Gradient of Mutual Information in Linear Vector Gaussian Channels

Daniel P. Palomar, *Member, IEEE,* and Sergio Verdú, *Fellow, IEEE*

***Abstract*—This paper considers a general linear vector Gaussian channel with arbitrary signaling and pursues two closely related goals: i) closed-form expressions for the gradient of the mutual information with respect to arbitrary parameters of the system, and ii) fundamental connections between information theory and estimation theory. Generalizing the fundamental relationship recently unveiled by Guo, Shamai, and Verdú, we show that the gradient of the mutual information with respect to the channel matrix is equal to the product of the channel matrix and the error covariance matrix of the best estimate of the input given the output. Gradients and derivatives with respect to other parameters are then found via the differentiation chain rule.**

***Index Terms*—De Bruijn's identity, divergence, Gaussian noise, minimum mean-square error (MMSE), multiple-input multiple-output (MIMO) channels, mutual information, nonlinear estimation, precoder optimization.**

I. Introduction and Motivation

THIS paper considers general linear vector channels with Gaussian noise and arbitrary input distributions. Our purpose is twofold: i) to find closed-form expressions for the gradient of the mutual information with respect to arbitrary parameters of the system, and ii) to explore the fundamental connections between information theory and estimation theory. In fact, both goals are achieved simultaneously since the gradient of the mutual information happens to be directly related to the performance of the conditional mean estimator.

Closed-form expressions for the gradient of the mutual information with respect to arbitrary parameters are useful in both analysis and design. The most direct application is to analyze and understand the sensitivity and robustness of a system to variations in certain parameters. Learning the weaknesses of a system may teach us how to make it more robust. Indeed, an engineer may have the freedom to modify or even to design a specific part of the system. In such a case, the availability of expressions for the gradient of an objective function with respect to the design parameters is of paramount importance to optimize the system. One common example is the design of a transmit precoder for a given communication system; in particular, the precoder can be flexible and adapt to the channel realization to increase the system performance [1].

Early connections between fundamental quantities in information theory and estimation theory were De Bruijn's identity, used by Stam [2] to prove the entropy-power inequality, the representation of mutual information as a function of causal filtering error by Duncan [3] and by Kadota, Zakai, and Ziv [4], and the rate-distortion bounds of Ziv and Zakai [5]. Recently, a fundamental relation between the mutual information and the minimum mean-square error (MMSE) was unveiled in [6] for discrete-time and continuous-time, scalar and vector channels with Gaussian noise. In the special case of the scalar Gaussian channel $y = \sqrt{\mathsf{snr}}x + n$ (complex-valued inputs/outputs are considered throughout this paper) and regardless of the input distribution:

$$\frac{d}{d\mathsf{snr}} I(x; \sqrt{\mathsf{snr}}x + n) = \mathsf{mmse}(\mathsf{snr}) \tag{1}$$

where mutual information is in nats and $\mathsf{mmse}(\mathsf{snr})$ is the MMSE corresponding to the best estimation of x upon the observation y for a given signal-to-noise ratio (SNR) snr, i.e.,

$$\mathsf{mmse}(\mathsf{snr}) = \mathbb{E}[|x - \mathbb{E}[x \,|\, \sqrt{\mathsf{snr}}\, x + n]|^2]. \tag{2}$$

An extension of (1) to the vector case was also given in [6] for the linear vector Gaussian channel $\boldsymbol{y} = \sqrt{\mathsf{snr}}\, \boldsymbol{H}\boldsymbol{x} + \boldsymbol{n}$ as

$$\frac{d}{d\,\mathsf{snr}} I(\boldsymbol{x}; \sqrt{\mathsf{snr}}\, \boldsymbol{H}\boldsymbol{x} + \boldsymbol{n}) = \mathbb{E}[\|\boldsymbol{H}\boldsymbol{x} - \mathbb{E}[\boldsymbol{H}\boldsymbol{x} \,|\, \sqrt{\mathsf{snr}}\, \boldsymbol{H}\boldsymbol{x} + \boldsymbol{n}]\|^2] \tag{3}$$

where the right-hand side is the expected squared Euclidean norm of the error in the estimation of $\boldsymbol{H}\boldsymbol{x}$ (rather than $\boldsymbol{x}$).

The representation of mutual information in (1) as an integral of $\mathsf{mmse}(\mathsf{snr})$ has found various interesting applications (cf. [6]). The derivative of the mutual information with respect to the signal-to-noise ratio (SNR) in a communication system is clearly a useful quantity for an engineer. In addition, it may also be of interest to generalize such sensitivity analysis to arbitrary parameters of the system rather than just the SNR. In particular, to obtain a full description of the sensitivity of the mutual information, one should obtain the partial derivatives with respect to arbitrary parameters affecting each combination of transmit–receive dimension. A compact way to describe the sensitivity of mutual information for the linear vector channel with independent Gaussian noise $\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{n}$ is via the gradient with respect

Manuscript received April 29, 2005; revised August 4, 2005. This work was supported in part by the Fulbright Program and the Ministry of Education and Science of Spain; the U.S. National Science Foundation under Grant NCR-0074277; and through collaborative participation in the Communications and Networks Consortium sponsored by the U.S. Army Research Laboratory under the Collaborative Technology Alliance Program, Cooperative Agreement DAAD19-01-2-0011. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation thereon.

The authors are with the Department of Electrical Engineering, Princeton University, Engineering Quadrangle, Princeton, NJ 08544 USA (e-mail: danielp@princeton.edu; verdu@princeton.edu).

Communicated by Y. Steinberg, Associate Editor for Shannon Theory.

Digital Object Identifier 10.1109/TIT.2005.860424

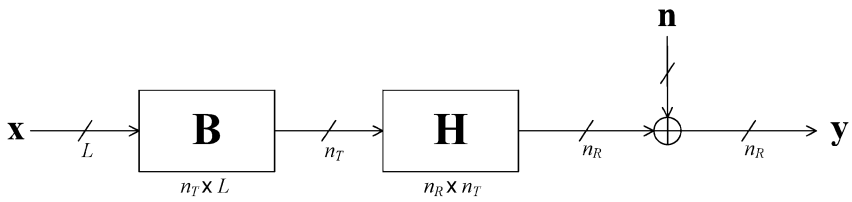

Fig. 1. Linear vector channel including linear precoding.

to the deterministic matrix $\boldsymbol{H}$. The main result of this paper is the following formula for the gradient:

$$\nabla_{\boldsymbol{H}} I(\boldsymbol{x}; \boldsymbol{H}\boldsymbol{x} + \boldsymbol{n}) = \boldsymbol{H}\boldsymbol{E} \tag{4}$$

where $\boldsymbol{E}$ is the covariance matrix of the estimation error vector, sometimes referred to as the MMSE matrix [7]. The MMSE matrix is the full generalization of the scalar MMSE in (1) to the vector case. As we will see, applying the chain rule for differentiation to the basic gradient in (4), one can obtain the sensitivity of the mutual information with respect to any arbitrary parameter such as the SNR, the transmit covariance matrix, or a precoding matrix.

The paper is organized as follows. Section II describes the signal model for the general linear vector Gaussian channel, along with several particularizations. The main results of the paper are given in Section III: the relation between various gradients of the mutual information and the MMSE matrix. Section IV contains a number of gradients of information measures (namely, particularization of the general result to SNR gradients, generalization of De Bruijn's identity, gradient of the divergence between the conditional and unconditional outputs, gradient of the divergence between the signal-plus-noise and noise distributions, and gradient of the non-Gaussianness). Section V illustrates the potential of the theoretical results with a practical application consisting of the numerical optimization of a linear precoder. Section VI draws the final conclusions and summarizes the paper.

Notation: Boldface upper case letters denote matrices, boldface lower case letters denote column vectors, and italics denote scalars. The superscripts $(\cdot)^T, (\cdot)^*$, and $(\cdot)^\dagger$ denote transpose, complex conjugate, and Hermitian operations, respectively. $[\boldsymbol{X}]_{ij}$ denotes the (ith, jth) element of matrix $\boldsymbol{X}$ and x_i denotes the ith element of vector $\boldsymbol{x}$. $\mathrm{Tr}(\cdot)$ and $\det(\cdot)$ denote the trace and determinant of a matrix, respectively. $\|\boldsymbol{X}\|$ denotes the Frobenius norm of matrix $\boldsymbol{X}$ and is given by $\|\boldsymbol{X}\| = \sqrt{\mathrm{Tr}(\boldsymbol{X}\boldsymbol{X}^\dagger)}$.

