Globally Optimal Linear Precoders for Finite Alphabet Signals Over Complex Vector Gaussian Channels

Chengshan Xiao, *Fellow, IEEE*, Yahong Rosa Zheng, *Senior Member, IEEE*, and Zhi Ding, *Fellow, IEEE*

***Abstract*—We study the design optimization of linear precoders for maximizing the mutual information between finite alphabet input and the corresponding output over complex-valued vector channels. This mutual information is a nonlinear and non-concave function of the precoder parameters, posing a major obstacle to precoder design optimization. Our work presents three main contributions: First, we prove that the mutual information is a concave function of a matrix which itself is a quadratic function of the precoder matrix. Second, we propose a parameterized iterative algorithm for finding optimal linear precoders to achieve the global maximum of the mutual information. The proposed iterative algorithm is numerically robust, computationally efficient, and globally convergent. Third, we demonstrate that maximizing the mutual information between a discrete constellation input and the corresponding output of a vector channel not only provides the highest practically achievable rate but also serves as an excellent criterion for minimizing the coded bit error rate. Our numerical examples show that the proposed algorithm achieves mutual information very close to the channel capacity for channel coding rate under 0.75, and also exhibits a large gain over existing linear precoding and/or power allocation algorithms. Moreover, our examples show that certain existing methods are susceptible to being trapped at locally optimal precoders.**

***Index Terms*—Finite alphabet input, linear precoding, mutual information, optimization, vector Gaussian noise channel.**

I. Introduction

LINEAR transmit precoding has been a popular research topic in multiple-input multiple-output (MIMO) system optimization, as evidenced by [1]–[17] and references therein. Existing methods typically belong to three categories: (a) diversity-driven designs; (b) rate-driven designs; and (c) designs based on minimum mean squared error (MMSE) or maximum signal-to-noise ratio (SNR). The first category applies pairwise error probability analysis to maximize diversity order as in [18] but may not achieve the highest coding gain [19]. The second category often utilizes (ergodic or outage) capacity as design criteria for precoder optimization. However, most such designs rely on the impractical Gaussian input assumption, which often leads to substantial performance degradation when applied with actual input of finite alphabet data [20], [21]. The third category uses MMSE or SNR as the figure of merit to design a linear precoder. Linear MMSE estimation strategy is globally optimal if the channel inputs and noise are both (independent) Gaussian [10]. However, when the inputs belong to finite alphabets, such strategy is also not optimal.

In fact, several recent works have begun to study MIMO precoding design for maximizing mutual information under discrete-constellation inputs [20]–[24]. In [20], a seminal result was presented for power allocation based on mercury/waterfilling (M/WF) that maximizes the input-output mutual information over parallel channels with finite alphabet inputs. The significance of M/WF is that, for given equal probable constellations, stronger channels may receive less allocated power; whereas for channels of equal strength, denser constellations may receive larger power allocation. These results indicate that power allocation depends not only on channel gain but also on the constellation for finite alphabet inputs. Therefore, M/WF is very different from the classic waterfilling (CWF) policy [25] for Gaussian inputs. However, M/WF inherits a feature of the CWF by constraining the M/WF precoding matrix $\mathbf{G}$ as diagonal if the channel matrix $\mathbf{H}$ is also diagonal [20]. Unfortunately, this diagonal constraint on $\mathbf{G}$ makes the M/WF power allocation *sub-optimal* even for parallel channels with discrete constellation inputs. The works in [21] and [22] proposed iterative algorithms based on necessary but not sufficient conditions. Hence, such algorithms do not guarantee global optimality. For real-valued vector channel models in which all signals and matrices [see (1)] have real-valued entries, [23] showed that: a) the mutual information between channel input and output, $\mathcal{I}(\mathbf{x};\mathbf{y})$, is a function of $\mathbf{G}^t\mathbf{H}^t\mathbf{H}\mathbf{G}$ with $(\cdot)^t$ denoting matrix transpose; b) the left singular vectors of optimal $\mathbf{G}$ can be the right singular vectors of $\mathbf{H}$; c) the mutual information is a concave function of the squared singular values of $\mathbf{G}$ if its right singular vectors are fixed. However, [23] pointed out that optimizing the right singular vectors of the precoder "*seems to be an extremely difficult problem*". Independently, [24] stated that $\mathcal{I}(\mathbf{x};\mathbf{y})$ is a concave function of $\mathbf{G}^t\mathbf{H}^t\mathbf{H}\mathbf{G}$ for real-valued signals and channels, with an incomplete proof. An iterative algorithm was further presented in [24] to solve the real-valued precoder $\mathbf{G}$. The

Manuscript received November 29, 2010; revised March 18, 2011; accepted March 18, 2011. Date of publication April 07, 2011; date of current version June 15, 2011. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Shahram Shahbazpanahi. The work of C. Xiao and Y. R. Zheng was supported in part by the NSF by Grants ECCS-0846486 and CCF-0915846, and by the Office of Naval Research by Grants N00014-09-1-0011 and N00014-10-1-0174. The work of Z. Ding was supported in part by the NSF by Grants CCF-0830706 and ECCS-1028729. The material in this paper has been presented at the IEEE ICASSP 2011.

C. Xiao and Y. R. Zheng are with the Department of Electrical and Computer Engineering, Missouri University of Science and Technology, Rolla, MO 65409 USA (e-mail: xiaoc@mst.edu; zhengyr@mst.edu).

Z. Ding is with the Department of Electrical and Computer Engineering, University of California, Davis, CA 95616 USA. He was on sabbatical leave at Southeast University, Nanjing, China (e-mail: zding@ucdavis.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2011.2140112

simulation results in [24] indicate that the algorithm in [24] converges to the same maximum mutual information as that of [22], but at more than twice the convergence rate. However, we will show that the test results in [24] is not globally optimal, either.

It should be pointed out that confining channel and precoder parameters as well as the associated signals to real-field as in [23] and [24] usually leads to suboptimal designs. Consequently, these methods may achieve mutual information results that fall far short of the achievable global maximum. In this paper, we reopen the exploration of linear precoding optimization for complex vector channels. Our work investigates general precoding in the complex field for maximizing input-output (I/O) mutual information. Our major contribution in this work consists of the proof of the aforementioned concavity property and our presentation of an iterative algorithm that converges globally to maximum I/O mutual information for complex vector Gaussian channels. As will become clearer, the concavity proof and the iterative algorithm for linear precoding in complex channels are radically different from their real-field counterpart.

Notation: Uppercase (lowercase) boldface letters denote matrices (column vectors), $\mathrm{vec}(\mathbf{A})$ represents the column vector obtained by stacking the columns of matrix $\mathbf{A}$, $\mathrm{diag}(\mathbf{X})$ stands for a diagonal matrix formed by the diagonal entries of $\mathbf{X}$, $\mathrm{trace}(\cdot)$ represents the trace of a matrix, $E(\cdot)$ stands for ensemble average, $\mathbf{A} \otimes \mathbf{B} (\mathbf{A} \odot \mathbf{B})$ denotes the Kronecker (Hadamard) product, and the superscripts $(\cdot)^t$, $(\cdot)^*$ and $(\cdot)^h$ represent transpose, conjugate, and conjugate transpose operations, respectively.

II. System Model and Preliminaries

Consider a discrete-time complex vector channel, in which the baseband input-output relationship is described by

$$\mathbf{y} = \mathbf{HGx} + \mathbf{v} \tag{1}$$

where $\mathbf{y} \in \mathbb{C}^{N_r \times 1}$ is the received channel output signal; $\mathbf{H} \in \mathbb{C}^{N_r \times N_t}$ is the channel matrix; $\mathbf{G} \in \mathbb{C}^{N_t \times N_t}$ is the linear precoder; $\mathbf{v} \in \mathbb{C}^{N_r \times 1}$ is the circularly symmetric white Gaussian noise vector with covariance matrix $\sigma^2 \mathbf{I}_{N_r}$; and $\mathbf{x} \in \mathbb{C}^{N_t \times 1}$ is the baseband channel input signal of zero mean and covariance $E[\mathbf{xx}^h] = \mathbf{I}_{N_t}$, with $\mathbf{I}_{N_t}$ being the $N_t \times N_t$ identity matrix.

Instead of the traditional assumption of Gaussian signal $\mathbf{x}$, we let the signal $\mathbf{x}$ be equiprobably drawn from well-known discrete constellations such as M-ary phase-shift keying (PSK), pulse-amplitude modulation (PAM), or quadrature amplitude modulation (QAM) of cardinality M. Let the channel $\mathbf{H}$ and the precoder $\mathbf{G}$ be known at the receiver. The corresponding channel output $\mathbf{y}$ has probability density functions of

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{(\pi\sigma^2)^{N_r}} \exp\left(-\frac{\|\mathbf{y} - \mathbf{HGx}\|^2}{\sigma^2}\right) \tag{2}$$

$$p(\mathbf{y}) = E_{\mathbf{x}}\{p(\mathbf{y}|\mathbf{x})\} = \frac{1}{M^{N_t}} \sum_{m=1}^{M^{N_t}} p(\mathbf{y}|\mathbf{x}_m) \tag{3}$$

where $\|\cdot\|$ denotes Euclidean norm. Each input realization $\mathbf{x}_m$ consists of N_t i.i.d. symbols from the M-ary constellation.

The MMSE estimate of $\mathbf{x}$ is the conditional mean

$$\hat{\mathbf{x}} = E\{\mathbf{x}|\mathbf{y}\} = E_{\mathbf{x}}\left\{\mathbf{x}\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right\}. \tag{4}$$

Define the MMSE matrix and the companion MMSE matrix, respectively, as

$$\mathbf{\Phi} \triangleq E\left\{(\mathbf{x} - \hat{\mathbf{x}})(\mathbf{x} - \hat{\mathbf{x}})^h\right\} \tag{5}$$

$$\mathbf{\Psi} \triangleq E\left\{(\mathbf{x} - \hat{\mathbf{x}})(\mathbf{x} - \hat{\mathbf{x}})^t\right\}. \tag{6}$$

Note that in the real field, these two matrices are identical. However, in the complex field, they differ. Also, the channel mutual information between the discrete input $\mathbf{x}$ and the channel output $\mathbf{y}$ is given by

$$\mathcal{I}(\mathbf{x};\mathbf{y}) = N_t \log_2 M - \frac{1}{M^{N_t}} \sum_{m=1}^{M^{N_t}} E_{\mathbf{v}}\left\{\log_2 \sum_{k=1}^{M^{N_t}} e^{-d_{m,k}}\right\} \tag{7}$$

where $d_{m,k} = \sigma^{-2}(\|\mathbf{HG}(\mathbf{x}_m - \mathbf{x}_k) + \mathbf{v}\|^2 - \|\mathbf{v}\|^2)$.

We are now in a position to state the following proposition while omitting the proof for brevity.

Proposition 1: Let $\mathbf{U}$ and $\mathbf{V}$ be unitary matrices with appropriate dimensions. Then the following relationships hold:

$$\mathcal{I}(\mathbf{Ux};\mathbf{y}) = \mathcal{I}(\mathbf{Ux};\mathbf{Vy}) \tag{8}$$

$$\mathcal{I}(\mathbf{x};\mathbf{y}) \neq \mathcal{I}(\mathbf{Ux};\mathbf{y}). \tag{9}$$

Apparently, (8) implies that a linear unitary transformation at the channel output does not alter mutual information. This is not surprising since unitary transformation of the channel output vector preserves the original output information. However, (9) implies that the mutual information can change if the discrete input vector undergoes a rotational transformation (by a unitary matrix). It is this relationship that motivates the quest for optimal linear precoding in MIMO wireless communications.

Unlike the cases involving Gaussian input, finding precoder $\mathbf{G}$ to maximize $\mathcal{I}(\mathbf{x};\mathbf{y})$ is much more complicated. Indeed, the complex expression of $\mathcal{I}(\mathbf{x};\mathbf{y})$ is the major reason that most works on capacity-maximizing precoders assumed Gaussian input $\mathbf{x}$. The goal of this paper is to present both theory and algorithms for globally optimal linear precoder $\mathbf{G}$, to maximize mutual information $\mathcal{I}(\mathbf{x};\mathbf{y})$ of complex-field vector Gaussian channels with finite alphabet inputs $\mathbf{x}$.

Before ending this section, we introduce a few notations and operators for the complex first- and second-order derivatives of real-valued and complex-valued functions and matrices.

We adopt the formal partial complex derivative of a real-valued scalar function f [26]

$$\frac{\partial f}{\partial z} \triangleq \frac{1}{2}\left[\frac{\partial f}{\partial \mathrm{Re}(z)} - j\frac{\partial f}{\partial \mathrm{Im}(z)}\right] \tag{10}$$

$$\frac{\partial f}{\partial z^*} \triangleq \frac{1}{2}\left[\frac{\partial f}{\partial \mathrm{Re}(z)} + j\frac{\partial f}{\partial \mathrm{Im}(z)}\right] \tag{11}$$

where z is a complex-valued variable. For a complex-valued matrix $\mathbf{Z}$, the partial derivatives of a real-valued scalar function f are matrices

$$\frac{\partial f}{\partial \mathbf{Z}} = \left[\frac{\partial f}{\partial \mathbf{Z}_{m,n}}\right] \quad \text{and} \quad \frac{\partial f}{\partial \mathbf{Z}^*} = \left[\frac{\partial f}{\partial \mathbf{Z}^*_{m,n}}\right] \tag{12}$$

where $\mathbf{Z}_{m,n}$ denotes the (m,n)th element of $\mathbf{Z}$.

For a complex-valued scale function denoted by $g = u + jv$, we adopt a simple notation of

$$\frac{\partial g}{\partial z} = \frac{\partial u}{\partial z} + j\frac{\partial v}{\partial z}$$

where $\frac{\partial u}{\partial z}$ and $\frac{\partial v}{\partial z}$ follow the definition given by (10). This notation simply follows the definitions given in [26].

Let $\mathbf{F}(\mathbf{Z}, \mathbf{Z}^*)$ be a complex matrix function of $\mathbf{Z}$ and $\mathbf{Z}^*$. Then the Jacobian matrices of $\mathbf{F}$ with respect to $\mathbf{Z}$ and $\mathbf{Z}^*$ are respectively given by [26]

$$\mathcal{D}_{\mathbf{Z}}\mathbf{F} \triangleq \frac{\partial \mathrm{vec}(\mathbf{F})}{\partial \mathrm{vec}^t(\mathbf{Z})} \quad \text{and} \quad \mathcal{D}_{\mathbf{Z}^*}\mathbf{F} \triangleq \frac{\partial \mathrm{vec}(\mathbf{F})}{\partial \mathrm{vec}^t(\mathbf{Z}^*)}. \tag{13}$$

Let $\mathbf{Z}_0$ and $\mathbf{Z}_1$ be two complex-valued matrices, and let f be a real-valued scalar function of $\mathbf{Z}_0$ and $\mathbf{Z}_1$. Then, the complex Hessian of f with respect to $\mathbf{Z}_0$ and $\mathbf{Z}_1$ is defined by [26]

$$\mathcal{H}_{\mathbf{Z}_0,\mathbf{Z}_1} f \triangleq \frac{\partial}{\partial \mathrm{vec}^t(\mathbf{Z}_0)}\left[\frac{\partial f}{\partial \mathrm{vec}^t(\mathbf{Z}_1)}\right]^t. \tag{14}$$

III. Theoretical Results

A. *Concavity of the Mutual Information*

In this section, we present main results with respect to the mutual information for the complex-field vector channel described by (1). Two theorems form the foundation for developing new algorithms to globally maximize $\mathcal{I}(\mathbf{x};\mathbf{y})$ with linear precoders. We utilize seven lemmas for proving the two theorems. The proofs of all the theorems and lemmas of this section are relegated to Appendix A.

