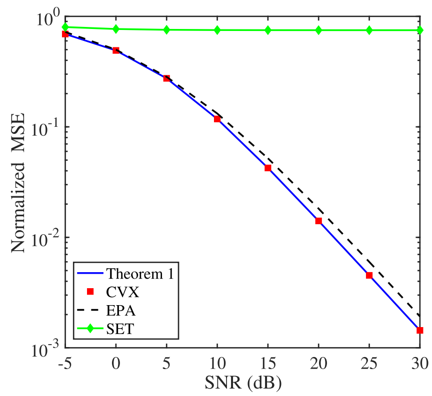
(c) $\mathrm{SNR} = 20$dB

Fig. 1. Achievable information rate of different algorithms versus SNR under the (a) total power constraint and (b) per-antenna power constraints; (c) shows the convergence of Algorithm 1.

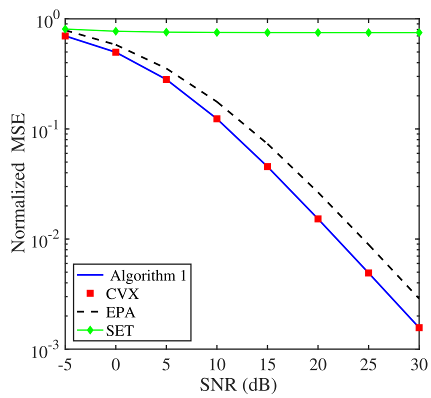
A. Perfect CSI

Fig. 1 (a) and Fig. 1 (b) show the achievable capacity versus SNR for all studied algorithms under the total and per-antenna power constraints, respectively. It is observed from Fig. 1 (a) that the unified closed-form solution given by Theorem 1 achieves the same maximum capacity as the CVX solution. The EPA and SET schemes perform close to the optimal closed-form solution in the high-SNR and low-SNR regime, respectively, as discussed above. Furthermore, in Fig. 1 (b) under per-antenna power constraints, Algorithm 1 is seen to converge to the maximum capacity achieved by the CVX solution, and its convergence behavior is shown in Fig. 1 (c), where the power margin is defined as the difference between the actual antenna transmit power and its corresponding maximum threshold, i.e. $\mathrm{Tr}(\boldsymbol{\Omega}_i\boldsymbol{Q}) - P_i$, $\forall i$. It is clear from Fig. 1 (c) that for each antenna the power margin converges to zero within 100 iterations, implying that Algorithm 1 is capable of finding the KKT conditions based optimal solution at an acceptable convergence speed. In particular, compared to Fig. 1 (a), we find a larger performance gap between Algorithm 1 and the EPT scheme in Fig. 1 (b), since the asymptotical optimality of the EPT scheme in the high-SNR regime is not established under the per-antenna power constraints.

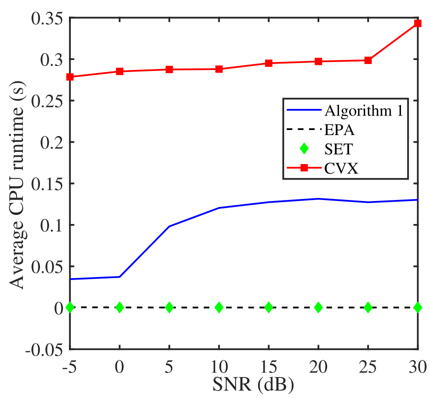
[4] In fact, our work is applicable to arbitrary transmit and receive correlation matrices. Note that the exponential correlation model may not be suitable when a distributed antenna array is considered.



(a) $\mathrm{Tr}(\boldsymbol{Q}) \leq P$



(b) $\mathrm{Tr}(\boldsymbol{\Omega}_i \boldsymbol{Q}) \leq P_i, 1 \leq i \leq N_t$



(c) $\mathrm{Tr}(\boldsymbol{\Omega}_i \boldsymbol{Q}) \leq P_i, 1 \leq i \leq N_t$

Fig. 2. Normalized MSE and average CPU runtime of different algorithms versus SNR under (a) total power constraint; (b)/(c) per-antenna power constraints.