II. Signal Model

Consider a general discrete-time linear vector Gaussian channel represented by the following vector signal model with n_T transmit dimensions and n_R receive dimensions

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{n} \tag{5}$$

where all quantities are complex-valued, $\boldsymbol{x}$ is the n_T-dimensional transmitted vector, $\boldsymbol{H}$ is the $n_R \times n_T$ matrix that denotes the linear transformation undergone by the signal, $\boldsymbol{y}$ is the n_R-dimensional received vector, and $\boldsymbol{n}$ is an n_R-dimensional proper complex Gaussian noise vector independent of $\boldsymbol{x}$. The input and the noise covariance matrices are

$$\boldsymbol{\Sigma}_x = \mathbb{E}[(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])^\dagger] \tag{6}$$

and

$$\boldsymbol{\Sigma}_n = \mathbb{E}[(\boldsymbol{n} - \mathbb{E}[\boldsymbol{n}])(\boldsymbol{n} - \mathbb{E}[\boldsymbol{n}])^\dagger]. \tag{7}$$

The general channel model in (5) describes many different communication systems, e.g., wireless multiple-antenna systems, code-division multiple-access (CDMA) systems, wireline digital subscriber line systems, or even single-antenna frequency-selective wideband channels.

It is interesting to further generalize the model in (5) to include an additional linear transformation represented by the $n_T \times L$ matrix $\boldsymbol{B}$, where L is now the dimension of the data vector $\boldsymbol{x}$ (Fig. 1)

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{B}\boldsymbol{x} + \boldsymbol{n}. \tag{8}$$

The additional matrix $\boldsymbol{B}$ can play different roles: i) in a wireless multiple-antenna system it may represent a *beamforming matrix* that uses some knowledge about the physical channel $\boldsymbol{H}$ to properly steer the transmitted signal through the best channel eigenmodes [8], [9]; ii) $\boldsymbol{B}$ can denote a *linear precoding matrix* or *shaping matrix* that adapts or shapes the transmitted signal to the channel realization [1], [10]–[12], [9]; iii) the overall input-output linear transformation may factor into two matrices $\boldsymbol{H}$ and $\boldsymbol{B}$ (one of which may be controllable by the designer of the system).

A special case of (8) by setting $\boldsymbol{B} = \sqrt{\mathsf{snr}}\boldsymbol{I}$ yields the model used in (3)

$$\boldsymbol{y} = \sqrt{\mathsf{snr}}\,\boldsymbol{H}\boldsymbol{x} + \boldsymbol{n}. \tag{9}$$

The output conditional probability density function (pdf) corresponding to the linear vector Gaussian model in (8) (assuming a zero-mean noise without loss of generality) is

$$p_{\boldsymbol{y}\,|\,\boldsymbol{x}}(\boldsymbol{y}) = \frac{1}{\det(\pi\boldsymbol{\Sigma}_n)} \exp\Big(-(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{B}\boldsymbol{x})^\dagger \boldsymbol{\Sigma}_n^{-1} (\boldsymbol{y} - \boldsymbol{H}\boldsymbol{B}\boldsymbol{x})\Big). \tag{10}$$

Consider the estimation of the input signal $\boldsymbol{x}$ based on the observation of the output $\boldsymbol{y}$. In the scalar case, the mean-square error (MSE) of an estimate $\hat{x}(y)$ of the input x based on the observation y is defined as $\mathbb{E}[|x - \hat{x}(y)|^2]$, which is the variance of the estimator error $x - \hat{x}(y)$ provided that the estimator is unbiased. The conditional mean $\hat{x}(y) = \mathbb{E}[x\,|\,y]$ achieves the

MMSE and is referred to as *MMSE estimator*. In the more general vector setup, the MMSE estimator $\hat{\boldsymbol{x}}(\boldsymbol{y})$ achieves simultaneously the MMSE for all components of the estimation error vector and is again given by the conditional mean estimator

$$\hat{\boldsymbol{x}}(\boldsymbol{y}) = \mathbb{E}[\boldsymbol{x}\,|\,\boldsymbol{y}]. \tag{11}$$

The full description of the performance of the vector MMSE estimator is given by the *MMSE matrix*, i.e., the covariance of the estimation error vector

$$\boldsymbol{E} \triangleq \mathbb{E}[(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}\,|\,\boldsymbol{y}])(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}\,|\,\boldsymbol{y}])^{\dagger}]. \tag{12}$$

An alternative limited description of the performance, which is sometimes useful, is given by the mean-squared norm of the estimation error $\mathbb{E}[\|\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}\,|\,\boldsymbol{y}]\|^2]$, which is equal to the trace of the MMSE matrix.

III. Main Results

A. Basic Result: Gradient With Respect to $\boldsymbol{H}$

To start with, consider the signal model in (5) with *Gaussian signaling*, where the covariance matrix of the transmitted signal $\boldsymbol{x}$ is denoted by $\boldsymbol{\Sigma}_x$ and the covariance matrix of the noise is the identity matrix. The mutual information, $I(\boldsymbol{x};\boldsymbol{y})$, is[1] (e.g., [13])

$$I = \log\det(\boldsymbol{I} + \boldsymbol{H}\boldsymbol{\Sigma}_x\boldsymbol{H}^{\dagger}). \tag{13}$$

The MMSE estimator is

$$\hat{\boldsymbol{x}} = \boldsymbol{\Sigma}_x\boldsymbol{H}^{\dagger}(\boldsymbol{I} + \boldsymbol{H}\boldsymbol{\Sigma}_x\boldsymbol{H}^{\dagger})^{-1}\boldsymbol{y} \tag{14}$$

(assuming $\boldsymbol{x}$ and $\boldsymbol{n}$ with zero mean), which is linear since the input is Gaussian, and the MMSE matrix is (e.g., [7, Theorem 11.1]

$$\boldsymbol{E} = \left(\boldsymbol{\Sigma}_x^{-1} + \boldsymbol{H}^{\dagger}\boldsymbol{H}\right)^{-1}. \tag{15}$$

Taking the gradient[2] of (13) we obtain

$$\begin{aligned}\nabla_{\boldsymbol{H}} I &= \nabla_{\boldsymbol{H}} \log\det(\boldsymbol{I} + \boldsymbol{H}^{\dagger}\boldsymbol{H}\boldsymbol{\Sigma}_x) \\ &= \boldsymbol{H}\boldsymbol{\Sigma}_x(\boldsymbol{I} + \boldsymbol{H}^{\dagger}\boldsymbol{H}\boldsymbol{\Sigma}_x)^{-1} = \boldsymbol{H}\boldsymbol{E}.\end{aligned} \tag{16}$$

The main result of the paper shows that the Gaussian assumption on the input is unnecessary for (4) to hold.

Theorem 1: Consider the signal model in (5), where $\boldsymbol{H}$ is an arbitrary deterministic matrix, the signaling $\boldsymbol{x}$ is arbitrarily distributed (with finite second-order moments), and the noise $\boldsymbol{n}$ is Gaussian, independent of the input $\boldsymbol{x}$, with normalized covariance matrix $\boldsymbol{\Sigma}_n = \boldsymbol{I}$. Then, the mutual information I and the MMSE matrix $\boldsymbol{E}$ satisfy

$$\nabla_{\boldsymbol{H}} I(\boldsymbol{x};\boldsymbol{H}\boldsymbol{x} + \boldsymbol{n}) = \boldsymbol{H}\boldsymbol{E}. \tag{17}$$

Proof: See Appendix B. □

[1]Throughout this paper nats are used as the information units and $\log$ denotes natural logarithm.

[2]See Appendix A for some observations on the definition of complex gradient.

B. Gradient With Respect to Arbitrary Parameters

With the basic gradient result of Theorem 1, one can easily find the gradient with respect to arbitrary parameters through a gradient chain rule as stated next.