Theorem 1: For the complex-field vector channel model of (1), the mutual information $\mathcal{I}(\mathbf{x};\mathbf{y})$ depends on the precoder $\mathbf{G}$ only through $\mathbf{W} = \mathbf{G}^h\mathbf{H}^h\mathbf{H}\mathbf{G}$. The partial derivative of $\mathcal{I}(\mathbf{x};\mathbf{y})$ with respect to $\mathbf{W}$ is

$$\frac{\partial}{\partial \mathbf{W}^*}\mathcal{I}(\mathbf{x};\mathbf{y}) = \mathbf{\Phi}. \tag{15}$$

The composite Hessian of $\mathcal{I}(\mathbf{x};\mathbf{y})$ with respect to $\mathbf{W}$ is

$$\begin{aligned}
&\mathcal{CH}_{\mathbf{W}}\mathcal{I}(\mathbf{x};\mathbf{y})\\
&\triangleq \begin{bmatrix} \mathcal{H}_{\mathbf{W},\mathbf{W}^*}\mathcal{I}(\mathbf{x};\mathbf{y}) & \mathcal{H}_{\mathbf{W}^*,\mathbf{W}^*}\mathcal{I}(\mathbf{x};\mathbf{y}) \\ \mathcal{H}_{\mathbf{W},\mathbf{W}}\mathcal{I}(\mathbf{x};\mathbf{y}) & \mathcal{H}_{\mathbf{W}^*,\mathbf{W}}\mathcal{I}(\mathbf{x};\mathbf{y}) \end{bmatrix}\\
&= \frac{-1}{\sigma^2}E\left\{\begin{bmatrix} \mathbf{\Phi}^*_{\mathbf{x}\mathbf{x}^h}(\mathbf{y})\otimes\mathbf{\Phi}_{\mathbf{x}\mathbf{x}^h}(\mathbf{y}) & \mathbf{\Psi}^*_{\mathbf{x}\mathbf{x}^t}(\mathbf{y})\otimes\mathbf{\Psi}_{\mathbf{x}\mathbf{x}^t}(\mathbf{y}) \\ \mathbf{\Psi}_{\mathbf{x}\mathbf{x}^t}(\mathbf{y})\otimes\mathbf{\Psi}^*_{\mathbf{x}\mathbf{x}^t}(\mathbf{y}) & \mathbf{\Phi}_{\mathbf{x}\mathbf{x}^h}(\mathbf{y})\otimes\mathbf{\Phi}^*_{\mathbf{x}\mathbf{x}^h}(\mathbf{y}) \end{bmatrix}\right\}\\
&= \frac{-1}{\sigma^2}E\left\{\begin{bmatrix} E\left\{\left[(\mathbf{x}-E\{\mathbf{x}|\mathbf{y}\})^*\otimes(\mathbf{x}-E\{\mathbf{x}|\mathbf{y}\})\right]|\mathbf{y}\right\} \\ E\left\{\left[(\mathbf{x}-E\{\mathbf{x}|\mathbf{y}\})\otimes(\mathbf{x}-E\{\mathbf{x}|\mathbf{y}\})^*\right]|\mathbf{y}\right\} \end{bmatrix}\right.\\
&\qquad\left.\times\begin{bmatrix} E\left\{\left[(\mathbf{x}-E\{\mathbf{x}|\mathbf{y}\})^*\otimes(\mathbf{x}-E\{\mathbf{x}|\mathbf{y}\})\right]|\mathbf{y}\right\} \\ E\left\{\left[(\mathbf{x}-E\{\mathbf{x}|\mathbf{y}\})\otimes(\mathbf{x}-E\{\mathbf{x}|\mathbf{y}\})^*\right]|\mathbf{y}\right\} \end{bmatrix}^h\right\}
\end{aligned} \tag{16}$$

and is negative (semi)-definite. Therefore, $\mathcal{I}(\mathbf{x};\mathbf{y})$ is a concave function of $\mathbf{W}$. □

Note that $\mathbf{\Phi}_{\mathbf{x}\mathbf{x}^h}(\mathbf{y})$ and $\mathbf{\Psi}_{\mathbf{x}\mathbf{x}^t}(\mathbf{y})$ are, respectively, MMSE and companion MMSE matrices conditioned on a specific observation of output $\mathbf{y}$

$$\mathbf{\Phi}_{\mathbf{x}\mathbf{x}^h}(\mathbf{y}) = E\left\{[\mathbf{x}-E\{\mathbf{x}|\mathbf{y}\}][\mathbf{x}-E\{\mathbf{x}|\mathbf{y}\}]^h|\mathbf{y}\right\} \tag{17}$$

$$\mathbf{\Psi}_{\mathbf{x}\mathbf{x}^t}(\mathbf{y}) = E\left\{[\mathbf{x}-E\{\mathbf{x}|\mathbf{y}\}][\mathbf{x}-E\{\mathbf{x}|\mathbf{y}\}]^t|\mathbf{y}\right\}. \tag{18}$$

Thus, $E\{\mathbf{\Phi}_{\mathbf{x}\mathbf{x}^h}(\mathbf{y})\} = \mathbf{\Phi}$ and $E\{\mathbf{\Psi}_{\mathbf{x}\mathbf{x}^t}(\mathbf{y})\} = \mathbf{\Psi}$.

Theorem 2: For the complex vector Gaussian channel model (1), if $\mathbf{H} = \mathbf{\Sigma}_H$ and $\mathbf{G} = \mathbf{\Sigma}_G\mathbf{V}_G^h$ with diagonal $\mathbf{\Sigma}_H$ and $\mathbf{\Sigma}_G$, the mutual information $\mathcal{I}(\mathbf{x};\mathbf{y})$ is a concave function of $\mathbf{\Sigma}_G^2$, and the partial derivative and the Hessian of the mutual information with respect to $\mathbf{\Sigma}_G^2$ are given by

$$\frac{\partial\mathcal{I}(\mathbf{x};\mathbf{y})}{\partial\mathbf{\Sigma}_G^2} = \mathrm{diag}\left(\mathbf{V}_G^h\mathbf{\Phi}\mathbf{V}_G\mathbf{\Sigma}_H^2\right) \tag{19}$$

$$\begin{aligned}
\mathcal{H}_{\mathbf{\Sigma}_G^2,\mathbf{\Sigma}_G^2}\mathcal{I}(\mathbf{x};\mathbf{y}) &= -\frac{1}{\sigma^2}\mathbf{\Sigma}_H^2E\left\{\mathbf{\Phi}_{\mathbf{g}\mathbf{g}^h}(\mathbf{y})\odot\mathbf{\Phi}^*_{\mathbf{g}\mathbf{g}^h}(\mathbf{y})\right\}\mathbf{\Sigma}_H^2\\
&\quad -\frac{1}{\sigma^2}\mathbf{\Sigma}_H^2E\left\{\mathbf{\Phi}_{\mathbf{g}\mathbf{g}^t}(\mathbf{y})\odot\mathbf{\Phi}^*_{\mathbf{g}\mathbf{g}^t}(\mathbf{y})\right\}\mathbf{\Sigma}_H^2\\
&= \frac{-1}{\sigma^2}\mathbf{\Sigma}_H^2\left[E\{\mathbf{Z}\mathbf{Z}^h\}+E\{\mathbf{Z}\mathbf{Z}^t\}\right]\mathbf{\Sigma}_H^2
\end{aligned} \tag{20}$$

where $\mathbf{g} = \mathbf{V}_G^h\mathbf{x}$ and

$$\mathbf{\Phi}_{\mathbf{g}\mathbf{g}^h}(\mathbf{y}) = \mathbf{V}_G^h\mathbf{\Phi}_{\mathbf{x}\mathbf{x}^h}(\mathbf{y})\mathbf{V}_G \tag{21}$$

$$\mathbf{\Phi}_{\mathbf{g}\mathbf{g}^t}(\mathbf{y}) = \mathbf{V}_G^h\mathbf{\Phi}_{\mathbf{x}\mathbf{x}^t}(\mathbf{y})\mathbf{V}_G^* \tag{22}$$

$$\mathbf{Z} = E\left\{\left[(\mathbf{g}-E\{\mathbf{g}|\mathbf{y}\})\odot(\mathbf{g}-E\{\mathbf{g}|\mathbf{y}\})^*\right]|\mathbf{y}\right\}. \tag{23}$$

□

These two theorems are based on the following seven lemmas.

Lemma 1: The probability density function $p(\mathbf{y})$ satisfies the following first-order derivative equations:

$$\begin{aligned}
\mathcal{D}_{\mathbf{y}}p(\mathbf{y}) &= E_{\mathbf{x}}\{\mathcal{D}_{\mathbf{y}}p(\mathbf{y}|\mathbf{x})\}\\
&= -E_{\mathbf{x}}\left\{p(\mathbf{y}|\mathbf{x})\frac{(\mathbf{y}-\mathbf{H}\mathbf{G}\mathbf{x})^h}{\sigma^2}\right\}
\end{aligned} \tag{24}$$

$$\frac{\partial p(\mathbf{y})}{\partial\mathbf{G}^*} = -\mathbf{H}^hE_{\mathbf{x}}\left\{[\mathcal{D}_{\mathbf{y}}p(\mathbf{y}|\mathbf{x})]^h\mathbf{x}^h\right\} \tag{25}$$

$$\mathcal{D}_{\mathbf{y}^*}p(\mathbf{y}) = [\mathcal{D}_{\mathbf{y}}p(\mathbf{y})]^* = -E_{\mathbf{x}}\left\{p(\mathbf{y}|\mathbf{x})\frac{(\mathbf{y}-\mathbf{H}\mathbf{G}\mathbf{x})^t}{\sigma^2}\right\} \tag{26}$$

$$\frac{\partial p(\mathbf{y})}{\partial\mathbf{G}} = \left[\frac{\partial p(\mathbf{y})}{\partial\mathbf{G}^*}\right]^* = -\mathbf{H}^tE_{\mathbf{x}}\left\{[\mathcal{D}_{\mathbf{y}}p(\mathbf{y}|\mathbf{x})]^t\mathbf{x}^t\right\}. \tag{27}$$

Lemma 2: The conditional mean $E\{\mathbf{x}|\mathbf{y}\}$ satisfies the following equalities:

$$\mathcal{D}_{\mathbf{y}}E\{\mathbf{x}|\mathbf{y}\} = \mathbf{\Phi}_{\mathbf{x}\mathbf{x}^h}(\mathbf{y})\mathbf{G}^h\mathbf{H}^h\frac{1}{\sigma^2} \tag{28}$$

$$\mathcal{D}_{\mathbf{y}^*}E\{\mathbf{x}|\mathbf{y}\} = \mathbf{\Psi}_{\mathbf{x}\mathbf{x}^t}(\mathbf{y})\mathbf{G}^t\mathbf{H}^t\frac{1}{\sigma^2}. \tag{29}$$

Lemma 3: Let x_i and x_j be the ith and jth elements of the transmitted signal vector $\mathbf{x}$, respectively. The conditional expectations, $E\{x_i|\mathbf{y}\}$ and $E\{x_j^*|\mathbf{y}\}$, respectively, satisfy the following:

$$\begin{aligned}
&\frac{\partial E\{x_i|\mathbf{y}\}}{\partial \mathbf{G}^*}\\
&= \frac{\partial}{\partial \mathbf{G}^*} E\left\{x_i \frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right\}\\
&= \frac{1}{p(\mathbf{y})}\mathbf{H}^h\left[-E\left\{x_i\left[\mathcal{D}_{\mathbf{y}}p(\mathbf{y}|\mathbf{x})\right]^h\mathbf{x}^h\right\}\right.\\
&\qquad\left.+E\{x_i|\mathbf{y}\}E_{\mathbf{x}}\left\{\left[\mathcal{D}_{\mathbf{y}}p(\mathbf{y}|\mathbf{x})\right]^h\mathbf{x}^h\right\}\right] \qquad (30)\\
&\frac{\partial E\left\{x_j^*|\mathbf{y}\right\}}{\partial \mathbf{G}^*}\\
&= \frac{\partial}{\partial \mathbf{G}^*} E\left\{x_j^* \frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right\}\\
&= \frac{1}{p(\mathbf{y})}\mathbf{H}^h\left[-E\left\{x_j^*\left[\mathcal{D}_{\mathbf{y}}p(\mathbf{y}|\mathbf{x})\right]^h\mathbf{x}^h\right\}\right.\\
&\qquad\left.+E\left\{x_j^*|\mathbf{y}\right\}E_{\mathbf{x}}\left\{\left[\mathcal{D}_{\mathbf{y}}p(\mathbf{y}|\mathbf{x})\right]^h\mathbf{x}^h\right\}\right]. \qquad (31)
\end{aligned}$$

Lemma 4: The first-order derivative of the mutual information $\mathcal{I}(\mathbf{x};\mathbf{y})$ with respect to $\mathbf{G}^*$ is given by

$$\frac{\partial}{\partial \mathbf{G}^*}\mathcal{I}(\mathbf{x};\mathbf{y}) = \mathbf{H}^h\mathbf{H}\mathbf{G}\left[\frac{\partial}{\partial \mathbf{W}^*}\mathcal{I}(\mathbf{x};\mathbf{y})\right]. \qquad (32)$$

Lemma 5: The following three conditional mean equalities are valid.

$$\begin{aligned}
&E\left\{\left[\mathbf{x}^*\otimes E\{\mathbf{x}|\mathbf{y}\}\right]|\mathbf{y}\right\}\\
&\quad = E\{\mathbf{x}|\mathbf{y}\}^*\otimes E\{\mathbf{x}|\mathbf{y}\} \qquad (33)\\
&E\left\{(\mathbf{x}^*\otimes\mathbf{x})|\mathbf{y}\right\}E\left\{(\mathbf{x}^*\otimes\mathbf{x})|\mathbf{y}\right\}^h\\
&\quad = E\{\mathbf{x}\mathbf{x}^h|\mathbf{y}\}^*\otimes E\{\mathbf{x}\mathbf{x}^h|\mathbf{y}\} \qquad (34)\\
&E\left\{(\mathbf{x}^*\otimes\mathbf{x})|\mathbf{y}\right\}\left[E\{\mathbf{x}|\mathbf{y}\}^*\otimes E\{\mathbf{x}|\mathbf{y}\}\right]^h\\
&\quad = E\{\mathbf{x}\mathbf{x}^h|\mathbf{y}\}^*\otimes\left[E\{\mathbf{x}|\mathbf{y}\}E\{\mathbf{x}|\mathbf{y}\}^h\right]. \qquad (35)
\end{aligned}$$

Lemma 6: The Jacobian matrix of MMSE matrix $\mathbf{\Phi}$ with respect to $\mathbf{G}^*$ satisfies the following equation:

$$\begin{aligned}
\mathcal{D}_{\mathbf{G}^*}\mathbf{\Phi} = &-\frac{1}{\sigma^2}E\left\{\mathbf{\Phi}^*_{\mathbf{x}\mathbf{x}^h}(\mathbf{y})\otimes\mathbf{\Phi}_{\mathbf{x}\mathbf{x}^h}(\mathbf{y})\right\}\mathbf{K}_{N_t}\\
&\times\left[\mathbf{I}_{N_t}\otimes(\mathbf{G}^t\mathbf{H}^t\mathbf{H}^*)\right]-\frac{1}{\sigma^2}E\left\{\mathbf{\Psi}^*_{\mathbf{x}\mathbf{x}^t}(\mathbf{y})\otimes\mathbf{\Psi}_{\mathbf{x}\mathbf{x}^t}(\mathbf{y})\right\}\\
&\times\left[\mathbf{I}_{N_t}\otimes(\mathbf{G}^t\mathbf{H}^t\mathbf{H}^*)\right] \qquad (36)
\end{aligned}$$

where $\mathbf{K}_{N_t}$ is an $N_t^2\times N_t^2$ commutation matrix [27], [28].

Lemma 7: The second-order derivatives of the mutual information $\mathcal{I}(\mathbf{x};\mathbf{y})$ contains four key complex Hessian matrices, and they are given by

$$\mathcal{H}_{\mathbf{W},\mathbf{W}^*}\mathcal{I}(\mathbf{x};\mathbf{y}) = -\frac{1}{\sigma^2}E\left\{\mathbf{\Phi}^*_{\mathbf{x}\mathbf{x}^h}(\mathbf{y})\otimes\mathbf{\Phi}_{\mathbf{x}\mathbf{x}^h}(\mathbf{y})\right\} \qquad (37)$$

$$\mathcal{H}_{\mathbf{W}^*,\mathbf{W}^*}\mathcal{I}(\mathbf{x};\mathbf{y}) = -\frac{1}{\sigma^2}E\left\{\mathbf{\Psi}^*_{\mathbf{x}\mathbf{x}^t}(\mathbf{y})\otimes\mathbf{\Psi}_{\mathbf{x}\mathbf{x}^t}(\mathbf{y})\right\} \qquad (38)$$

$$\mathcal{H}_{\mathbf{W}^*,\mathbf{W}}\mathcal{I}(\mathbf{x};\mathbf{y}) = -\frac{1}{\sigma^2}E\left\{\mathbf{\Phi}_{\mathbf{x}\mathbf{x}^h}(\mathbf{y})\otimes\mathbf{\Phi}^*_{\mathbf{x}\mathbf{x}^h}(\mathbf{y})\right\} \qquad (39)$$

$$\mathcal{H}_{\mathbf{W},\mathbf{W}}\mathcal{I}(\mathbf{x};\mathbf{y}) = -\frac{1}{\sigma^2}E\left\{\mathbf{\Psi}_{\mathbf{x}\mathbf{x}^t}(\mathbf{y})\otimes\mathbf{\Psi}^*_{\mathbf{x}\mathbf{x}^t}(\mathbf{y})\right\}. \qquad (40)$$

B. Remarks

We conclude this section with two remarks. Our first remark is to point out that it was first stated in [24] that the mutual information is a concave function of $\mathbf{G}^t\mathbf{H}^t\mathbf{H}\mathbf{G}$ for cases where the channel matrix, linear precoder, transmitted signal and received noise are all in *real field*. However, the proof of this result provided in [24] is incomplete, since it claimed $E\{\mathbf{\Phi}(\mathbf{y})\otimes\mathbf{\Phi}(\mathbf{y})\}$ to be positive semi-definite without a proof, where $\mathbf{\Phi}(\mathbf{y})$ is the MMSE matrix conditioned on a specific realization of the output $\mathbf{y}$. In this section, we not only extended the concave property to *complex field*, but also provided a rigorous proof, which is nontrivial as evidenced. Theorem 2 is also a nontrivial extension to the complex field systems from [28, Theorem 5] (or [23, Lemma 2]) which was originally derived for real field systems.

Our second remark is on the globally optimal solution for linear precoders. Having proven the concavity of the merit function (mutual information) for maximization, we are keen to develop ways to find globally optimal linear precoders. Unfortunately, existing algorithms, e.g., [29], cannot be straightforwardly used for finding a globally optimal $\mathbf{G}$. This is because even though $\mathcal{I}(\mathbf{x};\mathbf{y})$ is a concave function of $\mathbf{W}$, the power constraint $\mathrm{trace}(\mathbf{G}\mathbf{G}^h)=\mathcal{P}$ on our precoding parameter matrix $\mathbf{G}$ makes it a very difficult optimization problem.

More specifically, we take gradient ascent algorithm as an example. We may update

$$\mathbf{W}_{k+1} = \mathbf{W}_k + \delta\mathbf{W} \qquad (41)$$

with $\delta\mathbf{W} = \mu\frac{\partial}{\partial\mathbf{W}^*}\mathcal{I}(\mathbf{x};\mathbf{y}) = \mu\mathbf{\Phi}$, in which μ is a sufficiently small positive step size. Our iterative solution requires us to update

$$\mathbf{G}_{k+1} = \mathbf{G}_k + \delta\mathbf{G} \qquad (42)$$

where the linear precoder has to satisfy the power constraint:

$$\mathrm{trace}\left(\mathbf{G}_{k+1}\mathbf{G}_{k+1}^h\right) = \mathrm{trace}\left(\mathbf{G}_k\mathbf{G}_k^h\right) = \mathcal{P}.$$

Given the power constraint, directly solving for the incremental update $\delta\mathbf{G}$ can be difficult. Moreover, the convergence can also be painfully slow because the step size μ has to be very small to avoid divergence.