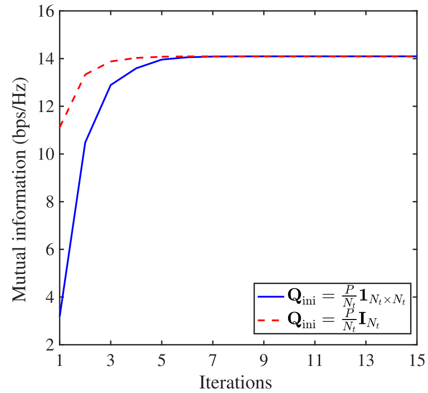
Fig. 2 (a) and Fig. 2 (b) plot the normalized MSE, denoted by $\mathrm{Tr}\big(\big(\boldsymbol{I} + \boldsymbol{R}_{\mathrm{n}}^{-1}\boldsymbol{H}\boldsymbol{Q}\boldsymbol{H}^{\mathrm{H}}\big)^{-1}\big)/d$, versus SNR under the total and per-antenna power constraints for all studied algorithms, respectively. Like in Fig. 1, both the unified closed-form solution in Theorem 1 directly aiming for satisfying the total power constraint and Algorithm 1 aiming for satisfying per-antenna power constraints achieve the almost identical normalized MSE performance to the CVX solution, as shown in Fig. 2 (a) and Fig. 2 (b), respectively, which again demonstrates their global optimality. Similarly, the EPT scheme performs closer to the optimal design in Fig. 2 (a) than that in Fig. 2 (b). Furthermore, in Fig. 2 (c), we illustrate the average CPU runtime of all studied algorithms. It is clear that the time overhead of Algorithm 1 is remarkably reduced compared to the CVX solution, which demonstrates its low complexity. Moreover, due to the suboptimal closed-form transmit covariance matrices adopted by both EPT and SET schemes, their average CPU runtime is much less than that of Algorithm 1.

B. Imperfect CSI

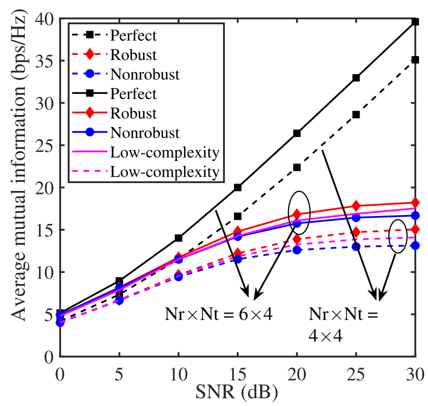
In this subsection, we consider the imperfect MIMO CSI, and focus our attention on evaluating both the average capacity and normalized MSE performance of the proposed robust designs under the total and per-antenna power constraints, respectively. In order to demonstrate the effectiveness of the proposed robust designs, a low-complexity approximate scheme in [5] is adopted as a benchmark. With this scheme, the term $\boldsymbol{K}_{\mathrm{n}} = \sigma_{\mathrm{n}}^2 \boldsymbol{I} + \mathrm{Tr}(\boldsymbol{R}_{\mathrm{T}}\boldsymbol{Q})\boldsymbol{R}_{\mathrm{R}}$ in (31) is relaxed to $\boldsymbol{K}_{\mathrm{n}} \preceq (\sigma_{\mathrm{n}}^2 + \lambda_{\max}(\boldsymbol{R}_{\mathrm{R}})\mathrm{Tr}(\boldsymbol{R}_{\mathrm{T}}\boldsymbol{Q}))\boldsymbol{I}$, based on which both average capacity maximization and average MSE minimization become convex and the closed-form solutions can be derived under the total power constraint, as shown in **P5.3** and **P6.3**.