Lemma 1: Let f be a real-valued function which depends on θ through $\boldsymbol{H}$. The following chain rules hold.[3]

- If θ is complex, then

$$\nabla_{\theta} f = \mathrm{Tr}(\nabla_{\boldsymbol{H}} f \times \nabla_{\theta}\boldsymbol{H}^{\dagger}) + \mathrm{Tr}\left(\nabla_{\boldsymbol{H}}^{\dagger} f \times \nabla_{\theta}\boldsymbol{H}\right). \tag{18}$$

- If θ is real, then

$$\nabla_{\theta} f = 2\mathrm{Re}\mathrm{Tr}(\nabla_{\boldsymbol{H}} f \times \nabla_{\theta}\boldsymbol{H}^{\dagger}). \tag{19}$$

For example, the gradient of the mutual information along a specific direction given by $\boldsymbol{D}$ is

$$\left.\frac{d}{dt} I(\boldsymbol{x};(\boldsymbol{H} + t\boldsymbol{D})\boldsymbol{x} + \boldsymbol{n})\right|_{t=0} = 2\mathrm{Re}\mathrm{Tr}(\boldsymbol{D}^{\dagger}\boldsymbol{H}\boldsymbol{E}). \tag{20}$$

We now give the gradient of the mutual information with respect to several quantities of interest.

Theorem 2: Consider the general signal model in (8) including a linear precoder $\boldsymbol{B}$ at the transmitter, where $\boldsymbol{H}$ is an arbitrary deterministic matrix, the signaling $\boldsymbol{x}$ is arbitrarily distributed with covariance matrix $\boldsymbol{\Sigma}_x$, the noise $\boldsymbol{n}$ is Gaussian, independent of the input $\boldsymbol{x}$, and has positive definite covariance matrix $\boldsymbol{\Sigma}_n$, the transmit covariance matrix is $\boldsymbol{Q} = \boldsymbol{B}\boldsymbol{\Sigma}_x\boldsymbol{B}^{\dagger}$ (which includes the precoding), and the squared linear precoder is $\boldsymbol{Q}_B = \boldsymbol{B}\boldsymbol{B}^{\dagger}$. Then, the mutual information I and the MMSE matrix $\boldsymbol{E}$ satisfy

$$\nabla_{\boldsymbol{H}} I(\boldsymbol{x};\boldsymbol{H}\boldsymbol{B}\boldsymbol{x} + \boldsymbol{n}) = \boldsymbol{\Sigma}_n^{-1}\boldsymbol{H}\boldsymbol{B}\boldsymbol{E}\boldsymbol{B}^{\dagger} \tag{21}$$

$$\nabla_{\boldsymbol{B}} I(\boldsymbol{x};\boldsymbol{H}\boldsymbol{B}\boldsymbol{x} + \boldsymbol{n}) = \boldsymbol{H}^{\dagger}\boldsymbol{\Sigma}_n^{-1}\boldsymbol{H}\boldsymbol{B}\boldsymbol{E} \tag{22}$$

$$\nabla_{\boldsymbol{Q}} I(\boldsymbol{x};\boldsymbol{H}\boldsymbol{B}\boldsymbol{x} + \boldsymbol{n})\,\boldsymbol{B}\boldsymbol{\Sigma}_x = \boldsymbol{H}^{\dagger}\boldsymbol{\Sigma}_n^{-1}\boldsymbol{H}\boldsymbol{B}\boldsymbol{E} \tag{23}$$

$$\nabla_{\boldsymbol{Q}_B} I(\boldsymbol{x};\boldsymbol{H}\boldsymbol{B}\boldsymbol{x} + \boldsymbol{n})\boldsymbol{B} = \boldsymbol{H}^{\dagger}\boldsymbol{\Sigma}_n^{-1}\boldsymbol{H}\boldsymbol{B}\boldsymbol{E} \tag{24}$$

$$\nabla_{\boldsymbol{\Sigma}_x} I(\boldsymbol{x};\boldsymbol{H}\boldsymbol{B}\boldsymbol{x} + \boldsymbol{n})\boldsymbol{\Sigma}_x = \boldsymbol{B}^{\dagger}\boldsymbol{H}^{\dagger}\boldsymbol{\Sigma}_n^{-1}\boldsymbol{H}\boldsymbol{B}\boldsymbol{E} \tag{25}$$

$$\nabla_{\boldsymbol{\Sigma}_n^{-1}} I(\boldsymbol{x};\boldsymbol{H}\boldsymbol{B}\boldsymbol{x} + \boldsymbol{n}) = \boldsymbol{H}\boldsymbol{B}\boldsymbol{E}\boldsymbol{B}^{\dagger}\boldsymbol{H}^{\dagger} \tag{26}$$

$$\nabla_{\boldsymbol{\Sigma}_n} I(\boldsymbol{x};\boldsymbol{H}\boldsymbol{B}\boldsymbol{x} + \boldsymbol{n}) = -\boldsymbol{\Sigma}_n^{-1}\boldsymbol{H}\boldsymbol{B}\boldsymbol{E}\boldsymbol{B}^{\dagger}\boldsymbol{H}^{\dagger}\boldsymbol{\Sigma}_n^{-1}. \tag{27}$$

Proof: See Appendix C. □

C. Random Channels

We now consider the interesting scenario in which the channel matrix $\boldsymbol{H}$ is random. In such a case, we can still obtain expressions for gradients with respect to different parameters of the system (with the exception of the channel matrix $\boldsymbol{H}$, of course). It will be instrumental to make the distinction according to whether the channel is known or unknown at the receiver. For concreteness, we consider the signal model

$$\boldsymbol{y} = \boldsymbol{C}\boldsymbol{H}\boldsymbol{x} + \boldsymbol{n} \tag{28}$$

[3]For the sake of clarity, in (18)–(19) we indicate the product between matrices by $\times$.

where $\boldsymbol{H}$ is a random matrix with finite second-order moments, and compute the gradient with respect to $\boldsymbol{C}$. Note that for a deterministic channel matrix the gradient is

$$\nabla_{\boldsymbol{C}} I(\boldsymbol{x}; \boldsymbol{CHx}+\boldsymbol{n}) = \boldsymbol{\Sigma}_n^{-1}\boldsymbol{CHEH}^{\dagger}. \tag{29}$$

The case of channel known at the receiver reduces to taking expectations with respect to the channel matrix in (29) with the following result (cf. Appendix H)

$$\begin{aligned}&\nabla_{\boldsymbol{C}} I(\boldsymbol{x}; \boldsymbol{CHx}+\boldsymbol{n}\,|\,\boldsymbol{H})\\ &\quad=\boldsymbol{\Sigma}_n^{-1}\boldsymbol{C}\mathbb{E}[(\boldsymbol{Hx}-\mathbb{E}[\boldsymbol{Hx}\,|\,\boldsymbol{y},\boldsymbol{H}])(\boldsymbol{Hx}-\mathbb{E}[\boldsymbol{Hx}\,|\,\boldsymbol{y},\boldsymbol{H}])^{\dagger}]\end{aligned} \tag{30}$$

where the outer expectation is with respect to $\boldsymbol{x}, \boldsymbol{y}$, and $\boldsymbol{H}$. Gradients with respect to other parameters of the system can be similarly derived (except with respect to the random channel matrix) similarly to Theorem 2. As an illustrative example, we can consider the signal model $\boldsymbol{y} = \boldsymbol{Hx}+\boldsymbol{n}$ with $\boldsymbol{\Sigma}_n = N_0\boldsymbol{I}$ and obtain the first-order expansion of the mutual information with respect to N_0 as

$$I(\boldsymbol{x}; \boldsymbol{Hx}+\boldsymbol{n}\,|\,\boldsymbol{H}) = N_0^{-1}\mathrm{Tr}(\mathbb{E}[\boldsymbol{H\Sigma}_x\boldsymbol{H}^{\dagger}]) + o\left(N_0^{-1}\right) \tag{31}$$

which agrees with [14, Theorem 3].

The case of channel unknown at the receiver is significantly more complicated and can be successfully tackled using the following result for an arbitrary random transformation of the input, which generalizes Theorem 1 and [6, Theorem 10].