We present a novel parametric algorithm for iterative solution of the optimum precoder $\mathbf{G}$ in the next section.

IV. Parameterized Iterative Algorithm

In this section, we propose a parameterized iterative algorithm for solving a linear precoder $\mathbf{G}$, which globally maximizes the mutual information $\mathcal{I}(\mathbf{x};\mathbf{y})$ given by (7), under the assumption that the channel state information is known to the transmitter and the receiver.

A. Parameterization

We first present the following proposition as a direct extension of [23, Proposition 1], which originally dealt with real-valued systems.

Proposition 2: The left singular vectors of the globally optimal precoder $\mathbf{G}\in\mathbb{C}^{N_t\times N_t}$ can always be chosen to be the same as the right singular vectors of channel matrix $\mathbf{H}$. $\diamond$

According to Propositions 1 and 2, the complex vector Gaussian channel (1) can be simplified to the following equivalent model

$$\overline{\mathbf{y}} = \mathbf{\Sigma}_H \mathbf{\Sigma}_G \mathbf{\Theta} \mathbf{x} + \mathbf{v} \tag{43}$$

where $\mathbf{\Sigma}_H$ and $\mathbf{\Sigma}_G$ are diagonal matrices containing the singular values of $\mathbf{H}$ and $\mathbf{G}$, respectively, $\mathbf{\Theta}$ is a unitary matrix, and $\bar{\mathbf{y}} = \mathbf{U}_H^h \mathbf{y}$ in which $\mathbf{U}_H$ is the left singular vectors of $\mathbf{H}$. We note that maximizing $\mathcal{I}(\mathbf{x}; \overline{\mathbf{y}})$ based on the model (43) is equivalent to maximizing $\mathcal{I}(\mathbf{x}; \mathbf{y})$ according to the model of (1).

From [4] and [30], any $N_t \times N_t$ unitary matrix can be written as

$$\mathbf{\Theta} = \mathbf{D} \prod_{p=N_t-1}^{1} \prod_{q=p+1}^{N_t} \mathbf{U}_{pq}(\omega_{pq}, \nu_{pq}) \tag{44}$$

where $\mathbf{D}$ is an $N_t \times N_t$ unitary diagonal matrix, the angles $\omega_{pq} \in (-\pi, \pi]$ and $\nu_{pq} \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ parameterize the complex Givens matrix $\mathbf{U}_{pq}(\omega_{pq}, \nu_{pq})$, formed by replacing the (p,p)th, (q,q)th, (p,q)th, and (q,p)th entries of identity matrix $\mathbf{I}_{N_t}$ with $\cos\omega_{pq}$, $\cos\omega_{pq}$, $\sin\omega_{pq} e^{-j\nu_{pq}}$, and $-\sin\omega_{pq} e^{j\nu_{pq}}$, respectively.

For the equivalent model (43), we have $\mathbf{W} = \mathbf{\Theta}^h \mathbf{\Sigma}_H^2 \mathbf{\Sigma}_G^2 \mathbf{\Theta}$. Using gradient ascent, if we choose a step size μ to obtain the incremental $\delta\mathbf{W} = \mu \frac{\partial}{\partial \mathbf{W}^*} \mathcal{I}(\mathbf{x}; \mathbf{y}) = \mu \mathbf{\Phi}$, we can find the incremental $\delta\omega_{pq}$ and $\delta\nu_{pq}$ via the following first-order approximation

$$\delta\mathbf{W} \simeq [\delta\mathbf{\Theta}]^h \mathbf{\Sigma}_H^2 \mathbf{\Sigma}_G^2 \mathbf{\Theta} + \mathbf{\Theta}^h \mathbf{\Sigma}_H^2 \mathbf{\Sigma}_G^2 [\delta\mathbf{\Theta}] \tag{45}$$

with

$$\delta\mathbf{\Theta} = \sum_{p=N_t-1}^{1} \sum_{q=p+1}^{N_t} \frac{\partial \mathbf{\Theta}}{\partial \omega_{pq}} \delta\omega_{pq} + \sum_{p=N_t-1}^{1} \sum_{q=p+1}^{N_t} \frac{\partial \mathbf{\Theta}}{\partial \nu_{pq}} \delta\nu_{pq}. \tag{46}$$

After finding increments $\delta\omega_{pq}$ and $\delta\nu_{pq}$, we can update unitary matrix $\mathbf{\Theta}$ and linear precoder $\mathbf{G}$. Note that the newly updated precoder $\mathbf{G}_{k+1}$ will automatically satisfy the power constraint as long as the singular values $\mathbf{\Sigma}_G$ remain unchanged, owing to the parameterization of unitary $\mathbf{\Theta}$ in (44).

B. Iterative Optimization Algorithm

We are now ready to summarize our algorithm steps.

Step 1: For a given channel $\mathbf{H}$, apply singular value decomposition (SVD) to decompose $\mathbf{H} = \mathbf{U}_H \mathbf{\Sigma}_H \mathbf{V}_H$ and convert the original channel model (1) into its equivalent model (43), where $\mathbf{x}$ is drawn from a pre-chosen finite alphabet set (constellation).

Step 2: Choose an initial set of non-zero values for ω_{pq}, ν_{pq} and non-negative diagonal matrix $\mathbf{\Sigma}_G$ with $\mathrm{trace}(\mathbf{\Sigma}_G^2) = N_t$. Choose a unitary diagonal matrix $\mathbf{D}$, which does not affect the maximization of mutual information $\mathcal{I}(\mathbf{x}; \overline{\mathbf{y}})$.

Step 3: Compute mutual information $\mathcal{I}(\mathbf{x}; \overline{\mathbf{y}})$ and MMSE matrix $\mathbf{\Phi}$. Use backtracking line search [29] to determine a step size μ_Θ to obtain the incremental $\delta\mathbf{W}$ before solving for $\delta\omega_{pq}$ and $\delta\nu_{pq}$ with fixed $\mathbf{\Sigma}_G$. Then update unitary matrix $\mathbf{\Theta}$.

Step 4: Update (or re-compute) mutual information $\mathcal{I}(\mathbf{x}; \overline{\mathbf{y}})$ and MMSE matrix $\mathbf{\Phi}$ according to the updated matrix $\mathbf{\Theta}$.

Step 5: Calculate the gradient for $\mathbf{\Sigma}_G^2$ as follows:

$$\mathbf{D}_G = \mathrm{diag}\left(\mathbf{\Sigma}_H^2 \mathbf{\Theta} \mathbf{\Phi} \mathbf{\Theta}^h\right) - \lambda \mathbf{I}_{N_t} \tag{47}$$

with $\lambda = \frac{\mathrm{trace}[\mathrm{diag}(\mathbf{\Sigma}_H^2 \mathbf{\Theta}\mathbf{\Phi}\mathbf{\Theta}^h)]}{N_t}$.

Step 6: Update $\mathbf{\Sigma}_G^2$ by $\mathbf{\Sigma}_G^2 + \mu_\Sigma \mathbf{D}_G$ with μ_Σ being a step size, which is determined by the backtracking line search. Set any negative diagonal entry of $\mathbf{\Sigma}_G^2$ to zero, before normalizing the updated non-negative $\mathbf{\Sigma}_G^2$ to satisfy the power constraint: $\mathrm{trace}(\mathbf{\Sigma}_G^2) = N_t$.

Step 7: If the step size μ_Σ in Step 6 has to be (nearly) zero, then update $\mathbf{\Sigma}_G^2$ by $\mathbf{\Sigma}_G^2 - \mu_\Sigma \mathbf{D}_G$ with μ_Σ being determined by the backtracking line search. Set any negative diagonal entry of $\mathbf{\Sigma}_G^2$ to zero, then normalize the updated non-negative $\mathbf{\Sigma}_G^2$ to satisfy the power constraint: $\mathrm{trace}(\mathbf{\Sigma}_G^2) = N_t$.

Repeat Step 3 through Step 7 until convergence or until a pre-set target is reached. Then a globally optimal precoder $\mathbf{G}^{\mathrm{opt}} = \mathbf{V}_H \mathbf{\Sigma}_G^{\mathrm{opt}} \mathbf{\Theta}^{\mathrm{opt}}$ is obtained. ◇

C. Remarks

We conclude this section with several remarks:

1) Based on the convergence result for block coordinate ascent/decent method [31, p. 273], the global convergence is guaranteed with properly chosen step size for the above iterative algorithm when SNR is finite and the initial unitary matrix $\mathbf{\Theta}$ is indeed complex-valued. This is because both Hessian matrices given by Theorems 1 and 2 will be negative definite, the mutual information is a strict concave function of $\mathbf{W}$, and a strict concave function of $\mathbf{\Sigma}_G^2$ for a fixed $\mathbf{\Theta}$, with SNR being finite.
2) Although there is only one global maximum of the mutual information $\mathcal{I}(\mathbf{x}; \overline{\mathbf{y}})$, which is equivalent to $\mathcal{I}(\mathbf{x}; \mathbf{y})$, it is possible to have non-unique globally optimal solutions in terms of $\mathbf{W}^{\mathrm{opt}}$, which give the same globally maximum mutual information. For example, a finite alphabet constellation such as PSK and QAM typically has certain symmetry (with respect to X axis and/or Y axis), allowing the same mutual information with various rotations.
3) Similarly, there exist many "equivalent" globally optimal linear precoders $\mathbf{G}^{\mathrm{opt}}$ corresponding to an optimal $\mathbf{W}^{\mathrm{opt}}$. For example, if we choose different unitary diagonal matrix $\mathbf{D}$ in (44), we will have different linear precoders $\mathbf{G}^{\mathrm{opt}}$.
4) The step sizes μ_Θ and μ_Σ are usually different for fast convergence purposes.
5) The gradient $\mathbf{D}_G$ in Step 5 contains power constraint information: if the stepsizes are large in previous iterations, Step 6 might get trapped in suboptimal power allocation points; Hence, Step 7 is introduced to guarantee the globally optimal power allocation under power constraints.

TABLE I
THE SMALLEST INTERVALS FOR THE ANGLES ω AND ν CORRESPONDING TO DIFFERENT MODULATIONS

	BPSK	QPSK	8PSK	16QAM	64QAM	256QAM
ω	$[0, \frac{\pi}{4}]$	$[0, \frac{\pi}{4}]$	$[0, \frac{\pi}{4}]$	$[0, \frac{\pi}{4}]$	$[0, \frac{\pi}{4}]$	$[0, \frac{\pi}{4}]$
ν	$[0, \frac{\pi}{2}]$	$[0, \frac{\pi}{4}]$	$[0, \frac{\pi}{8}]$	$[0, \frac{\pi}{4}]$	$[0, \frac{\pi}{4}]$	$[0, \frac{\pi}{4}]$

6) If we force $\lambda = 0$ in Step 5, then Step 6 may not get stuck to suboptimal power allocation points, in which case Step 7 is not needed. However, the global convergence will be much slower in general.
7) In principle, we can also employ Newton's method to develop an iterative algorithm, because we have obtained mathematical formulae for both the gradient and the Hessian matrix of mutual information $\mathcal{I}(\mathbf{x};\overline{\mathbf{y}})$ with respect to $\mathbf{W}$ (and $\mathbf{\Sigma}_G^2$ for a fixed $\mathbf{\Theta}$). However, computing the Hessian matrices numerically can be more demanding in terms of computations and memories, unlike computing the gradients. Thus, Newton's method may be less efficient.
8) If we only consider real-valued vector channels with real-valued linear precoders, signals and noise, which are the cases previously discussed in [23] and [24], then we need to set $\nu_{pq} = 0$ for all p and q, and the unitary matrix $\mathbf{\Theta}$ becomes a routine rotational matrix [30]. In this special case, our proposed iterative algorithm will still work with fast and robust convergence. Moreover, we should point out that in this case, all the four complex-valued Hessian matrices in Lemma 7 reduce to one single real-valued Hessian matrix, and the composite Hessian matrix in Theorem 1 needs to be re-defined as the regular Hessian matrix given by Lemma 7.

V. SIMULATION RESULTS

We now provide examples to illustrate the convergence properties of our proposed iterative optimization algorithm. Our examples can demonstrate the performance advantages of employing the proposed optimal linear precoders that maximize the I/O mutual information for complex vector Gaussian channels with finite discrete inputs.

A. Example 1

Consider a 2×2 static (non-fading) MIMO baseband system (1) with the channel matrix

$$\mathbf{H} = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$$

which was also used in [32]. The average SNR of a MIMO system $\mathbf{H}$ is given by $\mathrm{SNR} = \frac{\mathrm{trace}(\mathbf{H}\mathbf{H}^h)}{(N_t\sigma^2)}$.

Taking into account the symmetric property of commonly used constellations, it is proved in Appendix B that the intervals for the angles ω and ν of the unitary matrix $\mathbf{\Theta}$ can be reduced to much smaller intervals than those that followed (44). Table I lists the smallest intervals for ω and ν for 2×2 MIMO channels with different modulations.

We first consider BPSK inputs with $\mathrm{SNR} = 0\,\mathrm{dB}$. Since both the channel matrix and channel inputs are real-valued quantities, for comparison purpose, we maximize the mutual information by employing real-valued linear precoder and complex-valued linear precoder, respectively. The resulting precoders and corresponding mutual information values are listed in Table II, which shows that an optimal complex-valued precoder can provide much higher mutual information than an optimal real-valued precoder.

TABLE II
REAL-VALUED AND COMPLEX-VALUED SOLUTIONS

BPSK, SNR$=0$ dB	Optimal Precoder $\mathbf{G}$	Maximum $\mathcal{I}(\mathbf{x};\mathbf{y})$
Real-valued solution	$\begin{bmatrix} 1.0493 & 0.5884 \\ 0.6485 & 0.3637 \end{bmatrix}$	1.4639
Complex-valued solution	$\begin{bmatrix} 0.8507 & -0.8507j \\ 0.5257 & -0.5257j \end{bmatrix}$	1.8170

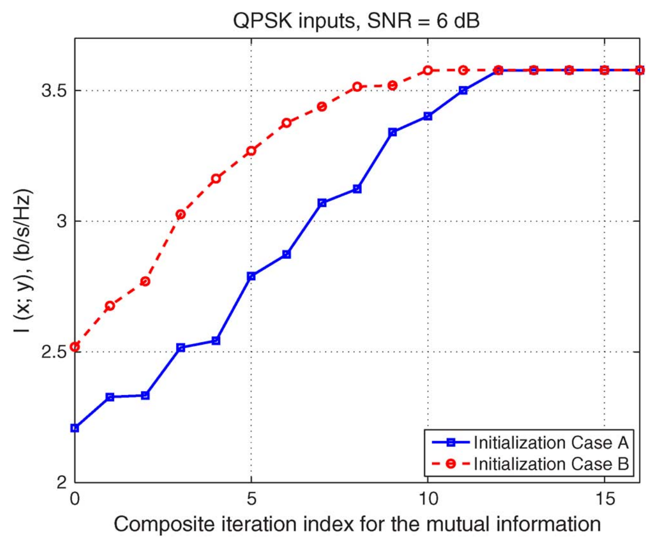


Fig. 1. Convergence trajectories for the mutual information.

We now test our iterative algorithm for QPSK inputs with $\mathrm{SNR} = 6$ dB. We chose two different initializations: Case A: $\mathbf{\Sigma}_G = \mathbf{I}_2$, $\omega = \frac{\pi}{100}$ and $\nu = \frac{\pi}{10}$; and Case B: $\mathbf{\Sigma}_G = \mathbf{I}_2$, $\omega = \frac{\pi}{15}$ and $\nu = \frac{\pi}{15}$. The convergence trajectories of the resulting mutual information is shown in Fig. 1.

As we can see, under either initialization, the algorithm quickly converged to the global maximum of the mutual information, $\mathcal{I}(\mathbf{x};\mathbf{y}) = 3.5785$ bits/s/Hz. Within the intervals $\omega \in [0, \frac{\pi}{4}]$ and $\nu \in [0, \frac{\pi}{4}]$, the global optimal solution $\mathbf{W}$ is

$$\mathbf{W}_1^{\mathrm{opt}} = \begin{bmatrix} 3.0126 & 1.5855 - 0.4577j \\ 1.5855 + 0.4577j & 0.9040 \end{bmatrix}$$

and a corresponding globally optimal linear precoder is

$$\mathbf{G}_1^{\mathrm{opt}} = \begin{bmatrix} 1.0551 & 0.5553 - 0.1603j \\ 0.6521 & 0.3432 - 0.0991j \end{bmatrix}.$$

For different intervals such as $\omega \in [\frac{\pi}{4}, \frac{\pi}{2}]$ and $\nu \in [\frac{\pi}{4}, \frac{\pi}{2}]$, the global optimal solution $\mathbf{W}$ is given by

$$\mathbf{W}_2^{\mathrm{opt}} = \begin{bmatrix} 0.9040 & 0.4577 - 1.5855j \\ 0.4577 + 1.5855j & 3.0126 \end{bmatrix}$$

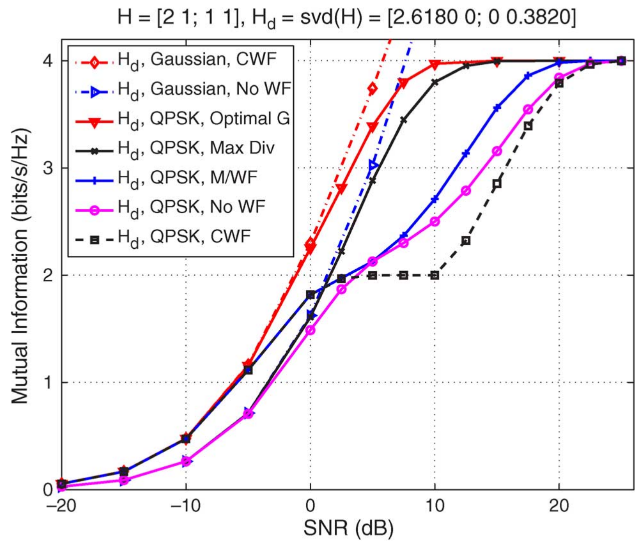


Fig. 2. Mutual information of the 2 × 2 diagonal channel matrix with Gaussian and QPSK inputs.