Firstly, Fig. 3 (a) plots the convergence behavior of the Prop-Alt algorithm in **P5.1** for the robust transceiver design with $p_t = p_r = 0.6$. It is observed that the Prop-Alt algorithm monotonically converges to the same maximum within 10 iterations, regardless of the initial value of $\boldsymbol{Q}$. This fact confirms the stability and rapid convergence of the Prop-Alt algorithm. Fig. 3 (b) and Fig. 3 (c) also show the average capacity achieved by different algorithms versus SNR for different antenna setups $N_r \times N_t$ and transmit-receive correlation $p_t \times p_r$, respectively. Specifically, in Fig. 3 (b) with two different antenna setups, we readily observe the higher average capacity of each algorithm for $N_r \times N_t = 6\times 4$ than for $N_r \times N_t = 4 \times 4$ due to the expansion of the antenna array. Naturally, the perfect design using the unified closed-form solution in Theorem 1 has the best capacity performance. However, as the SNR increases, the achievable average capacity of the robust design given by **P5.1** firstly increases and then becomes saturated, since a high SNR can lead to the increasing power of the equivalent noise $\boldsymbol{K}_n$. Moreover, the low-complexity design in [5] achieves lower average capacity than the robust design, since an upper bound of the original optimization objective is studied instead, but it still performs better than the nonrobust design. Furthermore, the performance gain of the robust design over the non-robust design is enlarged in the high-SNR regime. Similarly, in Fig. 3 (b) with different transmit-receive correlation, the robust designs given by **P5.1** for $p_t \neq 0, p_r \neq 0$, **P5.2** for $p_t = 0, p_r = 0.6$ and **P5.3** for $p_t = 0.6, p_r = 0$ all outperform the corresponding non-robust designs in terms of the average capacity. In particular, the capacity advantage of these robust designs is more evident for the strong transmit correlation $p_t = 0.6$. The high average capacity is clearly observed for $p_t = p_r = 0.6$, as compared to both $p_t = 0.6, p_r = 0$ and $p_t = 0, p_r = 0.6$, which may be attributed to the stronger correlation of the CSI error matrix.

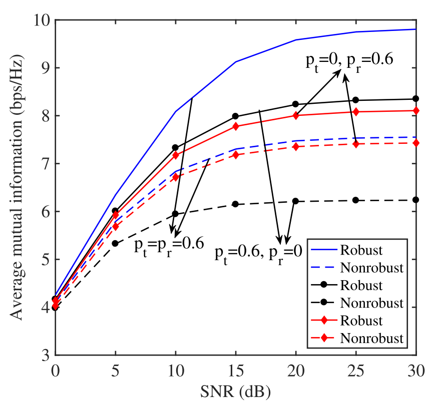
As for per-antenna power constraints, Fig. 4 shows the average capacity performance and CPU runtime versus SNR of different algorithms, where transmit-receive correlations of $p_t = p_r = 0.6$ and $p_t = 0.6, p_r = 0$ are considered. To be specific, it is observed from Fig. 4 (a) that for different



(a) $p_t = p_r = 0.6$

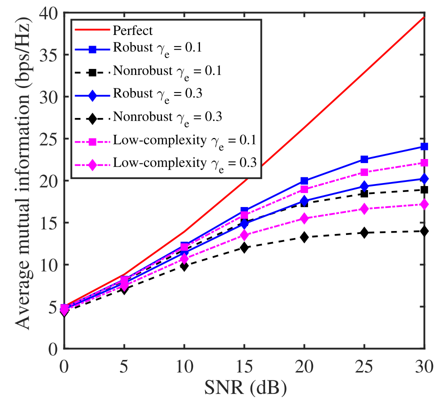


(b) $p_t = p_r = 0.6$

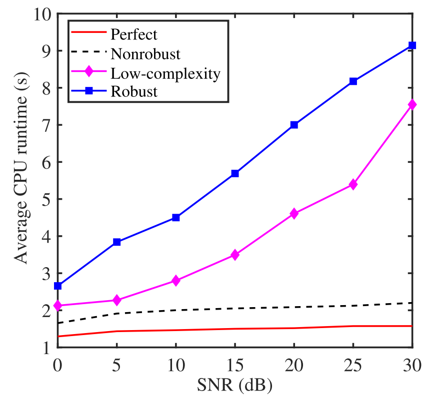


(c) $N_r \times N_t = 4 \times 6$

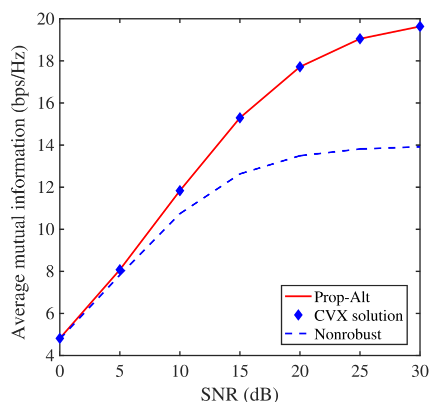
Fig. 3. (a) The convergence of the Prop-Alt algorithm in **P5.1** for $p_t = p_r = 0.6$; Average capacity achieved by different algorithms versus SNR under the total power constraint.