Theorem 3: Consider the following signal model including an arbitrary random transformation $\boldsymbol{z} = f(\boldsymbol{x})$ on the input (independent of the noise $\boldsymbol{n}$):

$$\boldsymbol{y} = \boldsymbol{H}f(\boldsymbol{x}) + \boldsymbol{n} \tag{32}$$

where all the terms are defined as in Theorem 1. Then, the gradient of the mutual information is

$$\begin{aligned}\nabla_{\boldsymbol{H}} I(\boldsymbol{x};\boldsymbol{y}) &= \boldsymbol{H}\mathbb{E}[(\boldsymbol{z}-\mathbb{E}[\boldsymbol{z}\,|\,\boldsymbol{y}])(\boldsymbol{z}-\mathbb{E}[\boldsymbol{z}\,|\,\boldsymbol{y}])^{\dagger}]\\ &\quad-\boldsymbol{H}\mathbb{E}[(\boldsymbol{z}-\mathbb{E}[\boldsymbol{z}\,|\,\boldsymbol{y},\boldsymbol{x}])(\boldsymbol{z}-\mathbb{E}[\boldsymbol{z}\,|\,\boldsymbol{y},\boldsymbol{x}])^{\dagger}].\end{aligned} \tag{33}$$

In addition, if the transformation f is deterministic, then the second term in the right-hand side of (33) vanishes, obtaining

$$\begin{aligned}&\nabla_{\boldsymbol{H}} I(\boldsymbol{x}; \boldsymbol{H}f(\boldsymbol{x})+\boldsymbol{n})\\ &\quad= \boldsymbol{H}\mathbb{E}[(f(\boldsymbol{x})-\mathbb{E}[f(\boldsymbol{x})\,|\,\boldsymbol{y}])(f(\boldsymbol{x})-\mathbb{E}[f(\boldsymbol{x})\,|\,\boldsymbol{y}])^{\dagger}]\end{aligned} \tag{34}$$

which happens to be equal to $\nabla_{\boldsymbol{H}} I(f(\boldsymbol{x}); \boldsymbol{H}f(\boldsymbol{x})+\boldsymbol{n})$.

Proof: The proof hinges on the fact that $\boldsymbol{x} \to \boldsymbol{z} \to \boldsymbol{y}$ is a Markov chain and then $I(\boldsymbol{x};\boldsymbol{y}) = I(\boldsymbol{z};\boldsymbol{y}) - I(\boldsymbol{z};\boldsymbol{y}\,|\,\boldsymbol{x})$ to which Theorem 1 can be applied. If f is deterministic, then $I(\boldsymbol{z};\boldsymbol{y}\,|\,\boldsymbol{x}) = 0$. □

Now, using Theorem 3, we can easily obtain the gradient for the case of channel unknown at the receiver (similarly to [6, Theorem 10] for the scalar case) by modeling the system as the Markov chain $\boldsymbol{x} \to \boldsymbol{z} \to \boldsymbol{y}$, where $\boldsymbol{z} = \boldsymbol{Hx}$ is the output of a random transformation applied to $\boldsymbol{x}$ that describes the effect to the random channel. The gradient is found as (cf. Appendix H)

$$\begin{aligned}&\nabla_{\boldsymbol{C}} I(\boldsymbol{x}; \boldsymbol{CHx}+\boldsymbol{n})\\ &\quad= \boldsymbol{\Sigma}_n^{-1}\boldsymbol{C}\mathbb{E}[(\boldsymbol{Hx}-\mathbb{E}[\boldsymbol{Hx}\,|\,\boldsymbol{y}])(\boldsymbol{Hx}-\mathbb{E}[\boldsymbol{Hx}\,|\,\boldsymbol{y}])^{\dagger}]\\ &\qquad-\boldsymbol{\Sigma}_n^{-1}\boldsymbol{C}\mathbb{E}[(\boldsymbol{Hx}-\mathbb{E}[\boldsymbol{Hx}\,|\,\boldsymbol{y},\boldsymbol{x}])(\boldsymbol{Hx}-\mathbb{E}[\boldsymbol{Hx}\,|\,\boldsymbol{y},\boldsymbol{x}])^{\dagger}]\end{aligned} \tag{35}$$

where the outer expectations are with respect to $\boldsymbol{x}, \boldsymbol{y}$, and $\boldsymbol{H}$. Gradients with respect to other parameters of the system can be similarly derived as long as they appear after the random transformation $\boldsymbol{Hx}$, i.e., as long as they occur in the last step $\boldsymbol{z} \to \boldsymbol{y}$ of the Markov chain $\boldsymbol{x} \to \boldsymbol{z} \to \boldsymbol{y}$. As an illustrative example, we can again consider the signal model $\boldsymbol{y} = \boldsymbol{Hx}+\boldsymbol{n}$ with $\boldsymbol{\Sigma}_n = N_0\boldsymbol{I}$ and obtain the first-order expansion of the mutual information with respect to N_0 as

$$I(\boldsymbol{x}; \boldsymbol{Hx}+\boldsymbol{n}) = N_0^{-1}\mathbb{E}[\|\mathbb{E}[\boldsymbol{H}](\boldsymbol{x}-\mathbb{E}[\boldsymbol{x}])\|^2] + o\left(N_0^{-1}\right) \tag{36}$$

which agrees with [14, Theorem 2].

D. First-Order Approximation of Mutual Information

In this subsection, we obtain first-order approximations of the mutual information as a function of different parameters using the gradients in Theorems 1 and 2.

Consider the signal model $\boldsymbol{y} = \boldsymbol{Hx}+\boldsymbol{n}$ (with $\boldsymbol{\Sigma}_n = \boldsymbol{I}$). The first-order expansion of the mutual information as a function of the transmit (positive definite) covariance matrix $\boldsymbol{\Sigma}_x$, denoted by $I(\boldsymbol{\Sigma}_x)$, is obtained from the gradient $\nabla_{\boldsymbol{\Sigma}_x} I = \boldsymbol{H}^{\dagger}\boldsymbol{H}\boldsymbol{E}\boldsymbol{\Sigma}_x^{-1}$ (Theorem 2) as [15]

$$\begin{aligned}I(\boldsymbol{\Sigma}_{x,0}+\boldsymbol{\Delta}) &= I(\boldsymbol{\Sigma}_{x,0}) + \mathrm{Tr}\left(\nabla_{\boldsymbol{\Sigma}_x}^{\dagger} I(\boldsymbol{\Sigma}_{x,0})\boldsymbol{\Delta}\right) + o(\|\boldsymbol{\Delta}\|)\\ &= I(\boldsymbol{\Sigma}_{x,0}) + \mathrm{Tr}\left(\boldsymbol{\Sigma}_{x,0}^{-1}\boldsymbol{E}\boldsymbol{H}^{\dagger}\boldsymbol{H}\boldsymbol{\Delta}\right) + o(\|\boldsymbol{\Delta}\|).\end{aligned} \tag{37}$$

This expansion can be easily particularized around $\boldsymbol{\Sigma}_{x,0} = \boldsymbol{0}$ (low-power regime) as follows. The gradient $\nabla_{\boldsymbol{\Sigma}_x} I$ in Theorem 2 is not fully characterized when $\boldsymbol{\Sigma}_x$ is singular, so we need the following more general result (see Appendix D)

$$\nabla_{\boldsymbol{\Sigma}_x} I = \boldsymbol{H}^{\dagger}\boldsymbol{H}\lim_{\epsilon\to 0}\boldsymbol{E}(\boldsymbol{\Sigma}_x+\epsilon\boldsymbol{I})^{-1} \tag{38}$$

where the MMSE matrix $\boldsymbol{E}$ implicitly depends on ϵ, and the limit is equal to $\boldsymbol{E}\boldsymbol{\Sigma}_x^{-1}$, when $\boldsymbol{\Sigma}_x$ is nonsingular, and to the identity matrix $\boldsymbol{I}$, when $\boldsymbol{\Sigma}_x = \boldsymbol{0}$ (in which case $\nabla_{\boldsymbol{\Sigma}_x} I = \boldsymbol{H}^{\dagger}\boldsymbol{H}$). More generally, one can work with the subspaces of $\boldsymbol{\Sigma}_x$ associated to the null eigenvalues and the positive eigenvalues to obtain a closed-form expression for the limit. For example, if $\boldsymbol{H} = \mathrm{diag}(\{\boldsymbol{H}_1, \boldsymbol{H}_2\})$ and $\boldsymbol{\Sigma}_x = \mathrm{diag}(\{\boldsymbol{\Sigma}_{x,1}, \boldsymbol{0}\})$ with nonsingular $\boldsymbol{\Sigma}_{x,1}$, then the limit still admits a simple expression

$$\lim_{\epsilon\to 0}\boldsymbol{E}(\boldsymbol{\Sigma}_x+\epsilon\boldsymbol{I})^{-1} = \begin{bmatrix}\boldsymbol{E}_1\boldsymbol{\Sigma}_{x,1}^{-1} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{I}\end{bmatrix} \tag{39}$$

where $\boldsymbol{E}_1$ is the MMSE matrix corresponding to the components of $\boldsymbol{x}$ associated to $\boldsymbol{\Sigma}_{x,1}$, i.e., the subspace of $\boldsymbol{\Sigma}_x$ with positive eigenvalues.