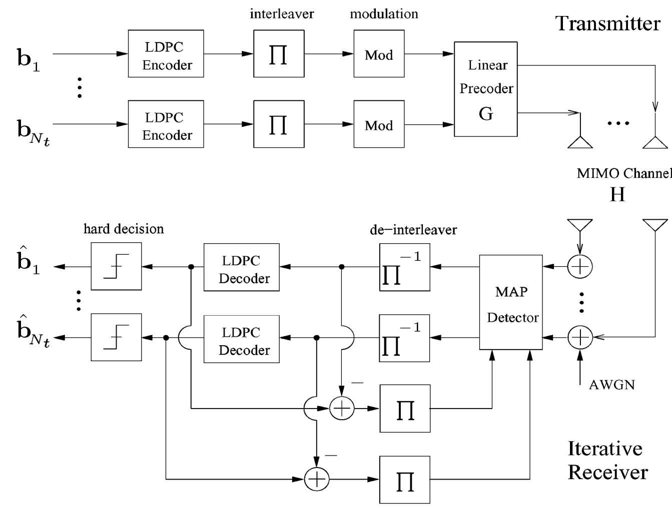


Fig. 3. MIMO transceiver with a linear precoder $\mathbf{G}$.

and a corresponding globally optimal linear precoder is

$$\mathbf{G}_2^{\mathrm{opt}} = \begin{bmatrix} 0.5780 & 0.2926 - 1.0137j \\ 0.3572 & 0.1809 - 0.6265j \end{bmatrix}.$$

Both $\mathbf{G}_1^{\mathrm{opt}}$ and $\mathbf{G}_2^{\mathrm{opt}}$ lead to the same global maximum mutual information.

It is noted that due to the periodicity of ω and ν, there are 16 "different" solutions for $\mathbf{W}$ in a 2 × 2 MIMO system with QPSK modulated signals. However, once given one of these 16 solutions, the remaining 15 solutions can be directly obtained from the first found solution using the periodicity property. It is further noted that for each optimal $\mathbf{W}^{\mathrm{opt}}$, there are many corresponding globally optimal linear precoders $\mathbf{G}^{\mathrm{opt}}$, which have the following structure:

$$\mathbf{G}^{\mathrm{opt}} = \mathbf{V}_H \mathbf{D} \boldsymbol{\Sigma}_G^{\mathrm{opt}} \boldsymbol{\Theta}^{\mathrm{opt}}$$

where $\mathbf{D}$ is any unitary diagonal matrix with appropriate dimensions. All these optimal precoders lead to the same global maximum mutual information.

For convenient comparison with other existing precoding or power allocation algorithms, we diagonalize the original channel matrix $\mathbf{H}$ via singular value decomposition (SVD): $\mathbf{H}_d = \begin{bmatrix} 2.6180 & 0 \\ 0 & 0.3820 \end{bmatrix}$. We apply, at the transmitter, classic waterfilling [25], mercury waterfilling [20], maximum diversity precoding [4], and our proposed optimal precoding for QPSK inputs. The resulting values of I/O mutual information as functions of the average SNR are comparatively shown in Fig. 2. As benchmarks, we also plot results for Gaussian inputs with classic waterfilling and without waterfilling.

From Fig. 2, we have the following observations:

1) Compared to the system without waterfilling, the classic waterfilling has positive gain on mutual information for low SNR (corresponding to channel coding rate under 0.5) but shows a negative gain (i.e., *loss*) at high SNR (corresponding to channel coding rate above 0.5), this is because the classic waterfilling algorithm allocates all the power to the stronger channel for $\mathrm{SNR} \le 10\ \mathrm{dB}$.
2) The mercury-waterfilling achieves up to 3-dB gain in terms of mutual information over the diagonal channel matrix without waterfilling throughout SNR test range.
3) Maximum diversity precoding shows substantial gain over classic waterfilling and mercury waterfilling when the channel coding rate is above 0.5. However, it is inferior to classic waterfilling and mercury waterfilling when the coding rate falls below 0.5.
4) Throughout the tested SNR range, our proposed globally optimal precoding is always better than maximum diversity precoding, classic waterfilling, and mercury waterfilling. It achieves a very large gain when channel coding rate is high.
5) Our proposed globally optimal precoding can achieve mutual information very close to channel capacity (under Gaussian inputs with waterfilling) when coding rate is below 0.75; Moreover, it outperforms Gaussian inputs without waterfilling when coding rate is below 0.9.

To further illustrate the benefit of globally maximizing the mutual information $\mathcal{I}(\mathbf{x}; \overline{\mathbf{y}})$, we test an MIMO system (1) with the transceiver structure depicted in Fig. 3.

As a more comprehensive system, we adopt the LDPC encoder and decoder simulation package [33], for coding rate $r = 0.75$ and code length $L = 2400$ bits, we obtained the bit error rate (BER) results shown in Fig. 4 for the precoding schemes discussed in Fig. 2. Apparently, the globally optimal precoder we obtained outperforms all the comparison precoding (or power allocation) schemes. In fact, the performance gains over other schemes at $\mathrm{BER} = 10^{-4}$ match the value predicted by the mutual information results of Fig. 2 when $\mathcal{I}(\mathbf{x}; \overline{\mathbf{y}}) = 3$. This observation suggests that maximizing mutual information is a very sound approach and that it has direct impact on coded BER performance. It is noted that the gap between the QPSK limit and the optimal precoder can be narrowed by using a longer length code. Moreover, the QPSK limit is only 0.7 dB off the Gaussian-input capacity limit.

Before concluding this example, we would like to point out that the linear precoder obtained in [32] is a sub-optimal

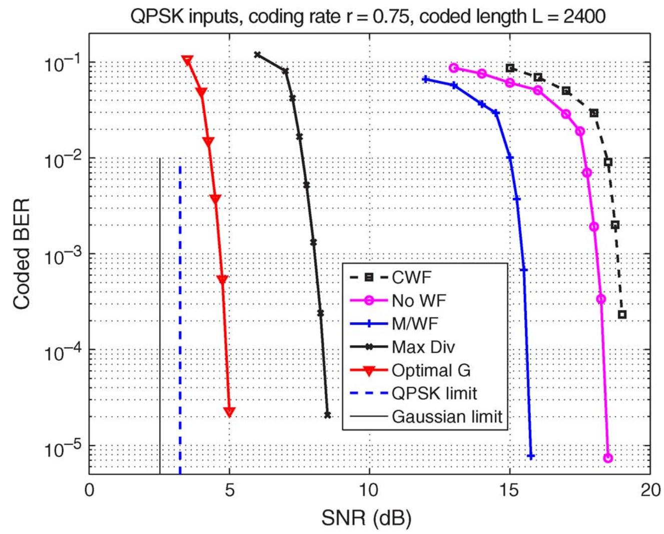


Fig. 4. Coded BER for QPSK modulation with different linear precoding (or power allocation) schemes.

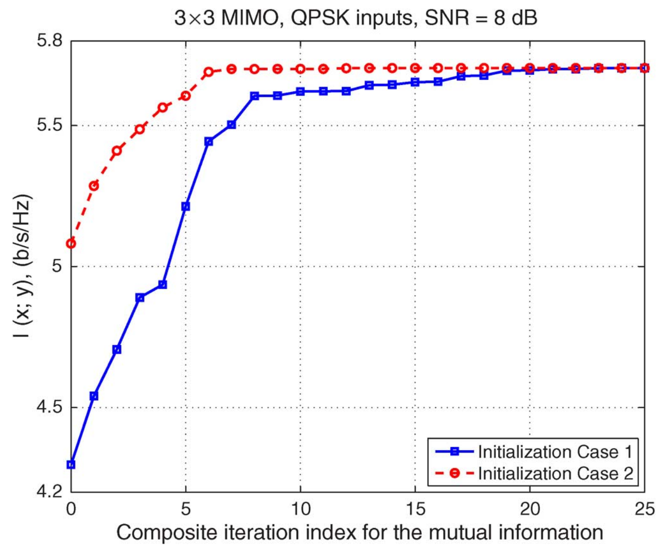


Fig. 5. Convergence trajectory for the mutual information.

solution. Its corresponding mutual information stayed below 3.6 b/s/Hz after 3000 iterations, while our proposed algorithm converged to the maximum mutual information being 3.9996 b/s/Hz within 15 iterations, at the same SNR as tested in [32].

B. Example 2

Next, we consider a 3×3 complex-valued MIMO channel

$$\mathbf{H} = \begin{bmatrix} 1 & 0.5j & 0.3 \\ -0.5j & 1.5 & -0.1j \\ 0.3 & 0.1j & 0.5 \end{bmatrix}. \tag{48}$$

A QPSK signal was transmitted at $\mathrm{SNR} = 8$ dB. We randomly picked two sets of initializations: Case 1: $\Sigma_G = \mathbf{I}_3$, $\omega_{pq} = \frac{\pi}{100}$, and $\nu_{pq} = \frac{\pi}{100}$ for all p and q; Case 2: $\Sigma_G = \mathbf{I}_3$, $\omega_{pq} = \frac{\pi}{10}$, and $\nu_{pq} = \frac{\pi}{10}$ for all p and q. The converging trajectories of the resulting mutual information $\mathcal{I}(\mathbf{x};\mathbf{y})$ are given in Fig. 5.

We observe that the mutual information grows monotonically and converges within 22 iterations and 12 iterations for Case 1 and Case 2, respectively. The global maximum mutual information is 5.7370 b/s/Hz. Upon convergence, our optimized matrix $\mathbf{W}$ is given by

$$\mathbf{W}^{\mathrm{opt}} = \begin{bmatrix} 3.0568 & 1.0142 - 0.3936j & 1.5399 - 0.4367j \\ 1.0142 + 0.3936j & 0.9758 & 0.5756 - 0.1145j \\ 1.5399 + 0.4367j & 0.5756 + 0.1145j & 0.8861 \end{bmatrix}$$

and our optimal linear precoder is given by the matrix shown at the bottom of the page. It is noted that there are other optimal solutions for matrix $\mathbf{W}$ and linear precoder $\mathbf{G}$ which lead to the same global maximum mutual information. Details are omitted here for brevity.

We also note that the above $\mathbf{G}^{\mathrm{opt}}$ satisfies the necessary optimality condition[21, (7)] and [22, (4)]. Thus, $\mathbf{G}^{\mathrm{opt}}$ is a solution to $\mathbf{W}^{\mathrm{opt}}$ as $\mathcal{I}(\mathbf{x};\mathbf{y})$ is a concave function of $\mathbf{W}$. Therefore, $\mathbf{G}^{\mathrm{opt}}$ is indeed a globally optimal precoder.

Finally, we would like to point out that we have also designed an optimal precoder for the 4×4 MIMO example used in [24] with the same SNR and digital modulation as used in [24]. Our global maximum mutual information reaches 3.9998 b/s/Hz, above the achieved rate (3.8 b/s/Hz) shown in Fig. 1 of [24]. In terms of speed, our algorithm converged to the global maximum mutual information within 30 iterations while the algorithm in [24] apparently required more than 150 iterations to converge to a lower data rate.

VI. Conclusion

In this work, we investigated the design of optimum precoders aimed at maximizing the I/O mutual information of complex-valued MIMO vector channels, driven by non-Gaussian inputs of finite alphabet. We first proved that the I/O mutual information of the linear MIMO system under finite alphabet inputs is a concave function of a composite matrix $\mathbf{W}$ which itself is a quadratic weighted matrix of a linear precoder matrix $\mathbf{G}$. Recognizing that the concavity property itself does not automatically provides an efficient optimization algorithm, we further developed a fast and effective iterative algorithm to optimize the linear precoder to globally maximize the resulting mutual information. Our numerical results demonstrate that the proposed

$$\mathbf{G}^{\mathrm{opt}} = \begin{bmatrix} 0.8366 - 0.0195j & -0.4440 - 0.0762j & 0.4192 + 0.0763j \\ 0.0160 - 0.7126j & -0.1148 - 0.9134j & -0.2866 - 0.3628j \\ 0.3689 - 0.0140j & -0.4083 - 0.0254j & 0.1839 + 0.0915j \end{bmatrix}.$$

algorithm is computationally efficient, numerically robust and globally convergent, while achieving optimum system performance in terms of highest achievable throughput. Test results also point out that some existing methods may be vulnerable to being trapped at suboptimal precoders, due to either numerical issues or multi-modality.

Appendix A
Proofs of Lemmas 1–7 and Theorems 1 and 2

Proof of Lemma 1: Employing $\mathcal{D}_{\mathbf{y}}\|\mathbf{y}-\mathbf{HGx}\|^2=(\mathbf{y}-\mathbf{HGx})^h$, one has $\mathcal{D}_{\mathbf{y}}p(\mathbf{y}|\mathbf{x})=-p(\mathbf{y}|\mathbf{x})\frac{(\mathbf{y}-\mathbf{HGx})^h}{\sigma^2}$. Invoking $p(\mathbf{y})=E_{\mathbf{x}}\{p(\mathbf{y}|\mathbf{x})\}$, one can easily prove (24).

From [26], we know that $\frac{\partial}{\partial \mathbf{G}^*}\|\mathbf{y}-\mathbf{HGx}\|^2=\mathbf{H}^h(\mathbf{y}-\mathbf{HGx})\mathbf{x}^h$, therefore,

$$\begin{aligned}\frac{\partial p(\mathbf{y})}{\partial \mathbf{G}^*}&=E_{\mathbf{x}}\left\{\frac{\partial p(\mathbf{y}|\mathbf{x})}{\partial \mathbf{G}^*}\right\}=E_{\mathbf{x}}\left\{p(\mathbf{y}|\mathbf{x})\frac{\mathbf{H}^h(\mathbf{y}-\mathbf{HGx})\mathbf{x}^h}{\sigma^2}\right\}\\&=-\mathbf{H}^hE_{\mathbf{x}}\left\{\left[\mathcal{D}_{\mathbf{y}}p(\mathbf{y}|\mathbf{x})\right]^h\mathbf{x}^h\right\}.\end{aligned}\tag{49}$$

The proofs of (26) and (27) follows from their relationships to (24) and (25), respectively. □

Proof of Lemma 2: We prove (28) first.

$$\begin{aligned}\mathcal{D}_{\mathbf{y}}E\{\mathbf{x}|\mathbf{y}\}&=\mathcal{D}_{\mathbf{y}}E_{\mathbf{x}}\left\{\mathbf{x}\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right\}\\&=E_{\mathbf{x}}\left\{\mathbf{x}\frac{\mathcal{D}_{\mathbf{y}}p(\mathbf{y}|\mathbf{x})\cdot p(\mathbf{y})-p(\mathbf{y}|\mathbf{x})\cdot\mathcal{D}_{\mathbf{y}}p(\mathbf{y})}{[p(\mathbf{y})]^2}\right\}\\&=E_{\mathbf{x}}\left\{\mathbf{x}\frac{-p(\mathbf{y}|\mathbf{x})(\mathbf{y}-\mathbf{HGx})^h}{p(\mathbf{y})\sigma^2}\right\}\\&\quad+E_{\mathbf{x}}\left\{\mathbf{x}\frac{p(\mathbf{y}|\mathbf{x})\left(\mathbf{y}-\mathbf{HG}E\{\mathbf{x}|\mathbf{y}\}\right)^h}{p(\mathbf{y})\sigma^2}\right\}\\&=E_{\mathbf{x}}\left\{\mathbf{x}\mathbf{x}^h\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\frac{\mathbf{G}^h\mathbf{H}^h}{\sigma^2}\right\}\\&\quad-E_{\mathbf{x}}\left\{\mathbf{x}\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}E\{\mathbf{x}|\mathbf{y}\}^h\frac{\mathbf{G}^h\mathbf{H}^h}{\sigma^2}\right\}\\&=\left[E\{\mathbf{x}\mathbf{x}^h|\mathbf{y}\}-E\{\mathbf{x}|\mathbf{y}\}E\{\mathbf{x}|\mathbf{y}\}^h\right]\frac{\mathbf{G}^h\mathbf{H}^h}{\sigma^2}\\&=\mathbf{\Phi}_{\mathbf{x}\mathbf{x}^h}(\mathbf{y})\mathbf{G}^h\mathbf{H}^h\frac{1}{\sigma^2}.\end{aligned}\tag{50}$$

The proof of (29) is similar to that of (28). We omit the details for brevity. □

Proof of Lemma 3: The proof of (30) is given here.