(a) $p_t = p_r = 0.6$



(b) $p_t = p_r = 0.6$

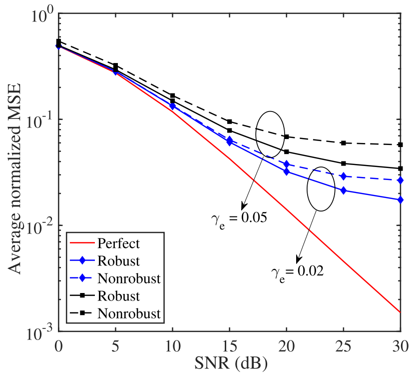


(c) $p_t = 0.6, p_r = 0$

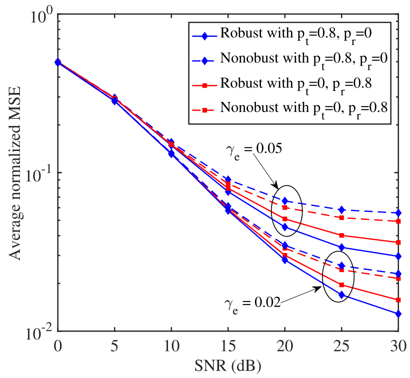
Fig. 4. Average capacity achieved by different algorithms versus SNR under the per-antenna power constraints.

error variance γ_e, the average mutual information achieved by the robust solution derived from the Prop-Alt algorithm decreases with the increasing error variance γ_e, and still outperforms both the non-robust design and the lower-bound design for each γ_e. Moreover, a larger performance improvement of the robust design over the non-robust design is observed in the high-SNR regime or for a large γ_e. Similar to Fig. 2 (c), we also compare the average CPU runtime of different algorithms in Fig. 4 (b). Obviously, the Prop-Alt algorithm has the largest time consumption due to the extra SCA-based iterations compared to other schemes. Actually, in the step 3 of Algorithm 1, the low-complexity, perfect and nonrobust designs all can lead to the optimal closed-form solution without the SCA operation, and thus exhibit lower time overhead. In particular, the time overhead of the perfect design is lowest due to the simple optimization problem structure.

Futhermore, as seen from Fig. 4 (c) with $p_t = 0.6, p_r = 0$, the robust solution derived from the Prop-Alt algorithm in **P7.1** achieves almost the same performance as the CVX solution, which further verifies the global optimality of the Prop-Alt algorithm for $p_t = 0.6, p_r = 0$. Furthermore, in Fig. 2 (c), we illustrate the average CPU runtime of all studied algorithms. It is clear that Algorithm 1 realizes the same MSE performance as the CVX solution with the remarkably reduced time overhead, which demonstrates the low complexity of

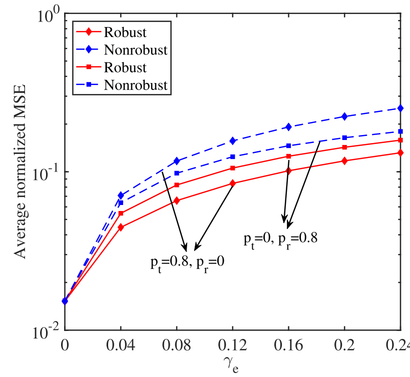


(a) $p_t = 0.6, p_r = 0.8$

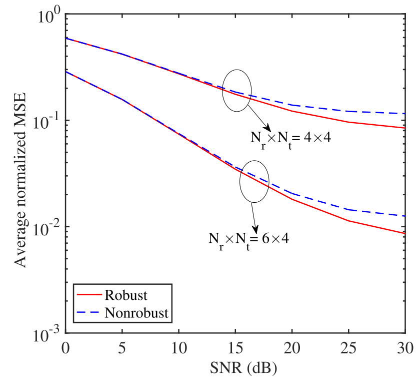


(b) $p_t = 0\ 8, p_r = 0$ and $p_t = 0, p_r = 0\ 8$

Fig. 5. Average normalized MSEs achieved by different algorithms versus SNR under the total power constraint.



(a) SNR $= 20$dB



(b) $p_r = p_t = 0.8, \gamma_e = 0.04$

Fig. 6. Average normalized MSE of robust and nonrobust algorithms versus SNR under individual per-antenna power constraints for (a) different antenna setups $N_t \times N_r$ and (b) different transmit-receive correlation (p_t, p_r).