Using the result in (38), we can immediately obtain the following first-order expansion of the mutual information around $\boldsymbol{\Sigma}_{x,0} = \boldsymbol{0}$:

$$I(\boldsymbol{\Sigma}_x) = \mathrm{Tr}(\boldsymbol{H\Sigma}_x\boldsymbol{H}^{\dagger}) + o(\|\boldsymbol{\Sigma}_x\|). \tag{40}$$

Note that this expression only depends on the signaling through the transmit covariance matrix $\boldsymbol{\Sigma}_x$, which implies the well-known result (e.g., [16]) that Gaussian signaling is not required to achieve $E_b/N_0|_{\min}$.

For illustrative purposes, we now particularize the signal model with $\boldsymbol{\Sigma}_n = N_0\boldsymbol{I}$ and obtain the first-order expansion of

the mutual information as a function of the inverse noise power N_0^{-1} as

$$I\left(N_0^{-1}\right) = N_0^{-1}\mathrm{Tr}(\boldsymbol{H}\boldsymbol{\Sigma}_x\boldsymbol{H}^\dagger) + o\left(N_0^{-1}\right) \tag{41}$$

which agrees with [14, Corollary 2].

IV. Gradients of Information Measures

In this section, we generalize some of the results of [6] by particularizing Theorem 1, and also deal with gradients of differential entropy (showing a new generalized version of De Bruijn's identity) and divergence.

A. Particularizations to SNR Gradients

The fundamental relation between the mutual information and the MMSE was thoroughly explored in [6] for the scalar Gaussian channel $y = \sqrt{\mathsf{snr}}\,x + n$ and, to some extent, also for the vector Gaussian channel $\boldsymbol{y} = \sqrt{\mathsf{snr}}\boldsymbol{x} + \boldsymbol{n}$ considering the trace of the MSE matrix as the estimation performance. We now show how the main results of this paper, namely Theorems 1 and 2, can be readily particularized to extend the results of [6]. The special case of the following result for $\boldsymbol{\Sigma}_n = \boldsymbol{I}$ is given in [6, Theorem 2].

Corollary 1: Consider the signal model in (9), where all the terms are defined as in Theorem 2. Then

$$\frac{dI}{d\mathsf{snr}} = \mathrm{Tr}\left(\boldsymbol{H}^\dagger\boldsymbol{\Sigma}_n^{-1}\boldsymbol{H}\boldsymbol{E}\right) \tag{42}$$

$$= \mathbb{E}\left[\left\|\boldsymbol{\Sigma}_n^{-1/2}\boldsymbol{H}\boldsymbol{x} - \mathbb{E}\left[\boldsymbol{\Sigma}_n^{-1/2}\boldsymbol{H}\boldsymbol{x}|\,\boldsymbol{y}\right]\right\|^2\right]. \tag{43}$$

Proof: This is just a particular case of (8) with $\boldsymbol{B} = \sqrt{\mathsf{snr}}\,\boldsymbol{I}$ (notice that $\boldsymbol{Q}_B = \boldsymbol{B}\boldsymbol{B}^\dagger$ in Theorem 2 plays the role of snr here). The result follows simply from the chain rule (note that $\boldsymbol{Q}_B = \mathsf{snr}\boldsymbol{I}$) and Theorem 2 ($\nabla_{\boldsymbol{Q}_B} I\,\boldsymbol{B} = \boldsymbol{H}^\dagger\boldsymbol{\Sigma}_n^{-1}\boldsymbol{H}\boldsymbol{B}\boldsymbol{E}$)

$$\frac{dI}{d\mathsf{snr}} = \mathrm{Tr}\left(\frac{\partial I}{\partial \boldsymbol{Q}_B^T}\frac{\partial \boldsymbol{Q}_B}{\partial \mathsf{snr}}\right)$$

$$= \mathrm{Tr}\left(\frac{\partial I}{\partial \boldsymbol{Q}_B^*}\right) = \mathrm{Tr}\left(\boldsymbol{H}^\dagger\boldsymbol{\Sigma}_n^{-1}\boldsymbol{H}\boldsymbol{E}\right). \tag{44}$$

□

Observe that (43) is decreasing in snr (since decreasing the noise variance improves the MMSE), which implies that $I(\mathsf{snr})$ is an (increasing) concave function.

Now we can use Corollary 1 to generalize in several ways the result (1) obtained in [6, Theorem 1] for the scalar case. One obvious generalization is to include a scalar channel h in the signal model: $y = \sqrt{\mathsf{snr}}\,hx + n$, obtaining

$$\frac{dI}{d\mathsf{snr}} = |h|^2\mathsf{mmse}(\mathsf{snr}) \tag{45}$$

where $\mathsf{mmse}(\mathsf{snr}) \triangleq \mathbb{E}[|x - \mathbb{E}[x\,|\,y]|^2]$. However, this signal model assumes a memoryless channel and also a memoryless source. Another interesting generalization is to allow a correlated input source (still with a memoryless channel) possibly originated by a code. Consider a block transmission corresponding to N channel uses where the inputs in different blocks are independent

$$y_k = \sqrt{\mathsf{snr}}\,h_k x_k + n_k, \qquad 1 \le k \le N. \tag{46}$$

A direct application of Corollary 1 (after rewriting the signal model in vector form $\boldsymbol{y} = \sqrt{\mathsf{snr}}\,\boldsymbol{H}\boldsymbol{x} + \boldsymbol{n}$ with $\boldsymbol{H} = \mathrm{diag}(\{h_k\})$) gives (see also [6, Corollary 3])

$$\frac{dI}{d\,\mathsf{snr}} = \sum_{k=1}^{N}|h_k|^2\,\mathsf{mmse}_k(\mathsf{snr}) \tag{47}$$

where $\mathsf{mmse}_k(\mathsf{snr}) \triangleq \mathbb{E}[|x_k - \mathbb{E}[x_k\,|\,\boldsymbol{y}]|^2]$.

Next, we state a result for a multiuser scenario, which generalizes [6, Theorem 4].

Corollary 2: Consider the following signal model:

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{\Gamma}\boldsymbol{x} + \boldsymbol{n} \tag{48}$$

where $\boldsymbol{\Gamma}$ is a square invertible matrix and the rest of the terms are defined as in Theorem 2. Then

$$\nabla_{\boldsymbol{\Gamma}\boldsymbol{\Gamma}^\dagger} I = \boldsymbol{H}^\dagger\boldsymbol{\Sigma}_n^{-1}\boldsymbol{H}\boldsymbol{\Gamma}\boldsymbol{E}\boldsymbol{\Gamma}^{-1}. \tag{49}$$

Proof: This result follows from Theorem 2 with $\boldsymbol{B} = \boldsymbol{\Gamma}$.□

Observe that, when (48) models a multiuser system with the symbols transmitted by all K users stacked in $\boldsymbol{x}$ and $\boldsymbol{\Gamma} = \mathrm{diag}(\{\sqrt{\mathsf{snr}_k}\}_{k=1}^K)$, then (49) is particularly useful since $\partial I/\partial\,\mathsf{snr}_k = [\nabla_{\boldsymbol{\Gamma}\boldsymbol{\Gamma}^\dagger} I]_{kk}$.