$$\begin{aligned}&\frac{\partial E\{x_i|\mathbf{y}\}}{\partial \mathbf{G}^*}\\&=\frac{\partial}{\partial \mathbf{G}^*}E\left\{x_i\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right\}\\&=E\left\{x_i\frac{1}{p(\mathbf{y})}\frac{\partial p(\mathbf{y}|\mathbf{x})}{\partial \mathbf{G}^*}-x_i\frac{p(\mathbf{y}|\mathbf{x})}{[p(\mathbf{y})]^2}\frac{\partial p(\mathbf{y})}{\partial \mathbf{G}^*}\right\}\\&=-E\left\{x_i\frac{\mathbf{H}^h\left[\mathcal{D}_{\mathbf{y}}p(\mathbf{y}|\mathbf{x})\right]^h\mathbf{x}^h}{p(\mathbf{y})}\right\}\\&\quad+E\left\{x_i\frac{p(\mathbf{y}|\mathbf{x})\mathbf{H}^hE_{\mathbf{x}}\left\{\left[\mathcal{D}_{\mathbf{y}}p(\mathbf{y}|\mathbf{x})\right]^h\mathbf{x}^h\right\}}{[p(\mathbf{y})]^2}\right\}\\&=\frac{1}{p(\mathbf{y})}\mathbf{H}^h\Big[-E\left\{x_i\left[\mathcal{D}_{\mathbf{y}}p(\mathbf{y}|\mathbf{x})\right]^h\mathbf{x}^h\right\}\\&\qquad+E\left\{x_i\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right\}E_{\mathbf{x}}\left\{\left[\mathcal{D}_{\mathbf{y}}p(\mathbf{y}|\mathbf{x})\right]^h\mathbf{x}^h\right\}\Big]\\&=\frac{1}{p(\mathbf{y})}\mathbf{H}^h\Big[-E\left\{x_i\left[\mathcal{D}_{\mathbf{y}}p(\mathbf{y}|\mathbf{x})\right]^h\mathbf{x}^h\right\}\\&\qquad+E\{x_i|\mathbf{y}\}E_{\mathbf{x}}\left\{\left[\mathcal{D}_{\mathbf{y}}p(\mathbf{y}|\mathbf{x})\right]^h\mathbf{x}^h\right\}\Big].\end{aligned}\tag{51}$$

Similarly, we can prove (31). □

Proof of Lemma 4: Let $[\mathbf{W}]_{mn}$ denote the (m,n)th element of $\mathbf{W}$. We then have

$$[\mathbf{W}]_{mn}=\mathbf{e}_m^h\mathbf{G}^h\mathbf{H}^h\mathbf{HG}\mathbf{e}_n=\mathrm{trace}\left(\mathbf{G}^h\mathbf{H}^h\mathbf{HG}\mathbf{e}_n\mathbf{e}_m^h\right)\tag{52}$$

where $\mathbf{e}_m$ is the mth column of an identity matrix with appropriate dimension. Based on the complex-valued matrix differentiation rules [26], one can easily obtain the following

$$\frac{\partial[\mathbf{W}]_{mn}}{\partial \mathbf{G}^*}=\mathbf{H}^h\mathbf{HG}\mathbf{e}_n\mathbf{e}_m^h\tag{53}$$

$$\begin{aligned}\frac{\partial[\mathbf{W}]_{mn}}{\partial[\mathbf{G}^*]_{kl}}&=\mathbf{e}_k^h\mathbf{H}^h\mathbf{HG}\mathbf{e}_n\mathbf{e}_m^h\mathbf{e}_l\\&=\mathbf{e}_m^h\mathbf{e}_l\mathbf{e}_k^h\mathbf{H}^h\mathbf{HG}\mathbf{e}_n\end{aligned}\tag{54}$$

$$\frac{\partial \mathbf{W}}{\partial[\mathbf{G}^*]_{kl}}=\mathbf{e}_l\mathbf{e}_k^h\mathbf{H}^h\mathbf{HG}.\tag{55}$$

Keeping in mind that $\mathbf{W}^h=\mathbf{W}$, one can use the chain rule to have

$$\begin{aligned}\frac{\partial\mathcal{I}(\mathbf{x};\mathbf{y})}{\partial[\mathbf{G}^*]_{kl}}&=\mathrm{trace}\left\{\left(\frac{\partial\mathcal{I}(\mathbf{x};\mathbf{y})}{\partial\mathbf{W}}\right)^t\frac{\partial\mathbf{W}}{\partial[\mathbf{G}^*]_{kl}}\right\}\\&=\mathrm{trace}\left\{\frac{\partial\mathcal{I}(\mathbf{x};\mathbf{y})}{(\partial\mathbf{W}^h)^t}\frac{\partial(\mathbf{G}^h\mathbf{H}^h\mathbf{HG})}{\partial[\mathbf{G}^*]_{kl}}\right\}\\&=\mathrm{trace}\left\{\frac{\partial\mathcal{I}(\mathbf{x};\mathbf{y})}{\partial\mathbf{W}^*}\left(\mathbf{e}_l\mathbf{e}_k^h\mathbf{H}^h\mathbf{HG}\right)\right\}\\&=\mathbf{e}_k^h\mathbf{H}^h\mathbf{HG}\frac{\partial\mathcal{I}(\mathbf{x};\mathbf{y})}{\partial\mathbf{W}^*}\mathbf{e}_l.\end{aligned}\tag{56}$$

$$\frac{\partial\mathcal{I}(\mathbf{x};\mathbf{y})}{\partial\mathbf{G}^*}=\mathbf{H}^h\mathbf{HG}\frac{\partial\mathcal{I}(\mathbf{x};\mathbf{y})}{\partial\mathbf{W}^*}.\tag{57}$$

This completes the proof of Lemma 4. □

Proof of Lemma 5: The first equality (33) is validated as follows:

$$\begin{aligned}&E\left\{\left[\mathbf{x}^*\otimes E\{\mathbf{x}|\mathbf{y}\}\right]|\mathbf{y}\right\}\\&\quad=E_{\mathbf{x}}\left\{\left[\mathbf{x}^*\otimes E_{\mathbf{x}}\left\{\mathbf{x}\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right\}\right]\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right\}\\&\quad=E_{\mathbf{x}}\left\{\mathbf{x}^*\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right\}\otimes E_{\mathbf{x}}\left\{\mathbf{x}\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\right\}\\&\quad=E\{\mathbf{x}|\mathbf{y}\}^*\otimes E\{\mathbf{x}|\mathbf{y}\}.\end{aligned}\tag{58}$$

The left-hand side and right-hand side (LHS and RHS) of the second (34) can be rewritten by

$$\begin{aligned}
&E\{(\mathbf{x}^*\otimes\mathbf{x})|\mathbf{y}\}\,E\{(\mathbf{x}^*\otimes\mathbf{x})|\mathbf{y}\}^h\\
&=\sum_{m=1}^{M^{N_t}}(\mathbf{x}_m^*\otimes\mathbf{x}_m)\,p(\mathbf{x}_m|\mathbf{y})\sum_{n=1}^{M^{N_t}}(\mathbf{x}_n^*\otimes\mathbf{x}_n)^h\,p(\mathbf{x}_n|\mathbf{y})\\
&=\sum_{m=1}^{M^{N_t}}\sum_{n=1}^{M^{N_t}}\left[\left(\mathbf{x}_m^*\mathbf{x}_n^t\right)\otimes\left(\mathbf{x}_m\mathbf{x}_n^h\right)\right]p(\mathbf{x}_m|\mathbf{y})p(\mathbf{x}_n|\mathbf{y}). \qquad (59)
\end{aligned}$$

$$\begin{aligned}
&E\left\{(\mathbf{x}\mathbf{x}^h)|\mathbf{y}\right\}^*\otimes E\left\{(\mathbf{x}\mathbf{x}^h)|\mathbf{y}\right\}\\
&=\sum_{m=1}^{M^{N_t}}\left(\mathbf{x}_m\mathbf{x}_m^h\right)^*p(\mathbf{x}_m|\mathbf{y})\otimes\sum_{n=1}^{M^{N_t}}\mathbf{x}_n\mathbf{x}_n^h p(\mathbf{x}_n|\mathbf{y})\\
&=\sum_{m=1}^{M^{N_t}}\sum_{n=1}^{M^{N_t}}\left[\left(\mathbf{x}_m^*\mathbf{x}_m^t\right)\otimes\left(\mathbf{x}_n\mathbf{x}_n^h\right)\right]p(\mathbf{x}_m|\mathbf{y})p(\mathbf{x}_n|\mathbf{y}). \qquad (60)
\end{aligned}$$

The only work now that remains is to show the indices can be swapped in the two different equations. If we partition the matrix $E\{(\mathbf{x}^*\otimes\mathbf{x})|\mathbf{y}\}E\{(\mathbf{x}^*\otimes\mathbf{x})|\mathbf{y}\}^h$ into blocks with each block being $N_t\times N_t$ dimensions, then its (i,j)th block, denoting as $\Delta_L^{(i,j)}$, is given by

$$\begin{aligned}
\Delta_L^{(i,j)}&=\sum_{m=1}^{M^{N_t}}\sum_{n=1}^{M^{N_t}}\left[\left(x_{m,i}^*\mathbf{x}_n^t\right)\otimes\left(\mathbf{x}_m x_{n,j}^*\right)\right]p(\mathbf{x}_m|\mathbf{y})p(\mathbf{x}_n|\mathbf{y})\\
&=\sum_{m=1}^{M^{N_t}}\sum_{n=1}^{M^{N_t}}\left[x_{m,i}^*x_{n,j}^*\mathbf{x}_m\mathbf{x}_n^t\right]p(\mathbf{x}_m|\mathbf{y})p(\mathbf{x}_n|\mathbf{y}). \qquad (61)
\end{aligned}$$

Similarly, the (i,j)th block of $E\{(\mathbf{x}\mathbf{x}^h)|\mathbf{y}\}^*\otimes E\{(\mathbf{x}\mathbf{x}^h)|\mathbf{y}\}$, denoting as $\Delta_R^{(i,j)}$, is given by

$$\begin{aligned}
\Delta_R^{(i,j)}&=\sum_{m=1}^{M^{N_t}}\sum_{n=1}^{M^{N_t}}\left[\left(x_{m,i}^*\mathbf{x}_m^t\right)\otimes\left(\mathbf{x}_n x_{n,j}^*\right)\right]p(\mathbf{x}_m|\mathbf{y})p(\mathbf{x}_n|\mathbf{y})\\
&=\sum_{m=1}^{M^{N_t}}\sum_{n=1}^{M^{N_t}}\left[x_{m,i}^*x_{n,j}^*\mathbf{x}_n\mathbf{x}_m^t\right]p(\mathbf{x}_m|\mathbf{y})p(\mathbf{x}_n|\mathbf{y}). \qquad (62)
\end{aligned}$$

If we exchange the two dummy summation indices m and n, we can easily see that these two blocks $\Delta_L^{(i,j)}$ and $\Delta_R^{(i,j)}$ are identical. Therefore, the second equality of Lemma 5 holds.

We can prove the third equality (35) in a similar manner. Let $\Theta_L^{(i,j)}$ and $\Theta_R^{(i,j)}$ be the (i,j)th block of the LHS and the RHS of the third equality, respectively. After some mathematical manipulations, we obtain the following:

$$\begin{aligned}
\Theta_L^{(i,j)}&=\sum_{m=1}^{M^{N_t}}\sum_{n=1}^{M^{N_t}}\sum_{l=1}^{M^{N_t}}\left[\left(x_{m,i}^*\otimes\mathbf{x}_m\right)\left(\mathbf{x}_n^*\otimes x_{l,j}\right)^h\right]\\
&\quad\times p(\mathbf{x}_m|\mathbf{y})p(\mathbf{x}_n|\mathbf{y})p(\mathbf{x}_l|\mathbf{y})\\
&=\sum_{m=1}^{M^{N_t}}\sum_{n=1}^{M^{N_t}}\sum_{l=1}^{M^{N_t}}\left(x_{m,i}^*x_{l,j}^*\mathbf{x}_m\mathbf{x}_n^t\right)p(\mathbf{x}_m|\mathbf{y})p(\mathbf{x}_n|\mathbf{y})\\
&\quad\times p(\mathbf{x}_l|\mathbf{y}). \qquad (63a)
\end{aligned}$$

$$\begin{aligned}
\Theta_R^{(i,j)}&=\sum_{m=1}^{M^{N_t}}\sum_{n=1}^{M^{N_t}}\sum_{l=1}^{M^{N_t}}\left[\left(x_{m,i}^*\mathbf{x}_m^t\right)\otimes\left(\mathbf{x}_n x_{l,j}^*\right)\right]p(\mathbf{x}_m|\mathbf{y})\\
&\quad\times p(\mathbf{x}_n|\mathbf{y})p(\mathbf{x}_l|\mathbf{y})\\
&=\sum_{m=1}^{M^{N_t}}\sum_{n=1}^{M^{N_t}}\sum_{l=1}^{M^{N_t}}\left(x_{m,i}^*x_{l,j}^*\mathbf{x}_n\mathbf{x}_m^t\right)p(\mathbf{x}_m|\mathbf{y})p(\mathbf{x}_n|\mathbf{y})\\
&\quad\times p(\mathbf{x}_l|\mathbf{y}). \qquad (63b)
\end{aligned}$$

Apparently, $\Theta_L^{(i,j)}$ and $\Theta_R^{(i,j)}$ are identical; therefore, the third equality holds, hence completing the proof of Lemma 5. □

Proof of Lemma 6: The proof of this lemma is lengthy.

The (i,j)th element of the MMSE matrix $\mathbf{\Phi}$ is given by $\mathbf{\Phi}_{ij}=E\{x_ix_j^*\}-E\{E\{x_i|\mathbf{y}\}E\{x_j^*|\mathbf{y}\}\}$. Therefore

$$\begin{aligned}
\frac{\partial\mathbf{\Phi}_{ij}}{\partial\mathbf{G}_{kl}^*}&=-\int\frac{\partial p(\mathbf{y})}{\partial\mathbf{G}_{kl}^*}E\{x_i|\mathbf{y}\}E\left\{x_j^*|\mathbf{y}\right\}d\mathbf{y}\\
&\quad-\int p(\mathbf{y})\frac{\partial E\{x_i|\mathbf{y}\}}{\partial\mathbf{G}_{kl}^*}E\left\{x_j^*|\mathbf{y}\right\}d\mathbf{y}\\
&\quad-\int p(\mathbf{y})E\{x_i|\mathbf{y}\}\frac{\partial E\left\{x_j^*|\mathbf{y}\right\}}{\partial\mathbf{G}_{kl}^*}d\mathbf{y}. \qquad (64)
\end{aligned}$$

Utilizing Lemmas 1 and 3, and $\frac{\partial}{\partial\mathbf{G}_{kl}^*}=\mathbf{e}_k^t\left[\frac{\partial}{\partial\mathbf{G}^*}\right]\mathbf{e}_l$, where $\mathbf{e}_k$ is the kth column of an identity matrix with appropriate dimension, one can obtain

$$\begin{aligned}
\frac{\partial\mathbf{\Phi}_{ij}}{\partial\mathbf{G}_{kl}^*}&=\int\mathbf{e}_k^t\mathbf{H}^hE_\mathbf{x}\left\{x_i\left[\mathcal{D}_\mathbf{y}p(\mathbf{y}|\mathbf{x})\right]^hx_l^*\right\}E\left\{x_j^*|\mathbf{y}\right\}d\mathbf{y}\\
&\quad+\int E\{x_i|\mathbf{y}\}\mathbf{e}_k^t\mathbf{H}^hE_\mathbf{x}\left\{x_j^*\left[\mathcal{D}_\mathbf{y}p(\mathbf{y}|\mathbf{x})\right]^hx_l^*\right\}d\mathbf{y}\\
&\quad-\int E\{x_i|\mathbf{y}\}E\left\{x_j^*|\mathbf{y}\right\}\mathbf{e}_k^t\mathbf{H}^h\\
&\qquad\times E_\mathbf{x}\left\{\left[\mathcal{D}_\mathbf{y}p(\mathbf{y}|\mathbf{x})\right]^hx_l^*\right\}d\mathbf{y}. \qquad (65)
\end{aligned}$$

We split the three integrations separately. The first integrand in (65) can be re-expressed as

$$\begin{aligned}
&\mathbf{e}_k^t\mathbf{H}^hE_\mathbf{x}\left\{x_i\left[\mathcal{D}_\mathbf{y}p(\mathbf{y}|\mathbf{x})\right]^hx_l^*\right\}E\left\{x_j^*|\mathbf{y}\right\}\\
&=\mathbf{e}_k^t\mathbf{H}^hE_\mathbf{x}\left\{x_ix_l^*\left[\frac{\partial p(\mathbf{y}|\mathbf{x})}{\partial\mathrm{vec}^t(\mathbf{y})}\right]^h\right\}E\left\{x_j^*|\mathbf{y}\right\}\\
&=\mathbf{e}_k^t\mathbf{H}^h\frac{\partial p(\mathbf{y})E\left\{x_ix_l^*|\mathbf{y}\right\}}{\partial\mathrm{vec}(\mathbf{y}^*)}E\left\{x_j^*|\mathbf{y}\right\}. \qquad (66)
\end{aligned}$$

Thus, the first term in the integration of (65) becomes

$$\begin{aligned}
&\int\mathbf{e}_k^t\mathbf{H}^h\frac{\partial p(\mathbf{y})E\left\{x_ix_l^*|\mathbf{y}\right\}}{\partial\mathrm{vec}(\mathbf{y}^*)}E\left\{x_j^*|\mathbf{y}\right\}d\mathbf{y}\\
&=-\int\mathbf{e}_k^t\mathbf{H}^hp(\mathbf{y})E\left\{x_ix_l^*|\mathbf{y}\right\}\frac{\partial E\left\{x_j^*|\mathbf{y}\right\}}{\partial\mathrm{vec}(\mathbf{y}^*)}d\mathbf{y}\\
&=\frac{-1}{\sigma^2}\int\mathbf{e}_k^t\mathbf{H}^h\mathbf{H}\mathbf{G}\mathbf{\Phi}_{\mathbf{x}\mathbf{x}^h}(\mathbf{y})\mathbf{e}_jE\left\{x_ix_l^*|\mathbf{y}\right\}p(\mathbf{y})d\mathbf{y}. \qquad (67)
\end{aligned}$$

The first step follows from integration by parts whereas the second step is due to Lemma 2.