Algorithm 1. Moreover, due to the suboptimal closed-form transmit covariance matrices adopted by both EPT and SET schemes, their average CPU runtime is much less than that of Algorithm 1.

On the other hand, Fig. 5 plots the average normalized MSE of the robust designs in **P6.1**, **P6.2**, **P6.3** versus SNR for different transmit-receive correlations, where the total power constraint is considered. As expected from the design philosophy, in Fig. 5 (a) with $p_t = 0.6, p_r = 0.8$, the average normalized MSE achieved by the robust design in **P6.1** increases with the error variance γ_e and still outperforms the non-robust design, however, it performs worse than the perfect design, similarly to that in Fig. 3 (b). Similar conclusions can also be obtained from Fig.5 (b) considering both $p_t = 0.8, p_r = 0$ and $p_t = 0, p_r = 0.8$. It is worth noting that using the unified closed-form solution in Theorem 1, the global optimality of our robust designs in **P6.2** for $p_t = 0.8, p_r = 0$ and **P6.3** for $p_t = 0, p_r = 0.8$ can both be validated in Fig. 5 (b).

Finally, upon considering per-antenna power constraints, Fig. 6 presents the average normalized MSE performance achieved by the Prop-Alt algorithm in **P7.2** for the general case of $p_r = p_t = 0.8$. More specifically, Fig. 6 (a) and Fig. 6 (b) respectively consider different error variance γ_e and antenna setups $N_r \times N_t$. It is clear from Fig.6 (a) that the larger γ_e is, the worse the average normalized MSE becomes. Moreover, as γ_e increases, the performance gap between the robust and non-robust designs becomes larger. Being consistent with Fig. 4 (a), the higher normalized MSE of the robust design under $N_r \times N_t = 6 \times 4$ can be observed from Fig. 6 (b).

VII. Conclusion

For MIMO systems, the key task of many transceiver optimization problems is to optimize the positive semidefinite transmit covariance matrix. By considering the realistic multiple weighted power constraints and the imperfect CSI, a unified framework was proposed for deriving the water-filling structure of the optimal covariance matrix, which helps reveal interesting underlying relationships among solutions of rather different MIMO transceiver designs. Specifically, for the case of multiple power constraints, we jointly applied Pareto optimization theory and Lagrangian dual theory to derive the corresponding optimal closed-form solution, which can be found by the proposed modified subgradient method. For the imperfect CSI case, various alternating optimization algorithms were proposed to address the non-convexity of the robust design and thus enable the unified framework directly applicable. Overall, this KKT conditions based unified framework can be applied to a wide range of applications in communications. Finally, our numerical results verified the rapid convergence and global optimality of the proposed algorithms.

References

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

[2] H. Sampath, P. Stoica, and A. Paulraj, “Generalized linear precoder and decoder design for MIMO channels using the weighted MMSE criterion,” *IEEE Trans. Commun.*, vol. 49, no. 12, pp. 2198–2206, Dec. 2001.

[3] D. P. Palomar, J. M. Cioffi, and M. A. Lagunas, “Joint Tx-Rx beamforming design for multicarrier MIMO channels: A unified framework for convex optimization,” *IEEE Trans. Signal Process.*, vol. 51, no. 9, pp. 2381–2401, Sep. 2003.

[4] A. Scaglione, P. Stoica, S. Barbarossa, G. B. Giannakis, and H. Sampath, “Optimal designs for space-time linear precoders and decoders,” *IEEE Trans. Signal Process.*, vol. 50, no. 5, pp. 1051–1064, May 2002.

[5] S. Wang, S. Ma, C. Xing, S. Gong, J. An, and H. V. Poor, “Optimal training design for MIMO systems with general power constraints,” *IEEE Trans. Signal Process.*, vol. 66, no. 14, pp. 3649–3664, Jul. 2018.

[6] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[7] D. P. Palomar, “Unified framework for linear MIMO transceivers with shaping constraints,” *IEEE Commun. Lett.*, vol. 8, no. 12, pp. 697–699, Dec. 2004.

[8] I. D. Schizas, G. B. Giannakis, and Z. Q. Luo, “Distributed estimation using reduced-dimensionality sensor observations,” *IEEE Trans. Signal Process.*, vol. 55, no. 8, pp. 4284–4299, Aug. 2007.