B. Matrix Generalization of De Bruijn's Identity

For a multivariate density function $p_{\boldsymbol{y}}(\boldsymbol{y})$, De Bruijn's identity [2], [17]–[19] relates the derivative of the differential entropy $h(\cdot)$ to the Fisher information matrix defined as

$$\boldsymbol{J}(\boldsymbol{y}) \triangleq \mathbb{E}_{\boldsymbol{y}}[\nabla_{\boldsymbol{y}}\log p_{\boldsymbol{y}}(\boldsymbol{y})\nabla_{\boldsymbol{y}}^\dagger\log p_{\boldsymbol{y}}(\boldsymbol{y})]. \tag{50}$$

Note that this is a special form of the Fisher information matrix (with respect to a translation parameter) which does not involve an explicit parameter as in its most general definition [19] .[4] The scalar version is defined as

$$J(\boldsymbol{y}) \triangleq \mathrm{Tr}(\boldsymbol{J}(\boldsymbol{y})). \tag{51}$$

The original De Bruijn's identity was obtained for the scalar case [2], [17] as

$$\frac{d}{dt}h(z + \sqrt{t}n) = J(z + \sqrt{t}n) \tag{52}$$

where the noise n is standard Gaussian. The scalar version of the multivariate De Bruijn's identity (assuming a normalized covariance matrix for the noise $\boldsymbol{\Sigma}_n = \boldsymbol{I}$) is [18], [19, Theorem 14], [6, eq. (51)]

$$\frac{d}{dt}h(\boldsymbol{z} + \sqrt{t}\boldsymbol{n}) = J(\boldsymbol{z} + \sqrt{t}\boldsymbol{n}). \tag{53}$$

[4]The Fisher information matrix with respect to a parameter $\boldsymbol{\theta}$ is defined as $\boldsymbol{J}(\boldsymbol{y}) \triangleq \mathbb{E}_{\boldsymbol{y}}[\nabla_{\boldsymbol{\theta}}\log p_{\boldsymbol{y}}(\boldsymbol{y};\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}}^\dagger\log p_{\boldsymbol{y}}(\boldsymbol{y};\boldsymbol{\theta})]$ [7].

A more general multivariate De Bruijn's identity is[5]

$$\frac{d}{dt}h(\boldsymbol{z}+\sqrt{t}\boldsymbol{A}\boldsymbol{n}) = \mathrm{Tr}(\boldsymbol{A}^\dagger \boldsymbol{J}(\boldsymbol{z}+\sqrt{t}\boldsymbol{A}\boldsymbol{n})\boldsymbol{A}) \tag{54}$$

where again the noise has a normalized noise covariance matrix and $\boldsymbol{A}$ is an arbitrary square invertible matrix. Note that in the context of the signal model in (5) we can set $\boldsymbol{z} = \boldsymbol{H}\boldsymbol{x}$ to particularize the results.

As can be observed from (52)–(54), the existing versions of De Bruijn's identity take a derivative with respect to a scalar parameter which takes the role of a noise-to-signal ratio. We now generalize these results by considering the gradient with respect to the noise covariance matrix and, more explicitly, with respect to an arbitrary linear transformation of the noise $\boldsymbol{T}^{1/2}\boldsymbol{n}$, where the linear transformation $\boldsymbol{T}^{1/2}$ plays the role of $\sqrt{t}$ in the scalar version of the identity.

Theorem 4 (Matrix Version of the Multivariate De Bruijn's Identity): Consider an arbitrary random variable (with finite second-order moments) $\boldsymbol{z}$ contaminated with a Gaussian noise $\boldsymbol{n}$ independent of $\boldsymbol{z}$ and with positive definite covariance matrix $\boldsymbol{\Sigma}_n$. Then

$$\nabla_{\boldsymbol{\Sigma}_n} h(\boldsymbol{z}+\boldsymbol{n}) = \boldsymbol{J}(\boldsymbol{z}+\boldsymbol{n}). \tag{55}$$

If $\boldsymbol{z}$ is contaminated with the Gaussian random vector $\boldsymbol{T}^{1/2}\boldsymbol{n}$, where $\boldsymbol{T}^{1/2}$ is an arbitrary square invertible matrix, then

$$\nabla_{\boldsymbol{T}} h(\boldsymbol{z}+\boldsymbol{T}^{1/2}\boldsymbol{n}) = \boldsymbol{J}(\boldsymbol{z}+\boldsymbol{T}^{1/2}\boldsymbol{n}) \times \boldsymbol{T}^{1/2}\boldsymbol{\Sigma}_n\boldsymbol{T}^{-1/2}. \tag{56}$$

Proof: See Appendix E. □

Observe that when the noise has a normalized covariance matrix $\boldsymbol{\Sigma}_n = \boldsymbol{I}$, (56) simplifies to

$$\nabla_{\boldsymbol{T}} h(\boldsymbol{z}+\boldsymbol{T}^{1/2}\boldsymbol{n}) = \boldsymbol{J}(\boldsymbol{z}+\boldsymbol{T}^{1/2}\boldsymbol{n}) \tag{57}$$

which is the natural generalization of the scalar identities (52)–(53).

The scalar expression of the multivariate De Bruijn's identity in (53) can be obtained from a simple application of the chain rule to Theorem 4 particularized for $\boldsymbol{T} = t\boldsymbol{I}$ as follows:

$$\frac{dh}{dt} = \mathrm{Tr}\left(\frac{\partial h}{\partial \boldsymbol{T}^T}\frac{\partial \boldsymbol{T}}{\partial t}\right) = \mathrm{Tr}\left(\frac{\partial h}{\partial \boldsymbol{T}^*}\right) = \mathrm{Tr}(\boldsymbol{J}(\boldsymbol{z}+\sqrt{t}\boldsymbol{n})). \tag{58}$$

The more general expression for the the multivariate De Bruijn's identity in (54) also follows directly from Theorem 4 by particularizing to $\boldsymbol{T} = t\boldsymbol{A}\boldsymbol{A}^\dagger (\boldsymbol{T}^{1/2} = \sqrt{t}\boldsymbol{A})$ and using the chain rule

$$\frac{dh}{dt} = \mathrm{Tr}\left(\frac{\partial h}{\partial \boldsymbol{T}^T}\frac{\partial \boldsymbol{T}}{\partial t}\right) = \mathrm{Tr}\left(\frac{\partial h}{\partial \boldsymbol{T}^*}\boldsymbol{A}\boldsymbol{A}^\dagger\right). \tag{59}$$

The multidimensional Cramer–Rao bound states that the Fisher information matrix is lower-bounded (i.e., the difference is nonnegative definite) by the inverse of its covariance matrix (e.g., [19]) (setting now $\boldsymbol{z} = \boldsymbol{H}\boldsymbol{x}$)

$$\begin{aligned}\boldsymbol{J}(\boldsymbol{y}) &\geq (\mathbb{E}[(\boldsymbol{y}-\mathbb{E}[\boldsymbol{y}])(\boldsymbol{y}-\mathbb{E}[\boldsymbol{y}])^\dagger])^{-1} \\ &= (\boldsymbol{H}\boldsymbol{\Sigma}_x\boldsymbol{H}^\dagger + \boldsymbol{\Sigma}_n)^{-1}\end{aligned} \tag{60}$$

with equality if and only if $\boldsymbol{y}$ is a Gaussian vector. Therefore, for Gaussian signaling, the following relation holds between the Fisher information matrix $\boldsymbol{J}(\boldsymbol{y})$ and the MMSE matrix $\boldsymbol{E}$:

$$\boldsymbol{\Sigma}_x\boldsymbol{H}^\dagger \boldsymbol{J}(\boldsymbol{H}\boldsymbol{x}+\boldsymbol{n})\boldsymbol{H}\boldsymbol{\Sigma}_x = \boldsymbol{\Sigma}_x - \boldsymbol{E} \tag{61}$$

where we have used (15) and the matrix inversion lemma.[6]

C. Divergence Gradients

1) Divergence for a Given Input: The mutual information $I(\boldsymbol{x};\boldsymbol{y})$ is equal to the conditional divergence $D(p_{\boldsymbol{y}\,|\,\boldsymbol{x}} \,\|\, p_{\boldsymbol{y}}\,|\,p_{\boldsymbol{x}})$. Since this quantity is an average of the divergence between conditional and unconditional output distributions, the following result is a refinement of Theorem 1.

Theorem 5: Consider the signal model in (5), where all the terms are defined as in Theorem 1. Then, the gradient of the divergence between the output conditioned on a given fixed input $\boldsymbol{x}_0$ and the unconditional output is

$$\begin{aligned}&\nabla_{\boldsymbol{H}} D(p_{\boldsymbol{y}\,|\,\boldsymbol{x}=\boldsymbol{x}_0} \,\|\, p_{\boldsymbol{y}}) \\ &\quad = \boldsymbol{H}\mathbb{E}[(\boldsymbol{x}_0 - \mathbb{E}[\boldsymbol{x}\,|\,\boldsymbol{y}])(\boldsymbol{x}_0 - \mathbb{E}[\boldsymbol{x}\,|\,\boldsymbol{y}])^\dagger \,|\, \boldsymbol{x}=\boldsymbol{x}_0]\end{aligned} \tag{62}$$

where the expectation is with respect to the conditional output distribution $p_{\boldsymbol{y}\,|\,\boldsymbol{x}=\boldsymbol{x}_0}$.