Similarly, the second integrand and the second term in (65) can be reexpressed, respectively, as

$$\begin{aligned}
&E\{x_i|\mathbf{y}\}\mathbf{e}_k^t\mathbf{H}^hE_\mathbf{x}\left\{x_j^*\left[\mathcal{D}_\mathbf{y}p(\mathbf{y}|\mathbf{x})\right]^hx_l^*\right\}\\
&=E\{x_i|\mathbf{y}\}\mathbf{e}_k^t\mathbf{H}^hE_\mathbf{x}\left\{x_j^*x_l^*\frac{\partial p(\mathbf{y}|\mathbf{x})}{\partial\mathrm{vec}(\mathbf{y}^*)}\right\}\\
&=E\{x_i|\mathbf{y}\}\mathbf{e}_k^t\mathbf{H}^h\frac{\partial p(\mathbf{y})E\left\{x_j^*x_l^*|\mathbf{y}\right\}}{\partial\mathrm{vec}(\mathbf{y}^*)} \qquad (68)
\end{aligned}$$

$$\int E\{x_i|\mathbf{y}\}\mathbf{e}_k^t\mathbf{H}^h\frac{\partial p(\mathbf{y})E\{x_j^*x_l^*|\mathbf{y}\}}{\partial \mathrm{vec}(\mathbf{y}^*)}d\mathbf{y}$$
$$= -\int \mathbf{e}_k^t\mathbf{H}^h p(\mathbf{y})E\{x_j^*x_l^*|\mathbf{y}\}\frac{\partial E\{x_i|\mathbf{y}\}}{\partial \mathrm{vec}(\mathbf{y}^*)}d\mathbf{y}$$
$$= \frac{-1}{\sigma^2}\int \mathbf{e}_k^t\mathbf{H}^h\mathbf{H}\mathbf{G}\boldsymbol{\Phi}_{\mathbf{x}\mathbf{x}^t}(\mathbf{y})\mathbf{e}_i E\{x_j^*x_l^*|\mathbf{y}\}p(\mathbf{y})d\mathbf{y}. \quad (69)$$

The third integrand in (65) is

$$E\{x_i|\mathbf{y}\}E\{x_j^*|\mathbf{y}\}\mathbf{e}_k^t\mathbf{H}^h E_{\mathbf{x}}\left\{[\mathcal{D}_{\mathbf{y}}p(\mathbf{y}|\mathbf{x})]^h x_l^*\right\}$$
$$= -E\{x_i|\mathbf{y}\}E\{x_j^*|\mathbf{y}\}\mathbf{e}_k^t\mathbf{H}^h E_{\mathbf{x}}\left\{x_l^*\frac{\partial p(\mathbf{y}|\mathbf{x})}{\partial \mathrm{vec}(\mathbf{y}^*)}\right\}$$
$$= -E\{x_i|\mathbf{y}\}E\{x_j^*|\mathbf{y}\}\mathbf{e}_k^t\mathbf{H}^h\frac{\partial p(\mathbf{y})E\{x_l^*|\mathbf{y}\}}{\partial \mathrm{vec}(\mathbf{y}^*)}. \quad (70)$$

Hence, the third integration becomes

$$-\int E\{x_i|\mathbf{y}\}E\{x_j^*|\mathbf{y}\}\mathbf{e}_k^t\mathbf{H}^h\frac{\partial p(\mathbf{y})E\{x_l^*|\mathbf{y}\}}{\partial \mathrm{vec}(\mathbf{y}^*)}d\mathbf{y}$$
$$= \int p(\mathbf{y})E\{x_l^*|\mathbf{y}\}\mathbf{e}_k^t\mathbf{H}^h\frac{\partial E\{x_i|\mathbf{y}\}E\{x_j^*|\mathbf{y}\}}{\partial \mathrm{vec}(\mathbf{y}^*)}d\mathbf{y}$$
$$= \int p(\mathbf{y})E\{x_l^*|\mathbf{y}\}\mathbf{e}_k^t\mathbf{H}^h\left[\frac{\partial E\{x_i|\mathbf{y}\}}{\partial \mathrm{vec}(\mathbf{y}^*)}E\{x_j^*|\mathbf{y}\} + E\{x_i|\mathbf{y}\}\frac{\partial E\{x_j^*|\mathbf{y}\}}{\partial \mathrm{vec}(\mathbf{y}^*)}\right]d\mathbf{y}$$
$$= \frac{1}{\sigma^2}\int p(\mathbf{y})E\{x_l^*|\mathbf{y}\}E\{x_j^*|\mathbf{y}\}\mathbf{e}_k^t\mathbf{H}^h\mathbf{H}\mathbf{G}\boldsymbol{\Phi}_{\mathbf{x}\mathbf{x}^t}(\mathbf{y})\mathbf{e}_i d\mathbf{y}$$
$$+\frac{1}{\sigma^2}\int p(\mathbf{y})E\{x_l^*|\mathbf{y}\}E\{x_i|\mathbf{y}\} \times \mathbf{e}_k^t\mathbf{H}^h\mathbf{H}\mathbf{G}\boldsymbol{\Phi}_{\mathbf{x}\mathbf{x}^h}(\mathbf{y})\mathbf{e}_j d\mathbf{y}. \quad (71)$$

Substituting (67), (69), and (71) into (65), we have

$$\frac{\partial \boldsymbol{\Phi}_{ij}}{\partial \mathbf{G}_{kl}^*} = -\frac{1}{\sigma^2}\int \mathbf{e}_k^t\mathbf{H}^h\mathbf{H}\mathbf{G}\boldsymbol{\Phi}_{\mathbf{x}\mathbf{x}^h}(\mathbf{y})\mathbf{e}_j p(\mathbf{y}) \times [E\{x_i x_l^*|\mathbf{y}\} - E\{x_i|\mathbf{y}\}E\{x_l^*|\mathbf{y}\}]d\mathbf{y}$$
$$-\frac{1}{\sigma^2}\int \mathbf{e}_k^t\mathbf{H}^h\mathbf{H}\mathbf{G}\boldsymbol{\Phi}_{\mathbf{x}\mathbf{x}^t}(\mathbf{y})\mathbf{e}_i p(\mathbf{y}) \times [E\{x_j^*x_l^*|\mathbf{y}\} - E\{x_j^*|\mathbf{y}\}E\{x_l^*|\mathbf{y}\}]d\mathbf{y}$$
$$= -\frac{1}{\sigma^2}E_{\mathbf{y}}\left\{\mathbf{e}_k^t\mathbf{H}^h\mathbf{H}\mathbf{G}\boldsymbol{\Phi}_{\mathbf{x}\mathbf{x}^h}(\mathbf{y})\mathbf{e}_j\mathbf{e}_i^t\boldsymbol{\Phi}_{\mathbf{x}\mathbf{x}^h}(\mathbf{y})\mathbf{e}_l\right\}$$
$$-\frac{1}{\sigma^2}E_{\mathbf{y}}\left\{\mathbf{e}_k^t\mathbf{H}^h\mathbf{H}\mathbf{G}\boldsymbol{\Phi}_{\mathbf{x}\mathbf{x}^t}(\mathbf{y})\mathbf{e}_i\mathbf{e}_j^t\boldsymbol{\Phi}_{\mathbf{x}\mathbf{x}^t}^*(\mathbf{y})\mathbf{e}_l\right\}$$
$$= -\frac{1}{\sigma^2}E_{\mathbf{y}}\left\{\mathbf{e}_i^t\boldsymbol{\Phi}_{\mathbf{x}\mathbf{x}^h}(\mathbf{y})\mathbf{e}_l\mathbf{e}_j^t\boldsymbol{\Phi}_{\mathbf{x}\mathbf{x}^h}^t(\mathbf{y})\mathbf{G}^t\mathbf{H}^t\mathbf{H}^*\mathbf{e}_k\right\}$$
$$-\frac{1}{\sigma^2}E_{\mathbf{y}}\left\{\mathbf{e}_j^t\boldsymbol{\Phi}_{\mathbf{x}\mathbf{x}^t}^*(\mathbf{y})\mathbf{e}_l\mathbf{e}_i^t\boldsymbol{\Phi}_{\mathbf{x}\mathbf{x}^t}(\mathbf{y})\mathbf{G}^t\mathbf{H}^t\mathbf{H}^*\mathbf{e}_k\right\}$$
$$= \frac{-1}{\sigma^2}E_{\mathbf{y}}\left\{\left[\mathbf{K}_{N_t}\left(\boldsymbol{\Phi}_{\mathbf{x}\mathbf{x}^h}(\mathbf{y}) \otimes \left[\boldsymbol{\Phi}_{\mathbf{x}\mathbf{x}^h}^t(\mathbf{y})\mathbf{G}^t\mathbf{H}^t\mathbf{H}^*\right]\right)\right]_{p,q}\right\}$$
$$-\frac{1}{\sigma^2}E_{\mathbf{y}}\left\{\left(\boldsymbol{\Phi}_{\mathbf{x}\mathbf{x}^t}^*(\mathbf{y}) \otimes \left[\boldsymbol{\Phi}_{\mathbf{x}\mathbf{x}^t}(\mathbf{y})\mathbf{G}^t\mathbf{H}^t\mathbf{H}^*\right]\right)_{p,q}\right\}. \quad (72)$$

with $p = i + (j-1)N_t$ and $q = k + (l-1)N_t$.

From $\frac{\partial \boldsymbol{\Phi}_{ij}}{\partial \mathbf{G}_{kl}^*} = [\mathcal{D}_{\mathbf{G}^*}\boldsymbol{\Phi}]_{i+(j-1)N_t, k+(l-1)N_t}$, we have

$$\mathcal{D}_{\mathbf{G}^*}\boldsymbol{\Phi} = -\frac{1}{\sigma^2}\mathbf{K}_{N_t}E_{\mathbf{y}}\left\{\boldsymbol{\Phi}_{\mathbf{x}\mathbf{x}^h}(\mathbf{y}) \otimes \left[\boldsymbol{\Phi}_{\mathbf{x}\mathbf{x}^h}^t(\mathbf{y})\mathbf{G}^t\mathbf{H}^t\mathbf{H}^*\right]\right\}$$
$$-\frac{1}{\sigma^2}E_{\mathbf{y}}\left\{\boldsymbol{\Phi}_{\mathbf{x}\mathbf{x}^t}^*(\mathbf{y}) \otimes \left[\boldsymbol{\Phi}_{\mathbf{x}\mathbf{x}^t}(\mathbf{y})\mathbf{G}^t\mathbf{H}^t\mathbf{H}^*\right]\right\}$$
$$= \frac{-1}{\sigma^2}\mathbf{K}_{N_t}E_{\mathbf{y}}\left\{\boldsymbol{\Phi}_{\mathbf{x}\mathbf{x}^h}(\mathbf{y}) \otimes \boldsymbol{\Phi}_{\mathbf{x}\mathbf{x}^h}^*(\mathbf{y})\right\} \times \left[\mathbf{I}_{N_t} \otimes \mathbf{G}^t\mathbf{H}^t\mathbf{H}^*\right]$$
$$-\frac{1}{\sigma^2}E_{\mathbf{y}}\left\{\boldsymbol{\Phi}_{\mathbf{x}\mathbf{x}^t}^*(\mathbf{y}) \otimes \boldsymbol{\Phi}_{\mathbf{x}\mathbf{x}^t}(\mathbf{y})\right\} \times \left[\mathbf{I}_{N_t} \otimes \mathbf{G}^t\mathbf{H}^t\mathbf{H}^*\right]. \quad (73)$$

Note that $\mathbf{K}_{N_t}E_{\mathbf{y}}\{\boldsymbol{\Phi}_{\mathbf{x}\mathbf{x}^h}(\mathbf{y}) \otimes \boldsymbol{\Phi}_{\mathbf{x}\mathbf{x}^h}^*(\mathbf{y})\} = E_{\mathbf{y}}\{\boldsymbol{\Phi}_{\mathbf{x}\mathbf{x}^h}^*(\mathbf{y}) \otimes \boldsymbol{\Phi}_{\mathbf{x}\mathbf{x}^h}(\mathbf{y})\}\mathbf{K}_{N_t}$, (73) becomes (36), which finally completes the proof of Lemma 6. □

Proof of Lemma 7: Based on the chain rule provided by Theorem 1 of [26], we have

$$\mathcal{D}_{\mathbf{G}^*}\boldsymbol{\Phi} = [\mathcal{D}_{\mathbf{W}}\boldsymbol{\Phi}][\mathcal{D}_{\mathbf{G}^*}\mathbf{W}] + [\mathcal{D}_{\mathbf{W}^*}\boldsymbol{\Phi}][\mathcal{D}_{\mathbf{G}^*}\mathbf{W}^*]. \quad (74)$$

From [26], we know that

$$d\mathbf{W} = (d\mathbf{G}^h)\mathbf{H}^h\mathbf{H}\mathbf{G} + \mathbf{G}^h\mathbf{H}^h\mathbf{H}(d\mathbf{G}) \quad (75)$$

and

$$d\mathrm{vec}(\mathbf{W}) = \left[(\mathbf{G}^t\mathbf{H}^t\mathbf{H}^*) \otimes \mathbf{I}_{N_t}\right]d\mathrm{vec}(\mathbf{G}^h) + \left[\mathbf{I}_{N_t} \otimes (\mathbf{G}^h\mathbf{H}^h\mathbf{H})\right]d\mathrm{vec}(\mathbf{G})$$
$$= \left[(\mathbf{G}^t\mathbf{H}^t\mathbf{H}^*) \otimes \mathbf{I}_{N_t}\right]\mathbf{K}_{N_t}d\mathrm{vec}(\mathbf{G}^*) + \left[\mathbf{I}_{N_t} \otimes (\mathbf{G}^h\mathbf{H}^h\mathbf{H})\right]d\mathrm{vec}(\mathbf{G}). \quad (76)$$

Thus

$$\frac{\partial \mathrm{vec}(\mathbf{W})}{\partial \mathrm{vec}^t(\mathbf{G}^*)} = \left[(\mathbf{G}^t\mathbf{H}^t\mathbf{H}^*) \otimes \mathbf{I}_{N_t}\right]\mathbf{K}_{N_t} = \mathbf{K}_{N_t}\left[\mathbf{I}_{N_t} \otimes (\mathbf{G}^t\mathbf{H}^t\mathbf{H}^*)\right] \quad (77)$$

which is equivalent to

$$\mathcal{D}_{\mathbf{G}^*}\mathbf{W} = \mathbf{K}_{N_t}\left[\mathbf{I}_{N_t} \otimes (\mathbf{G}^t\mathbf{H}^t\mathbf{H}^*)\right]. \quad (78)$$

Similarly, we can obtain

$$d\mathrm{vec}(\mathbf{W}^*) = \left[(\mathbf{G}^h\mathbf{H}^h\mathbf{H}) \otimes \mathbf{I}_{N_t}\right]d\mathrm{vec}(\mathbf{G}^t) + \left[\mathbf{I}_{N_t} \otimes (\mathbf{G}^t\mathbf{H}^t\mathbf{H}^*)\right]d\mathrm{vec}(\mathbf{G}^*) \quad (79)$$

and

$$\mathcal{D}_{\mathbf{G}^*}\mathbf{W}^* = \mathbf{I}_{N_t} \otimes (\mathbf{G}^t\mathbf{H}^t\mathbf{H}^*). \quad (80)$$

Based on the definition of complex Hessian matrix, we can easily obtain

$$\mathcal{H}_{\mathbf{W},\mathbf{W}^*}\mathcal{I}(\mathbf{x};\mathbf{y}) = \mathcal{D}_{\mathbf{W}}\boldsymbol{\Phi} \quad (81a)$$
$$\mathcal{H}_{\mathbf{W}^*,\mathbf{W}^*}\mathcal{I}(\mathbf{x};\mathbf{y}) = \mathcal{D}_{\mathbf{W}^*}\boldsymbol{\Phi}. \quad (81b)$$

Comparing (36), (74), (78), and (80), we can obtain (37) and (38). Utilizing conjugate properties, we can easily show (39) and (40). This completes the proof of Lemma 7. □

We are now is a position to prove the two theorems.

Proof of Theorem 1: The proof of $\mathcal{I}(\mathbf{x};\mathbf{y})$'s dependence on $\mathbf{G}$ through $\mathbf{W}$ is similar to that of Lemma 1 of [23], which considered real-valued vector Gaussian channels. Note that $\mathbf{G}^h\mathbf{H}^h\mathbf{y} = \mathbf{G}^h\mathbf{H}^h\mathbf{H}\mathbf{G}\mathbf{x} + \mathbf{G}^h\mathbf{H}^h\mathbf{v}$ is a sufficient statistics of $\mathbf{y}$ [34]. The first term on the RHS depends on $\mathbf{G}$ through $\mathbf{W} = \mathbf{G}^h\mathbf{H}^h\mathbf{H}\mathbf{G}$. The second term $\mathbf{G}^h\mathbf{H}^h\mathbf{v}$ is a zero-mean complex Gaussian random vector, and its statistical behavior is completely characterized by its covariance matrix, which equals $\sigma^2\mathbf{W}$. Since $\mathbf{x}$ and $\mathbf{v}$ are independent, $\mathcal{I}(\mathbf{x};\mathbf{y})$ depends on the precoder $\mathbf{G}$ only through $\mathbf{W}=\mathbf{G}^h\mathbf{H}^h\mathbf{H}\mathbf{G}$.

According to [32, eqn. (22)], $\frac{\partial \mathcal{I}(\mathbf{x};\mathbf{y})}{\partial \mathbf{G}^*}=\mathbf{H}^h\mathbf{H}\mathbf{G}\mathbf{\Phi}$. Invoking Lemma 4, one also has

$$\mathbf{H}^h\mathbf{H}\mathbf{G}\left[\frac{\partial \mathcal{I}(\mathbf{x};\mathbf{y})}{\partial \mathbf{W}^*}-\mathbf{\Phi}\right]=\mathbf{0}. \tag{82}$$

Because (82) is valid for any channel matrix and precoder matrix, we get $\frac{\partial \mathcal{I}(\mathbf{x};\mathbf{y})}{\partial \mathbf{W}^*}=\mathbf{\Phi}$, which validates (15).