[9] W. Guan and H. Luo, “Joint MMSE transceiver design in non-regenerative MIMO relay systems,” *IEEE Commun. Lett.*, vol. 12, no. 7, pp. 517–519, Jul. 2008.

[10] H. Sun and Z. Ding, “Iterative transceiver design for MIMO ARQ retransmissions with decision feedback detection,” *IEEE Trans. Signal Process.*, vol. 55, no. 7, pp. 3405–3416, Jul. 2007.

[11] W. Yu and J. M. Cioffi, “Sum capacity of Gaussian vector broadcast channels,” *IEEE Trans. Inf. Theory*, vol. 50, no. 9, pp. 1875–1892, Sep. 2004.

[12] C. Xing, Z. Fei, Y. Zhou, Z. Pan, and H. Wang, “Transceiver designs with matrix-version water-filling architecture under mixed power constraints,” *Sci. China Inf. Sci.*, vol. 59, no. 10, pp. 1–12, Oct. 2016.

[13] A. Liu, Y. Liu, H. Xiang, and W. Luo, “MIMO B-MAC interference network optimization under rate constraints by polite water-filling and duality,” *IEEE Trans. Signal Process.*, vol. 59, no. 1, pp. 263–276, Jan. 2011.

[14] O. Munoz-Medina, J. Vidal, and A. Agustin, “Linear transceiver design in nonregenerative relays with channel state information,” *IEEE Trans. Signal Process.*, vol. 55, no. 6, pp. 2593–2604, Jun. 2007.

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

[16] Y. Yu and Y. Hua, “Power allocation for a MIMO relay system with multiple-antenna users,” *IEEE Trans. Signal Process.*, vol. 58, no. 5, pp. 2823–2835, May 2010.

[17] C. Xing, D. Li, S. Ma, Z. Fei, and J. Kuang, “Robust transceiver design for MIMO-OFDM systems based on cluster water-filling,” *IEEE Commun. Lett.*, vol. 17, no. 7, pp. 1451–1454, Jul. 2013.

[18] F. Gao, T. Cui, and A. Nallanathan, “Optimal training design for channel estimation in decode-and-forward relay networks with individual and total power constraints,” *IEEE Trans. Signal Process.*, vol. 56, no. 12, pp. 5937–5949, Dec. 2008.

[19] 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.

[20] A. Liu, Y. Liu, H. Xiang, and W. Luo, “Polite water-filing for weighted sum-rate maximization in MIMO B-MAC networks under multiple linear constraints,” *IEEE Trans. Signal Process.*, vol. 60, no. 2, pp. 834–847, Feb. 2012.

[21] C. Xing, S. Ma, and Y. Zhou, “Matrix-monotonic optimization for MIMO systems,” *IEEE Trans. Signal Process.*, vol. 63, no. 2, pp. 334–348, Jan. 2015.

[22] Q. Shi, M. Razaviyayn, Z.-Q. Luo, and C. He, “An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel,” *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4331–4340, Sep. 2011.

[23] C. Xing, S. Ma, Z. Fei, Y.-C. Wu, and H. V. Poor, “A general robust linear transceiver design for multi-hop amplify-and-forward MIMO relaying systems,” *IEEE Trans. Signal Process.*, vol. 61, no. 5, pp. 1196–1209, Mar. 2013.

[24] M. Ding and S. D. Blostein, “MIMO minimum total MSE transceiver design with imperfect CSI at both ends,” *IEEE Trans. Signal Process.*, vol. 57, no. 3, pp. 1141–1150, Mar. 2009.

[25] M. Ding and S. D. Blostein, “Maximum mutual information design for MIMO systems with imperfect channel knowledge,” *IEEE Trans. Inf. Theory*, vol. 56, no. 10, pp. 4793–4801, Oct. 2010.

[26] A. W. Marshall, I. Olkin, and B. C. Arnold, *Inequalities: Theory of Majorization and Its Applications*. New York, NY, USA: Academic, 1979.

[27] C. Xing, Y. Ma, Y. Zhou, and F. Gao, “Transceiver optimization for multi-hop communications with per-antenna power constraints,” *IEEE Trans. Signal Process.*, vol. 64, no. 6, pp. 1519–1534, Mar. 2016.