Proof: See Appendix F. □

Taking the expectation over $\boldsymbol{x}_0$ in (62) with respect to $p_{\boldsymbol{x}}$, we get

$$\mathbb{E}[\nabla_{\boldsymbol{H}} D(p_{\boldsymbol{y}\,|\,\boldsymbol{x}=\boldsymbol{x}_0} \,\|\, p_{\boldsymbol{y}})] = \boldsymbol{H}\boldsymbol{E}. \tag{63}$$

Interchanging gradient and expectation in (63), we obtain Theorem 1.[7] We consider now the gradient of the conditional divergence but with the arguments swapped, i.e., of $D(p_{\boldsymbol{y}} \,\|\, p_{\boldsymbol{y}\,|\,\boldsymbol{x}})$.

Theorem 6: Consider the signal model in (5), where all the terms are defined as in Theorem 1. Then, the gradient of the divergence between the unconditional output and the output conditioned on a given fixed input $\boldsymbol{x}_0$ is

$$\nabla_{\boldsymbol{H}} D(p_{\boldsymbol{y}} \,\|\, p_{\boldsymbol{y}\,|\,\boldsymbol{x}=\boldsymbol{x}_0}) = \boldsymbol{H}(\mathbb{E}[(\boldsymbol{x}-\boldsymbol{x}_0)(\boldsymbol{x}-\boldsymbol{x}_0)^\dagger] - \boldsymbol{E}) \tag{64}$$

$$= \boldsymbol{H}\mathbb{E}[(\boldsymbol{x}_0 - \mathbb{E}[\boldsymbol{x}\,|\,\boldsymbol{y}])(\boldsymbol{x}_0 - \mathbb{E}[\boldsymbol{x}\,|\,\boldsymbol{y}])^\dagger] \tag{65}$$

where the expectation in (65) is with respect to the unconditional output distribution $p_{\boldsymbol{y}}$.

Proof: See Appendix G. □

Observe the similarity between (62), where the expectation over $\boldsymbol{y}$ is with respect to the conditional $p_{\boldsymbol{y}\,|\,\boldsymbol{x}=\boldsymbol{x}_0}$, and (65), where the expectation over $\boldsymbol{y}$ is with respect to the unconditional $p_{\boldsymbol{y}}$.

Taking the expectation over $\boldsymbol{x}_0$ in (64) with respect to $p_{\boldsymbol{x}}$, we get

$$\mathbb{E}[\nabla_{\boldsymbol{H}} D(p_{\boldsymbol{y}} \,\|\, p_{\boldsymbol{y}\,|\,\boldsymbol{x}=\boldsymbol{x}_0})] = \boldsymbol{H}(2\boldsymbol{\Sigma}_x - \boldsymbol{E}). \tag{66}$$

[5]The special case of (54) for $t = 0$ is given in [20]. An alternative expression for the multivariate De Bruijn's identity was obtained in [21].

[6]Matrix inversion lemma:

$$(\boldsymbol{A}+\boldsymbol{B}\boldsymbol{C}\boldsymbol{D})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{D}\boldsymbol{A}^{-1}\boldsymbol{B} + \boldsymbol{C}^{-1})^{-1}\boldsymbol{D}\boldsymbol{A}^{-1}.$$

[7]The validity of the interchange between gradient and expectation follows from the boundeness of (62) and the Lebesgue Dominated Convergence Theorem [22].

2) Divergence of Signal-Plus-Noise Versus Noise: The following result relates two fundamental quantities: the likelihood ratio (between the hypotheses that the random signal $\boldsymbol{x}$ has been transmitted and nothing has been transmitted) and the MMSE estimation, which play central roles in detection and estimation theory, respectively, [7], [23] (see [6, Theorem 5] for the scalar version).

Theorem 7: Consider the signal model in (5), where all the terms are defined as in Theorem 1. Then, the gradient of the divergence between the signal-plus-noise and noise distributions is

$$\nabla_{\boldsymbol{H}} D(p_{\boldsymbol{y}} \,\|\, p_{\boldsymbol{n}}) = \boldsymbol{H}(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\dagger] - \boldsymbol{E}) \tag{67}$$

$$= \boldsymbol{H}\mathbb{E}[\mathbb{E}[\boldsymbol{x}\,|\,\boldsymbol{y}]\mathbb{E}[\boldsymbol{x}^\dagger\,|\,\boldsymbol{y}]]. \tag{68}$$

Proof: Set $\boldsymbol{x}_0 = 0$ in Theorem 6. □

3) Non-Gaussianness: The non-Gaussianness of a random variable is defined in [24] as the divergence between its distribution and a Gaussian distribution with the same first- and second-order moments. The gradient of the non-Gaussianness can be directly obtained from Theorem 1 (similarly, Theorem 2) by noting that the mutual information admits the following decomposition (see, for example, [24]):

$$I(\boldsymbol{x};\boldsymbol{y}) = I(\boldsymbol{x}_G;\boldsymbol{y}_G) - D\left(p_{\boldsymbol{y}} \,\|\, p_{\boldsymbol{y}_G}\right) \tag{69}$$

where $(\boldsymbol{x}_G, \boldsymbol{y}_G)$ denote jointly Gaussian-distributed random vectors with the same mean and covariance matrix as the joint distribution of $(\boldsymbol{x}, \boldsymbol{y})$ and $D(p_{\boldsymbol{y}} \,\|\, p_{\boldsymbol{y}_G})$ is the non-Gaussianness of $\boldsymbol{y}$.

Theorem 8: Consider the signal model in (5), where all the terms are defined as in Theorem 2. Then, the gradient of the non-Gaussianness of $\boldsymbol{y}$ is

$$\nabla_{\boldsymbol{H}} D\left(p_{\boldsymbol{y}} \,\|\, p_{\boldsymbol{y}_G}\right) = \boldsymbol{\Sigma}_n^{-1}\boldsymbol{H}(\boldsymbol{E}_{\text{lin}} - \boldsymbol{E}) \tag{70}$$

where $\boldsymbol{E}_{\text{lin}} = (\boldsymbol{\Sigma}_x^{-1} + \boldsymbol{H}^\dagger\boldsymbol{\Sigma}_n^{-1}\boldsymbol{H})^{-1}$ is the linear MMSE matrix, i.e., the MMSE matrix among the class of *linear* estimators.

Using (70) and the fact that

$$D(p_{\boldsymbol{H}\boldsymbol{x}} \,\|\, p_{\boldsymbol{H}\boldsymbol{x}_G}) = \int_0^\infty (dD(p_{\boldsymbol{y}} \,\|\, p_{\boldsymbol{y}_G})/d\mathsf{snr})\, d\mathsf{snr}$$

for the signal model in (9) (since $D(p_{\boldsymbol{H}\boldsymbol{x}} \,\|\, p_{\boldsymbol{H}\boldsymbol{x}_G}) = \lim_{\mathsf{snr}\to\infty} D(p_{\boldsymbol{y}} \,\|\, p_{\boldsymbol{y}_G})$ [6, Lemma 7]), we can write the following alternative expression for the differential entropy of a random vector (setting $\boldsymbol{H} = \boldsymbol{I}$ and $\boldsymbol{\Sigma}_n = \boldsymbol{I}$) which generalizes the corresponding scalar version in [6, Theorem 14]:

$$h(\boldsymbol{x}) = h(\boldsymbol{x}_G) - D(p_{\boldsymbol{x}} \,\|\, p_{\boldsymbol{x}_G})$$

$$= \log\det(\pi e\boldsymbol{\Sigma}_x) - \int_0^\infty \mathrm{Tr}(\boldsymbol{E}_{\text{lin}}(\mathsf{snr}) - \boldsymbol{E}(\mathsf{snr}))\, d\mathsf{snr} \tag{71}$$

where in this case $\boldsymbol{E}_{\text{lin}}(\mathsf{snr}) = (\boldsymbol{\Sigma}_x^{-1} + \mathsf{snr}\,\boldsymbol{I})^{-1}$. Note that the integral is always a nonnegative quantity since $\boldsymbol{E}_{\text{lin}}(\mathsf{snr}) \geq \boldsymbol{E}(\mathsf{snr})$, which agrees with the fact that the Gaussian distribution has the highest differential entropy for given second-order moments. Recall that $\mathrm{Tr}(\boldsymbol{E}(\mathsf{snr})) = \mathbb{E}[\|\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}\,|\,\boldsymbol{y}]\|^2]$ is the mean-squared norm of the estimation error vector.