To check the concavity of $\mathcal{I}(\mathbf{x};\mathbf{y})$ with respect to $\mathbf{W}$, the composite Hessian matrix of the mutual information $\mathcal{I}(\mathbf{x};\mathbf{y})$, which is given here, must be negative (semi)-definite [26].

$$\begin{aligned}&\mathcal{CH}_{\mathbf{W}}\mathcal{I}(\mathbf{x};\mathbf{y})\\ &\triangleq\begin{bmatrix}\mathcal{H}_{\mathbf{W},\mathbf{W}^*}\mathcal{I}(\mathbf{x};\mathbf{y}) & \mathcal{H}_{\mathbf{W}^*,\mathbf{W}^*}\mathcal{I}(\mathbf{x};\mathbf{y})\\ \mathcal{H}_{\mathbf{W},\mathbf{W}}\mathcal{I}(\mathbf{x};\mathbf{y}) & \mathcal{H}_{\mathbf{W}^*,\mathbf{W}}\mathcal{I}(\mathbf{x};\mathbf{y})\end{bmatrix}\\ &=\frac{-1}{\sigma^2}E\left\{\begin{bmatrix}\mathbf{\Phi}^*_{\mathbf{x}\mathbf{x}^h}(\mathbf{y})\otimes\mathbf{\Phi}_{\mathbf{x}\mathbf{x}^h}(\mathbf{y}) & \mathbf{\Psi}^*_{\mathbf{x}\mathbf{x}^t}(\mathbf{y})\otimes\mathbf{\Psi}_{\mathbf{x}\mathbf{x}^t}(\mathbf{y})\\ \mathbf{\Psi}_{\mathbf{x}\mathbf{x}^t}(\mathbf{y})\otimes\mathbf{\Psi}^*_{\mathbf{x}\mathbf{x}^t}(\mathbf{y}) & \mathbf{\Phi}_{\mathbf{x}\mathbf{x}^h}(\mathbf{y})\otimes\mathbf{\Phi}^*_{\mathbf{x}\mathbf{x}^h}(\mathbf{y})\end{bmatrix}\right\}.\end{aligned} \tag{83}$$

Utilizing Lemma 5, we can show the following:

$$\begin{aligned}\mathbf{\Phi}^*_{\mathbf{x}\mathbf{x}^h}(\mathbf{y})\otimes\mathbf{\Phi}_{\mathbf{x}\mathbf{x}^h}(\mathbf{y})=&E\left\{\left[(\mathbf{x}-E\{\mathbf{x}|\mathbf{y}\})^*\otimes(\mathbf{x}-E\{\mathbf{x}|\mathbf{y}\})\right]|\mathbf{y}\right\}\\ &\times E\left\{\left[(\mathbf{x}-E\{\mathbf{x}|\mathbf{y}\})^*\otimes(\mathbf{x}-E\{\mathbf{x}|\mathbf{y}\})\right]|\mathbf{y}\right\}^h.\end{aligned} \tag{84}$$

Similarly, we can also prove the following:

$$\begin{aligned}\mathbf{\Psi}^*_{\mathbf{x}\mathbf{x}^t}(\mathbf{y})\otimes\mathbf{\Psi}_{\mathbf{x}\mathbf{x}^t}(\mathbf{y})=&E\left\{\left[(\mathbf{x}-E\{\mathbf{x}|\mathbf{y}\})^*\otimes(\mathbf{x}-E\{\mathbf{x}|\mathbf{y}\})\right]|\mathbf{y}\right\}\\ &\times E\left\{\left[(\mathbf{x}-E\{\mathbf{x}|\mathbf{y}\})^*\otimes(\mathbf{x}-E\{\mathbf{x}|\mathbf{y}\})\right]|\mathbf{y}\right\}^t.\end{aligned} \tag{85}$$

Based on (83)–(85), we can rewrite

$$\begin{aligned}&\mathcal{CH}_{\mathbf{W}}\mathcal{I}(\mathbf{x};\mathbf{y})\\ &=\frac{-1}{\sigma^2}E\\ &\quad\times\left\{\begin{bmatrix}E\left\{\left[(\mathbf{x}-E\{\mathbf{x}|\mathbf{y}\})^*\otimes(\mathbf{x}-E\{\mathbf{x}|\mathbf{y}\})\right]|\mathbf{y}\right\}\\ E\left\{\left[(\mathbf{x}-E\{\mathbf{x}|\mathbf{y}\})\otimes(\mathbf{x}-E\{\mathbf{x}|\mathbf{y}\})^*\right]|\mathbf{y}\right\}\end{bmatrix}\right.\\ &\qquad\left.\times\begin{bmatrix}E\left\{\left[(\mathbf{x}-E\{\mathbf{x}|\mathbf{y}\})^*\otimes(\mathbf{x}-E\{\mathbf{x}|\mathbf{y}\})\right]|\mathbf{y}\right\}\\ E\left\{\left[(\mathbf{x}-E\{\mathbf{x}|\mathbf{y}\})\otimes(\mathbf{x}-E\{\mathbf{x}|\mathbf{y}\})^*\right]|\mathbf{y}\right\}\end{bmatrix}^h\right\}.\end{aligned} \tag{86}$$

Clearly, the composite Hessian matrix $\mathcal{CH}_{\mathbf{W}}\mathcal{I}(\mathbf{x};\mathbf{y})$ is negative (semi)-definite. Therefore, the mutual information $\mathcal{I}(\mathbf{x};\mathbf{y})$ is a concave function of $\mathbf{W}$. This completes the proof. □

Proof of Theorem 2: The proof of the partial derivative is similar to that of Lemma 4. An outline is provided as follows. From $\mathbf{W}=\mathbf{V}_G\mathbf{\Sigma}_H^2\mathbf{\Sigma}_G^2\mathbf{V}_G^h$, one can obtain

$$\frac{\partial \mathbf{W}}{\partial\left[\mathbf{\Sigma}_G^2\right]_{kk}}=\mathbf{V}_G\mathbf{\Sigma}_H^2\mathbf{e}_k\mathbf{e}_k^h\mathbf{V}_G^h. \tag{87}$$

Using the differential chain rule and $\mathbf{W}^t=\mathbf{W}^*$ leads to

$$\begin{aligned}\frac{\partial \mathcal{I}(\mathbf{x};\mathbf{y})}{\partial\left[\mathbf{\Sigma}_G^2\right]_{kk}}&=\mathrm{trace}\left\{\left(\frac{\partial \mathcal{I}(\mathbf{x};\mathbf{y})}{\partial \mathbf{W}}\right)^t\frac{\partial \mathbf{W}}{\partial\left[\mathbf{\Sigma}_G^2\right]_{kk}}\right\}\\ &=\mathbf{e}_k^h\mathbf{V}_G^h\frac{\partial \mathcal{I}(\mathbf{x};\mathbf{y})}{\partial \mathbf{W}^*}\mathbf{V}_G\mathbf{\Sigma}_H^2\mathbf{e}_k\\ &=\mathbf{e}_k^h\mathbf{V}_G^h\mathbf{\Phi}\mathbf{V}_G\mathbf{\Sigma}_H^2\mathbf{e}_k.\end{aligned} \tag{88}$$

Therefore, we have $\frac{\partial \mathcal{I}(\mathbf{x};\mathbf{y})}{\partial \mathbf{\Sigma}_G^2}=\mathrm{diag}\{\mathbf{V}_G^h\mathbf{\Phi}\mathbf{V}_G\mathbf{\Sigma}_H^2\}$, which validates the partial derivative given by (19).

To derive the Hessian of mutual information with respect to $\mathbf{\Sigma}_G^2$, we will need to first find the partial derivative of $\mathbf{\Phi}$ with respect to $[\mathbf{\Sigma}_G^2]_{ll}$, which can be achieved via finding the Jacobian $\mathcal{D}_{[\mathbf{\Sigma}_G^2]_{ll}}\mathbf{\Phi}$ through the derivative chain rule [26, (14)], as follows:

$$\mathcal{D}_{[\mathbf{\Sigma}_G^2]_{ll}}\mathbf{\Phi}=(\mathcal{D}_{\mathbf{W}}\mathbf{\Phi})\left(\mathcal{D}_{[\mathbf{\Sigma}_G^2]_{ll}}\mathbf{W}\right)+(\mathcal{D}_{\mathbf{W}^*}\mathbf{\Phi})\left(\mathcal{D}_{[\mathbf{\Sigma}_G^2]_{ll}}\mathbf{W}^*\right). \tag{89}$$

From (87), we get

$$\mathcal{D}_{[\mathbf{\Sigma}_G^2]_{ll}}\mathbf{W}=\left(\mathbf{V}_G^*\otimes\mathbf{V}_G\mathbf{\Sigma}_H^2\right)\mathrm{vec}\left(\mathbf{e}_l\mathbf{e}_l^h\right) \tag{90a}$$

$$\mathcal{D}_{[\mathbf{\Sigma}_G^2]_{ll}}\mathbf{W}^*=\left(\mathbf{V}_G\otimes\mathbf{V}_G^*\mathbf{\Sigma}_H^2\right)\mathrm{vec}\left(\mathbf{e}_l\mathbf{e}_l^h\right). \tag{90b}$$

Also from the proof of Lemma 7

$$\mathcal{D}_{\mathbf{W}}\mathbf{\Phi}=-\frac{1}{\sigma^2}E\left\{\mathbf{\Phi}^*_{\mathbf{x}\mathbf{x}^h}(\mathbf{y})\otimes\mathbf{\Phi}_{\mathbf{x}\mathbf{x}^h}(\mathbf{y})\right\} \tag{91a}$$

$$\mathcal{D}_{\mathbf{W}^*}\mathbf{\Phi}=-\frac{1}{\sigma^2}E\left\{\mathbf{\Psi}^*_{\mathbf{x}\mathbf{x}^t}(\mathbf{y})\otimes\mathbf{\Psi}_{\mathbf{x}\mathbf{x}^t}(\mathbf{y})\right\}. \tag{91b}$$

Substituting (90a)–(91b) to (89), and after some manipulations, one obtains

$$\begin{aligned}\mathcal{D}_{[\mathbf{\Sigma}_G^2]_{ll}}\mathbf{\Phi}=&\frac{-1}{\sigma^2}\mathrm{vec}\left[E\left\{\mathbf{\Phi}_{\mathbf{x}\mathbf{x}^h}(\mathbf{y})\mathbf{V}_G\mathbf{\Sigma}_H^2\mathbf{e}_l\mathbf{e}_l^h\mathbf{V}_G^h\mathbf{\Phi}^h_{\mathbf{x}\mathbf{x}^h}(\mathbf{y})\right\}\right]\\ &-\frac{1}{\sigma^2}\mathrm{vec}\left[E\left\{\mathbf{\Phi}_{\mathbf{x}\mathbf{x}^t}(\mathbf{y})\mathbf{V}_G^*\mathbf{\Sigma}_H^2\mathbf{e}_l\mathbf{e}_l^h\mathbf{V}_G^t\mathbf{\Phi}^h_{\mathbf{x}\mathbf{x}^t}(\mathbf{y})\right\}\right]\end{aligned} \tag{92}$$

and

$$\begin{aligned}\frac{\partial\mathbf{\Phi}}{\partial\left[\mathbf{\Sigma}_G^2\right]_{ll}}=&\frac{-1}{\sigma^2}E\left\{\mathbf{\Phi}_{\mathbf{x}\mathbf{x}^h}(\mathbf{y})\mathbf{V}_G\mathbf{\Sigma}_H^2\mathbf{e}_l\mathbf{e}_l^h\mathbf{V}_G^h\mathbf{\Phi}^h_{\mathbf{x}\mathbf{x}^h}(\mathbf{y})\right\}\\ &-\frac{1}{\sigma^2}E\left\{\mathbf{\Phi}_{\mathbf{x}\mathbf{x}^t}(\mathbf{y})\mathbf{V}_G^*\mathbf{\Sigma}_H^2\mathbf{e}_l\mathbf{e}_l^t\mathbf{V}_G^t\mathbf{\Phi}^h_{\mathbf{x}\mathbf{x}^t}(\mathbf{y})\right\}.\end{aligned} \tag{93}$$

According to (88) and (93), we can obtain

$$\begin{aligned}&\frac{\partial}{\partial\left[\mathbf{\Sigma}_G^2\right]_{ll}}\frac{\partial \mathcal{I}(\mathbf{x};\mathbf{y})}{\partial\left[\mathbf{\Sigma}_G^2\right]_{kk}}\\ &=-\frac{1}{\sigma^2}\left[\mathbf{\Sigma}_H^2\right]_{kk}E\left\{\mathbf{e}_k^h\mathbf{\Phi}_{\mathbf{g}\mathbf{g}^h}(\mathbf{y})\mathbf{e}_l\mathbf{e}_k^t\mathbf{\Phi}^*_{\mathbf{g}\mathbf{g}^h}(\mathbf{y})\mathbf{e}_l\right\}\left[\mathbf{\Sigma}_H^2\right]_{ll}\\ &\quad-\frac{1}{\sigma^2}\left[\mathbf{\Sigma}_H^2\right]_{kk}E\left\{\mathbf{e}_k^h\mathbf{\Phi}_{\mathbf{g}\mathbf{g}^t}(\mathbf{y})\mathbf{e}_l\mathbf{e}_k^t\mathbf{\Phi}^*_{\mathbf{g}\mathbf{g}^t}(\mathbf{y})\mathbf{e}_l\right\}\left[\mathbf{\Sigma}_H^2\right]_{ll}\end{aligned} \tag{94}$$

where $\mathbf{\Phi}_{\mathbf{g}\mathbf{g}^h}(\mathbf{y})=\mathbf{V}_G^h\mathbf{\Phi}_{\mathbf{x}\mathbf{x}^h}(\mathbf{y})\mathbf{V}_G$ with $\mathbf{g}=\mathbf{V}_G^h\mathbf{x}$, and $\mathbf{\Phi}_{\mathbf{g}\mathbf{g}^t}(\mathbf{y})=\mathbf{V}_G^h\mathbf{\Phi}_{\mathbf{x}\mathbf{x}^t}(\mathbf{y})\mathbf{V}_G^*$.

Thus, the Hessian of mutual information with respect to $\mathbf{\Sigma}_G^2$ is given by

$$\begin{aligned}\mathcal{H}_{\mathbf{\Sigma}_G^2,\mathbf{\Sigma}_G^2}\mathcal{I}(\mathbf{x};\mathbf{y})=&-\frac{1}{\sigma^2}\mathbf{\Sigma}_H^2E\left\{\mathbf{\Phi}_{\mathbf{g}\mathbf{g}^h}(\mathbf{y})\odot\mathbf{\Phi}^*_{\mathbf{g}\mathbf{g}^h}(\mathbf{y})\right\}\mathbf{\Sigma}_H^2\\ &-\frac{1}{\sigma^2}\mathbf{\Sigma}_H^2E\left\{\mathbf{\Phi}_{\mathbf{g}\mathbf{g}^t}(\mathbf{y})\odot\mathbf{\Phi}^*_{\mathbf{g}\mathbf{g}^t}(\mathbf{y})\right\}\mathbf{\Sigma}_H^2\end{aligned} \tag{95}$$

which is identical to (20).

Utilizing the identity $\mathbf{r}\mathbf{s}^h\odot(\mathbf{r}\mathbf{s}^h)^*=[\mathbf{r}\odot\mathbf{r}^*][\mathbf{s}\odot\mathbf{s}^*]^h$ with $\mathbf{r}$ and $\mathbf{s}$ being complex-valued column vectors, one can obtain the following:

$$\mathbf{\Phi}_{\mathbf{g}\mathbf{g}^h}(\mathbf{y})\odot\mathbf{\Phi}^*_{\mathbf{g}\mathbf{g}^h}(\mathbf{y})=\mathbf{Z}\mathbf{Z}^h \tag{96a}$$

$$\mathbf{\Phi}_{\mathbf{g}\mathbf{g}^t}(\mathbf{y})\odot\mathbf{\Phi}^*_{\mathbf{g}\mathbf{g}^t}(\mathbf{y})=\mathbf{Z}\mathbf{Z}^t \tag{96b}$$

where $\mathbf{Z}=E\{[(\mathbf{g}-E\{\mathbf{g}|\mathbf{y}\})\odot(\mathbf{g}-E\{\mathbf{g}|\mathbf{y}\})^*]|\mathbf{y}\}$.

Substituting the above two equations to (95), one can easily conclude that $\mathcal{H}_{\boldsymbol{\Sigma}_G^2,\boldsymbol{\Sigma}_G^2}\mathcal{I}(\mathbf{x};\mathbf{y})$ is negative (semi)-definite. Therefore, the mutual information $\mathcal{I}(\mathbf{x};\mathbf{y})$ is a concave function of $\boldsymbol{\Sigma}_G^2$, which completes the proof of Theorem 2. □

APPENDIX B
PROOF FOR THE SMALLEST INTERVALS LISTED IN TABLE I

It is known [30] that 2×2 unitary matrix $\boldsymbol{\Theta}$ can be split into

$$\boldsymbol{\Theta} = \mathbf{D}\begin{bmatrix}\cos(\omega) & \sin(\omega)e^{-j\nu}\\ -\sin(\omega)e^{j\nu} & \cos(\omega)\end{bmatrix}$$

where $\mathbf{D}$ is a 2×2 diagonal unitary matrix, $\omega\in[-\pi,\pi)$, and $\nu\in\left[-\frac{\pi}{2},\frac{\pi}{2}\right]$. This unitary matrix can also be equivalently decomposed into $\boldsymbol{\Theta}=\tilde{\mathbf{D}}\mathbf{U}_\omega\mathbf{U}_\nu$ with $\tilde{\mathbf{D}}$ being another 2×2 diagonal unitary matrix, $\mathbf{U}_\omega$ and $\mathbf{U}_\nu$ given by

$$\mathbf{U}_\omega = \begin{bmatrix}\cos(\omega) & \sin(\omega)\\ -\sin(\omega) & \cos(\omega)\end{bmatrix},\quad \mathbf{U}_\nu = \begin{bmatrix}1 & 0\\ 0 & e^{-j\nu}\end{bmatrix} \tag{97}$$

where the intervals of ω and ν are changed to $\omega\in\left[0,\frac{\pi}{2}\right]$ and $\nu\in[0,2\pi)$.