[28] C. Xing, Y. Jing, and Y. Zhou, “On weighted MSE model for MIMO transceiver optimization,” *IEEE Trans. Veh. Technol.*, vol. 66, no. 8, pp. 7072–7085, Aug. 2017.

[29] M. Vu, “MIMO capacity with per-antenna power constraint,” in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, Houston, TX, USA, Dec. 2011, pp. 1–5.

[30] X. Zhang, D. P. Palomar, and B. Ottersten, “Statistically robust design of linear MIMO transceivers,” *IEEE Trans. Signal Process.*, vol. 56, no. 8, pp. 3678–3689, Aug. 2008.

[31] A. Garg, M. R. Bhatnagar, and O. B. B. Vrigneau, “Performance analysis of erroneous feedback-based MIMO system over Nakagami-m fading channels,” *IEEE Trans. Commun.*, vol. 67, no. 8, pp. 5403–5418, Aug. 2019.

[32] M. Razaviyayn, M. Hong, and Z.-Q. Luo, “A unified convergence analysis of block successive minimization methods for nonsmooth optimization,” *SIAM J. Optim.*, vol. 23, no. 2, pp. 1126–1153, Jun. 2013.

[33] S. A. Jafar and A. Goldsmith, “Multiple-antenna capacity in correlated Rayleigh fading with channel covariance information,” *IEEE Trans. Wireless Commun.*, vol. 4, no. 3, pp. 990–997, May 2005.

[34] M. C. Grant and S. P. Boyd, “The CVX users’ guide release 2.1,” CVX Res., Inc., Tech. Rep., 2015. [Online]. Available: http://cvxr.com/cvx/doc/CVX.pdf

Shiqi Gong received the B.S. and Ph.D. degrees in electronic engineering from Beijing Institute of Technology, China, in 2014 and 2020, respectively. She is currently a Post-Doctoral Fellow with the State Key Laboratory of Internet of Things for Smart City, University of Macau, China. Her research interests include intelligent reflecting surface, signal processing, physical-layer security, resource allocation, and convex optimization. She was a recipient of the Best Ph.D. Thesis Award of Beijing Institute of Technology in 2020.

Chengwen Xing (Member, IEEE) received the B.Eng. degree from Xidian University, Xi’an, China, in 2005, and the Ph.D. degree from the University of Hong Kong, Hong Kong, China, in 2010. Since September 2010, he has been with the School of Information and Electronics, Beijing Institute of Technology, Beijing, China, where he is currently a Full Professor. From September 2012 to December 2012, he was a Visiting Scholar with the University of Macau. His research interests include machine learning, statistical signal processing, convex optimization, multivariate statistics, array signal processing, and high-frequency band communication systems.

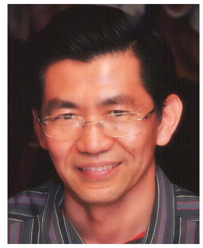
Yindi Jing (Senior Member, IEEE) received the B.Eng. and M.Eng. degrees in automatic control from the University of Science and Technology of China, Hefei, China, in 1996 and 1999, respectively, and the M.Sc. and Ph.D. degrees in electrical engineering from California Institute of Technology, Pasadena, CA, USA, in 2000 and 2004, respectively.

From October 2004 to August 2005, she was a Post-Doctoral Scholar with the Department of Electrical Engineering, California Institute of Technology. From February 2006 to June 2008, she was a Post-Doctoral Scholar with the Department of Electrical Engineering and Computer Science, University of California at Irvine, Irvine, CA, USA. She joined the Electrical and Computer Engineering Department, University of Alberta, in 2008, where she is currently a Professor. Her research interests include signal processing and wireless communications, focusing on massive MIMO, cooperative relay networks, fault detection, and channel estimation. She was a member of the IEEE Signal Processing Society Signal Processing for Communications and Networking (SPCOM) Technical Committee from 2015 to 2020, and a member of the NSERC Discover Grant Evaluation Group for Electrical and Computer Engineering from 2017 to 2020. From 2011 to 2016, she was an Associate Editor of IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. She is also a Senior Area Editor of IEEE SIGNAL PROCESSING LETTERS.

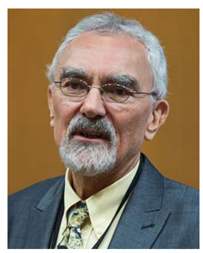
Shuai Wang received the B.Eng. degree (Hons.) in communications engineering from Zhengzhou University in 2005, and the Ph.D. degree (Hons.) in communications engineering from Beijing Institute of Technology, China, in 2012. From September 2010 to September 2011, he was a Visiting Ph.D. Student with the School of Electronics and Computer Science, University of Southampton, U.K. Since July 2012, he has been with the School of Information Science and Electronics, Beijing Institute of Technology, where he is currently an Associate Professor. He served or is serving as the Principal Investigator for 12 research funds, including two granted by the National Science Foundation of China. He has published more than 20 peer-reviewed articles, mainly in leading IEEE journals or conferences. He also holds 17 patents. His research interests include channel estimation, anti-jamming transmission, synchronization techniques, and beamforming. He was a recipient of the (Second Class) Scientific and Technical Progress Award granted by the Ministry of Industry and Information Technology of China. He received the Award for Outstanding Ph.D. Dissertation granted by Beijing Municipal Education Commission in 2013 with other 49 co-winners, nominated from all the Ph.D. graduates that received their degree in Beijing that year.

Jiaheng Wang (Senior Member, IEEE) received the B.E. and M.S. degrees from Southeast University, Nanjing, China, in 2001 and 2006, respectively, and the Ph.D. degree in electronic and computer engineering from the Hong Kong University of Science and Technology, Hong Kong, in 2010.

From 2010 to 2011, he was with the Signal Processing Laboratory, KTH Royal Institute of Technology, Stockholm, Sweden. He also held visiting positions at the Friedrich Alexander University Erlangen-Nürnberg, Nürnberg, Germany, and the University of Macau, Macau. He is currently a Full Professor with the National Mobile Communications Research Laboratory (NCRL), Southeast University. He has published more than 140 articles on international journals and conferences. His research interests include optimization in signal processing and wireless communications. He was a recipient of the Humboldt Fellowship for Experienced Researchers and the Best Paper Award of IEEE GLOBECOM 2019, ADHOCNETS 2019, and WCSP 2014. From 2014 to 2018, he served as an Associate Editor for IEEE SIGNAL PROCESSING LETTERS. Since 2018, he has been serving as a Senior Area Editor for IEEE SIGNAL PROCESSING LETTERS.



Sheng Chen (Fellow, IEEE) received the B.Eng. degree in control engineering from East China Petroleum Institute, Dongying, China, in 1982, and the Ph.D. degree in control engineering from the City University, London, in 1986, and the Higher Doctoral degree and the Doctor of Sciences (D.Sc.) degree from the University of Southampton, Southampton, U.K., in 2005. From 1986 to 1999, he held research and academic appointments at The University of Sheffield, U.K., University of Edinburgh, U.K., and University of Portsmouth, U.K. Since 1999, he has been with the School of Electronics and Computer Science, University of Southampton, where he is a Professor of intelligent systems and signal processing. He is a Distinguished Adjunct Professor with King Abdulaziz University, Jeddah, Saudi Arabia. He has published over 650 research articles. His research interests include adaptive signal processing, wireless communications, modeling and identification of nonlinear systems, neural network and machine learning, intelligent control system design, and evolutionary computation methods and optimization. He is a fellow of the U.K. Royal Academy of Engineering, a fellow of IET, and an original ISI Highly Cited Researcher in engineering in March 2004. He has more than 13 000 Web of Science citations and more than 27 000 Google Scholar citations.



Lajos Hanzo (Fellow, IEEE) received the master's and Ph.D. degrees from the Technical University (TU) of Budapest in 1976 and 1983, respectively, the Doctor of Sciences (D.Sc.) degree from the University of Southampton in 2004, the Honorary Doctorate from the TU of Budapest in 2009, and the Honorary Doctorates from the University of Edinburgh in 2015. He is a Foreign Member of the Hungarian Academy of Sciences and a former Editor-in-Chief of the IEEE Press. He has served several terms as a Governor for IEEE ComSoc and VTS. He has published more than 2000 contributions at IEEE Xplore, 19 Wiley-IEEE Press books and has helped the fast-track career of 123 Ph.D. students. Over 40 of them are professors at various stages of their careers in academia and many of them are leading scientists in the wireless industry. He is also a fellow of the Royal Academy of Engineering (FREng), IET, and EURASIP (http://www-mobile.ecs.soton.ac.uk; https://en.wikipedia.org/wiki/Lajos_Hanzo).