Since $\mathbf{v}$ and $\tilde{\mathbf{D}}\mathbf{v}$ are statistically equivalent when $\mathbf{v}$ is circularly symmetric zero-mean complex Gaussian noise, the mutual information $\mathcal{I}(\mathbf{x};\mathbf{y})$ can be reexpressed by

$$\begin{aligned}\mathcal{I}(\mathbf{x};\mathbf{y}) &\equiv \mathcal{I}(\mathbf{x};\overline{\mathbf{y}})\\ &= N_t\log_2 M\\ &\quad -\frac{1}{M^{N_t}}\sum_{m=1}^{M^{N_t}} E_{\mathbf{v}}\left\{\log_2\sum_{k=1}^{M^{N_t}} e^{-\frac{f_{m,k}-\|\mathbf{v}\|^2}{\sigma^2}}\right\}\end{aligned} \tag{98}$$

with

$$f_{m,k} = \left\|\boldsymbol{\Sigma}_H\boldsymbol{\Sigma}_G\mathbf{U}_\omega\mathbf{U}_\nu[s_1,s_2]^t+\mathbf{v}\right\|^2 \tag{99}$$

where $[s_1,s_2]^t=\left[s_1^{(m,k)},s_2^{(m,k)}\right]^t=\mathbf{x}_m-\mathbf{x}_k$ are difference vectors for $m\neq k$. As a minor notation abuse, we have dropped the superscript (m,k) index for simplicity.

From (97), we can prove the following:

$$\mathbf{U}_\omega\mathbf{U}_\nu[s_1,s_2]^t = \mathrm{diag}\{e^{-j\nu},-e^{-j\nu}\}\mathbf{U}_{\frac{\pi}{2}-\omega}\mathbf{U}_\nu^*[s_2,s_1]^t \tag{100}$$

$$\begin{aligned}&\left\|\boldsymbol{\Sigma}_H\boldsymbol{\Sigma}_G\mathbf{U}_\omega\mathbf{U}_\nu[s_1,s_2]^t+\mathbf{v}\right\|^2\\ &\quad=\left\|\boldsymbol{\Sigma}_H\boldsymbol{\Sigma}_G\mathbf{U}_{\frac{\pi}{2}-\omega}\mathbf{U}_\nu^*[s_2,s_1]^t+\hat{\mathbf{v}}\right\|^2\\ &\quad=\left\|\boldsymbol{\Sigma}_H\boldsymbol{\Sigma}_G\mathbf{U}_{\frac{\pi}{2}-\omega}\mathbf{U}_\nu\left[s_2^*,s_1^*\right]^t+\hat{\mathbf{v}}^*\right\|^2\end{aligned} \tag{101}$$

with $\hat{\mathbf{v}}=\mathrm{diag}\{e^{j\nu},-e^{j\nu}\}\mathbf{v}$.

Furthermore, for MPSK and M-QAM modulations, if $[s_1,s_2]^t$ is a valid difference vector, then $[s_2,s_1]^t$, $[s_2^*,s_1^*]^t$ and $[s_1^*,s_2^*]^t$ are also valid difference vectors. Therefore, the mutual information computed from (98) and (99) is the same as that computed from (98) by replacing ω with $\frac{\pi}{2}-\omega$. This means that, if ω is used, then $\frac{\pi}{2}-\omega$ is not needed in computing the mutual information. Therefore, the interval $\left[0,\frac{\pi}{2}\right]$ of ω can be reduced to $\left[0,\frac{\pi}{4}\right]$.

Similarly, we can proof the interval reduction for ν, as briefly outlined below.

From (97), we have $\mathbf{U}_\omega\mathbf{U}_\nu[s_1,s_2]^t=\mathbf{U}_\omega\mathbf{U}_{\nu+\pi}[s_1,-s_2]^t$. For BPSK, if $[s_1,s_2]^t$ is a valid difference vector, then $[s_1,-s_2]^t$ is also a valid difference vector. Therefore, if ν is used for computing the mutual information, then $\nu+\pi$ no longer needs to be tested. Hence the interval $[0,2\pi)$ of ν can be reduced to $[0,\pi]$. Moreover, we know that $\mathbf{U}_\omega\mathbf{U}_\nu[s_1,s_2]^t=\mathbf{U}_\omega\mathbf{U}_{\pi-\nu}^*[s_1,-s_2]^t$. This means, if ν is used, then $\pi-\nu$ is no longer needed for computing the mutual information. Thus, the interval of ν for BPSK can be further reduced to $\left[0,\frac{\pi}{2}\right]$.

It is known that QPSK possesses the rotational properties of BPSK, and additionally, if $[s_1,s_2]^t$ is a valid difference vector, then $\left[s_1,s_2e^{\frac{j\pi}{2}}\right]^t$, $\left[s_1,s_2e^{-\frac{j\pi}{2}}\right]^t$ and $\left[s_1^*,s_2^*e^{\frac{j\pi}{2}}\right]^t$ are also valid difference vectors. Utilizing the equalities $\mathbf{U}_\omega\mathbf{U}_\nu[s_1,s_2]^t=\mathbf{U}_\omega\mathbf{U}_{\nu+\frac{\pi}{2}}\left[s_1,s_2e^{\frac{j\pi}{2}}\right]^t$ and $\mathbf{U}_\omega\mathbf{U}_\nu[s_1,s_2]^t=\mathbf{U}_\omega\mathbf{U}_{\frac{\pi}{2}-\nu}^*\left[s_1,s_2e^{-\frac{j\pi}{2}}\right]^t$, we can find the smallest interval of ν for QPSK to be $\left[0,\frac{\pi}{4}\right]$.

Similarly, we can find the smallest interval of ν for 8PSK to be $\left[0,\frac{\pi}{8}\right]$ by recognizing that $[s_1,s_2]^t$, $\left[s_1,s_2e^{\frac{j\pi}{4}}\right]^t$, $\left[s_1,s_2e^{-\frac{j\pi}{4}}\right]^t$ and $\left[s_1^*,s_2^*e^{\frac{j\pi}{4}}\right]^t$ are valid difference vectors.

Because the (16, 64, 256) QAM constellations share the same $\frac{\pi}{2}$ rotational invariance property as QPSK, they have the same smallest interval for ν as QPSK. This completes the proof. □

ACKNOWLEDGMENT

The authors are grateful to Weiliang Zeng and Mingxi Wang for pointing out [24], [33] and for drafting Matlab codes helpful for generating Fig. 4.

REFERENCES

[1] E. Visotsky and U. Madhow, "Space-time transmit precoding with imperfect feedback," *IEEE Trans. Inf. Theory*, vol. 47, pp. 2634–2639, Sep. 2001.

[2] 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, pp. 1051–1064, May 2002.

[3] B. Hassibi and B. M. Hochwald, "High-rate codes that are linear in space and time," *IEEE Trans. Inf. Theory*, vol. 48, pp. 1804–1824, Jul. 2002.

[4] Y. Xin, Z. Wang, and G. B. Giannakis, "Space-time diversity systems based on linear constellation precoding," *IEEE Trans. Wireless Commun.*, vol. 2, pp. 294–309, Mar. 2003.

[5] D. P. Palomar, J. M. Cioffi, and M. A. Lagunas, "Joint Tx-Rx beamforming design for multicarrier MIMO channels: A unified framework for convex optimizaiton," *IEEE Trans. Signal Process.*, vol. 51, pp. 2381–2401, Sep. 2003.

[6] D. J. Love and R. W. Heath, Jr., "Multimode precoding for MIMO wireless systems," *IEEE Trans. Signal Process.*, vol. 53, pp. 3674–3687, Oct. 2005.

[7] Y. Jiang, J. Li, and W. W. Hager, "Joint transceiver design for MIMO communications using geometric mean decomposition," *IEEE Trans. Signal Process.*, vol. 53, pp. 3791–3803, Oct. 2005.

[8] R. Chen, R. W. Heath, Jr., and J. G. Andrews, "Transmit selection diversity for unitary precoded multiuser spatial multiplexing systems with linear receivers," *IEEE Trans. Signal Process.*, vol. 55, pp. 1159–1171, Mar. 2007.

[9] E. Yoon, J. Hansen, and A. Paulraj, "Space-frequency precoding with space-tap correlation information at the transmitter," *IEEE Trans. Commun.*, vol. 55, pp. 1702–1711, Sep. 2007.

[10] S. Shi, M. Schubert, and H. Boche, "Downlink MMSE transceiver optimization for multiuser MIMO systems: Duality and sum-MSE minimization," *IEEE Trans. Signal Process.*, vol. 55, pp. 5436–5446, Nov. 2007.

[11] G. Scutari, D. P. Palomar, and S. Barbarossa, "Optimal linear precoding strategies for wideband noncooperative systems based on game theory," *IEEE Trans. Signal Process.*, vol. 56, pp. 1230–1267, Mar. 2008.

[12] G. Yue, X. Wang, and M. Madihian, "Outage performance and IRA code design for MIMO block fading with unitary scrambling," *IEEE Trans. Veh. Technol.*, vol. 57, pp. 2354–2363, Jul. 2008.

[13] X. Zhang, D. P. Palomar, and B. Ottersten, "Statistically robust design of linear MIMO transceivers," *IEEE Trans. Signal Process.*, vol. 56, pp. 3678–3689, Aug. 2008.

[14] M. Ghogho, T. Whitworth, A. Swami, and D. McLernon, "Full-rank and rank-deficient precoding schemes for single-carrier block transmissions," *IEEE Trans. Signal Process.*, vol. 57, pp. 4433–4442, Nov. 2009.

[15] F. Rey, M. Lamarca, and G. Vazquez, "Linear precoder design through cut-off rate maximization in MIMO-OFDM coded systems with imperfect CSIT," *IEEE Trans. Signal Process.*, vol. 58, pp. 1741–1755, Mar. 2010.

[16] P. Xiao and M. Sellathurai, "Improved linear transmit processing for single-user and multi-user MIMO communications systems," *IEEE Trans. Signal Process.*, vol. 58, pp. 1768–1779, Mar. 2010.

[17] M. R. Bhatnagar and A. Hjørungnes, "Linear precoding of STBC over correlated ricean MIMO channels," *IEEE Trans. Wireless Commun.*, vol. 9, pp. 1832–1836, June 2010.

[18] V. Tarokh, N. Seshadri, and A. R. Calderbank, "Space-time codes for high data rate wireless communication: Performance criterion and code construction," *IEEE Trans. Inf. Theory*, vol. 44, pp. 744–765, Mar. 1998.

[19] H. Jafarkhani, *Space-Time Coding: Theory and Practice*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[20] A. Lozano, A. M. Tulino, and S. Verdú, "Optimum power allocation for parallel Gaussian channels with arbitrary input distributions," *IEEE Trans. Inf. Theory*, vol. 52, pp. 3033–3051, Jul. 2006.

[21] C. Xiao and Y. R. Zheng, "On the mutual information and power allocation for vector Gaussian channels with finite discrete inputs," in *Proc. IEEE Globecom'08*, New Orleans, LA, Nov. 30–Dec. 4 2008.

[22] F. Pérez-Cruz, M. R. D. Rodrigues, and S. Verdú, "MIMO gaussian channels with arbitrary inputs: Optimal precoding and power allocations," *IEEE Trans. Inf. Theory*, vol. 56, pp. 1070–1084, Mar. 2010.

[23] M. Payaró and D. P. Palomar, "On optimal precoding in linear vector Gaussian channels with arbitrary input distribution," in *Proc. IEEE Int. Symp. Inf. Theory*, 2009, pp. 1085–1089.

[24] M. Lamarca, "Linear precoding for mutual information maximization in MIMO systems," in *Proc. Int. Symp. Wireless Commun. Syst.*, 2009, pp. 26–30.

[25] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York: Wiley, 2006.

[26] A. Hjørungnes and D. Gesbert, "Complex-valued matrix differentiation: Techniques and key results," *IEEE Trans. Signal Process.*, vol. 55, pp. 2740–2746, June 2007, (for an extensively expanded version, readers are referred to A. Hjrungnes, *Complex-Valued Matrix Derivatives*, Cambridge Univ. Press, 2011).

[27] J. R. Magnus and H. Neudecker, *Matrix Differential Calculus With Applications in Statistics and Econometrics*, 2nd ed. New York: Wiley, 1999.

[28] M. Payaró and D. P. Palomar, "Hessian and concavity of mutual information, differential entropy, and entropy power in linear vector Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 55, pp. 3613–3628, Aug. 2009.

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

[30] F. D. Murnaghan, *Lectures on Applied Mathematics, Vol. III: The Unitary and Rotation Groups*. Wash., DC: Spartan, 1962.

[31] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed. New York: Athena Scientific, 2008, 3rd printing with corrections.

[32] D. P. Palomar and S. Verdu, "Gradient of mutual information in linear vector Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 52, pp. 141–154, Jan. 2006.

[33] M. Valenti, The Coded Modulation Library [Online]. Available: http://www.iterativesolutions.com/

[34] S. M. Kay, *Fundamentals of Statistical Signal Processing: Detection Theory*. Engelwood Cliffs, NJ: Prentice-Hall, 1998.

Chengshan Xiao (M'99–SM'02–F'10) received the B.S. degree in electronic engineering from the University of Electronic Science and Technology of China, Chengdu, in 1987, the M.S. degree in electronic engineering from Tsinghua University, Beijing, China, in 1989, and the Ph.D. degree in electrical engineering from the University of Sydney, Sydney, Australia, in 1997.

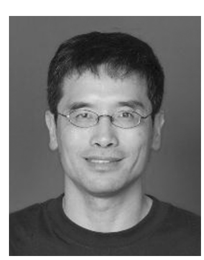
From 1989 to 1993, he was with the Department of Electronic Engineering, Tsinghua University, where he was a member of Research Staff and then a Lecturer. From 1997 to 1999, he was a Senior Member of Scientific Staff with Nortel, Ottawa, ON, Canada. From 1999 to 2000, he was a Faculty Member with the University of Alberta, Edmonton, AB, Canada. From 2000 to 2007, he was with the University of Missouri, Columbia, where he was an Assistant Professor and then an Associate Professor. He is currently a Professor with the Department of Electrical and Computer Engineering, Missouri University of Science and Technology, Rolla (formerly University of Missouri, Rolla). His research interests include wireless communications, signal processing, and underwater acoustic communications. He is the holder of three U.S. patents. His algorithms have been implemented into Nortel's base station radios after successful technical field trials and network integration.

Dr. Xiao is the Editor-in-Chief of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, and a Distinguished Lecturer of the IEEE Communications Society. Previously, he served as the founding Area Editor for Transmission Technology of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS; an Associate Editor of the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS-I, and the international journal of *Multidimensional Systems and Signal Processing*. He was the Technical Program Chair of the 2010 IEEE International Conference on Communications (ICC), the Lead Co-Chair of the 2008 IEEE ICC Wireless Communications Symposium, and a Phy/MAC Program Co-Chair of the 2007 IEEE Wireless Communications and Networking Conference. He served as the founding Chair of the IEEE Technical Committee on Wireless Communications, and the Vice-Chair of the IEEE Technical Committee on Personal Communications.

Yahong Rosa Zheng (S'99–M'03–SM'07) received the B.S. degree from the University of Electronic Science and Technology of China, Chengdu, China, in 1987, and the M.S. degree from Tsinghua University, Beijing, China, in 1989, both in electrical engineering. She received the Ph.D. degree from the Department of Systems and Computer Engineering, Carleton University, Ottawa, Canada, in 2002.

She was an NSERC Postdoctoral Fellow from 2003 to 2005 with the University of Missouri-Columbia. Since 2005, she has been a faculty member with the Department of Electrical and Computer Engineering, Missouri University of Science and Technology, Rolla. Her research interests include array signal processing, wireless communications, and wireless sensor networks.

Dr. Zheng has served as a Technical Program Committee (TPC) member for many IEEE international conferences, including VTC, GlobeCom, ICC, WCNC, and Aerospace Conference, etc. She served as an Associate Editor for IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS during 2006–2008. She is the recipient of an NSF CAREER award in 2009.



Zhi Ding (S'88–M'90–SM'95–F'03) received the Ph.D. degree in electrical engineering from Cornell University, Ithaca, NY, in 1990.

He is currently Child Family Endowed Professor of Engineering and Entrepreneurship at the University of California, Davis. From 1990 to 2000, he was a faculty member of Auburn University and later, the University of Iowa, Ames. He has held visiting positions with the Australian National University, Hong Kong University of Science and Technology, NASA Lewis Research Center, and USAF Wright Laboratory. He was named Visiting Professor of Northwestern Polytechnical University, Xi'an, China, and also received a 1000 Plan appointment from Southeast University, Nanjing, China. He has active collaboration with researchers from several countries including Australia, China, Japan, Canada, Taiwan, Korea, Singapore, and Hong Kong. He is a coauthor of the text, *Modern Digital and Analog Communication Systems*, 4th ed., (Oxford, U.K.: Oxford Univ. Press, 2009).

Dr. Ding is an active member of the IEEE, serving on Technical Programs of several workshops and conferences. He was Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING from 1994–1997, 2001–2004, and Associate Editor of IEEE SIGNAL PROCESSING LETTERS during 2002–2005. He was a member of the Technical Committee on Statistical Signal and Array Processing and a member of the Technical Committee on Signal Processing for Communications (1994–2003). He was the Technical Program Chair of the 2006 IEEE Globecom. He is also an IEEE Distinguished Lecturer (Circuits and Systems Society, 2004–06, Communications Society, 2008–09).