V. Practical Application: Precoder Optimization

The availability of the mutual information gradient enables us to optimize with respect to parameters of interest. One can always compute the gradient numerically and use it in a gradient-based algorithm to search the maximizing solution (which may be only locally optimum if the problem is not convex). In some special cases, it is even possible to derive optimality conditions in closed form. Recall that for a Gaussian input, the maximizing solution is the well-known channel diagonalization plus water-filling power allocation (e.g., [13]). In the following, we focus on a numerical optimization (for an arbitrary input) based on a simple gradient projection method.

Consider the optimization of the linear precoder $\boldsymbol{B}$ to maximize the mutual information of the general signal model $\boldsymbol{y} = \boldsymbol{H}\boldsymbol{B}\boldsymbol{x} + \boldsymbol{n}$ subject to the power constraint $\mathrm{Tr}(\boldsymbol{B}\boldsymbol{B}^\dagger) \leq P_T$. To deal with this constrained optimization problem, we can use a simple gradient update with an additional projection to guarantee the feasibility of the new solution [25]–[27]

$$\boldsymbol{B}(k+1) = [\boldsymbol{B}(k) + \mu\nabla_{\boldsymbol{B}} I(k)]^+_{\mathrm{Tr}(\boldsymbol{B}\boldsymbol{B}^\dagger)\leq P_T} \tag{72}$$

where k denotes the iteration index, μ is the stepsize of the update, the gradient of the mutual information at the kth iteration is given by $\nabla_{\boldsymbol{B}} I(k) = \boldsymbol{H}^\dagger\boldsymbol{\Sigma}_n^{-1}\boldsymbol{H}\boldsymbol{B}(k)\boldsymbol{E}(k)$ (from Theorem 2), and $[\,\cdot\,]^+_{\mathrm{Tr}(\boldsymbol{B}\boldsymbol{B}^\dagger)\leq P_T}$ denotes the projection onto the feasible set $\mathrm{Tr}(\boldsymbol{B}\boldsymbol{B}^\dagger) \leq P_T$. For simplicity, we choose the stepsize μ as a fixed small number; however, other more sophisticated choices can be made with better convergence properties [26].

For the numerical computation of the MMSE matrix, which is required to evaluate the gradient $\nabla_{\boldsymbol{B}} I(k)$ in (72), closed-form expressions may be used in case they are available; otherwise, one can resort to a Monte Carlo method to approximate the MMSE matrix as follows:

$$\boldsymbol{E} = \mathbb{E}[(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}\,|\,\boldsymbol{y}])(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}\,|\,\boldsymbol{y}])^\dagger]$$

$$\approx \frac{1}{S}\sum_{\boldsymbol{x}_0,\boldsymbol{y}_0} (\boldsymbol{x}_0 - \mathbb{E}[\boldsymbol{x}\,|\,\boldsymbol{y} = \boldsymbol{y}_0])(\boldsymbol{x}_0 - \mathbb{E}[\boldsymbol{x}\,|\,\boldsymbol{y} = \boldsymbol{y}_0])^\dagger \tag{73}$$

where the S random samples $\boldsymbol{x}_0, \boldsymbol{y}_0$ are generated from $p_{\boldsymbol{x}}p_{\boldsymbol{y}\,|\,\boldsymbol{x}}$ and the conditional estimate is computed as

$$\mathbb{E}[\boldsymbol{x}\,|\,\boldsymbol{y} = \boldsymbol{y}_0] = \sum_{\boldsymbol{x}_0} \boldsymbol{x}_0 p_{\boldsymbol{x}\,|\,\boldsymbol{y}}(\boldsymbol{x}_0\,|\,\boldsymbol{y}_0) \tag{74}$$

where the summation is over all possible input vectors.

The projection of a point $\boldsymbol{B}_0$ onto the feasible set $\mathrm{Tr}(\boldsymbol{B}\boldsymbol{B}^\dagger) \leq P_T$ (as required in (72)) is in turn the solution to the following convex optimization problem:

$$\begin{aligned} \underset{\boldsymbol{B}}{\text{minimize}} \quad & \|\boldsymbol{B} - \boldsymbol{B}_0\|^2 \\ \text{subject to} \quad & \|\boldsymbol{B}\|^2 \leq P_T. \end{aligned} \tag{75}$$

By forming the Lagrangian and deriving the Karush–Kuhn–Tucker (KKT) optimality conditions [28], one can easily find that the optimal solution to (75) is proportional to $\boldsymbol{B}_0$ up to a scaling factor chosen to satisfy the power constraint.

Summarizing, we propose the following iterative numerical algorithm to optimize the linear precoder $\boldsymbol{B}$.

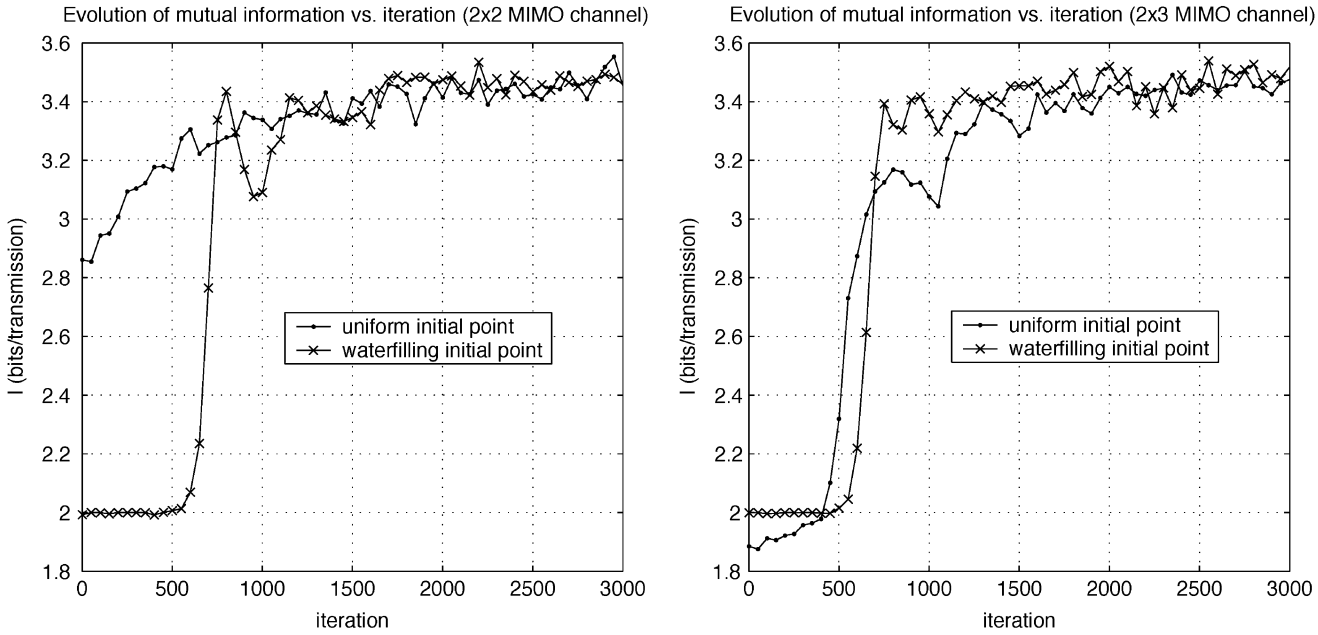

Fig. 2. Evolution of the mutual information (for a 2×2 channel matrix and for a 2×3 channel matrix) as the linear precoder is iteratively optimized with the proposed gradient algorithm.

Algorithm:

0. Initialization: set $k = 0$ and initialize the precoder $\boldsymbol{B}(0)$ to some sensible choice such as
   - uniform power allocation (equivalent to no precoder): $\boldsymbol{B}(0) = P_T/L[\boldsymbol{I}\ \boldsymbol{0}]^T$;
   - channel-diagonalization plus water filling (optimum for Gaussian inputs):

     $$\boldsymbol{B}(0) = \boldsymbol{U}\mathrm{diag}(\{\sqrt{p_i}\}),$$

     where $\boldsymbol{U}$ contains as columns the right singular vectors of the channel matrix corresponding to the L largest squared singular values λ_i and $p_i = (\mu - \lambda_i^{-1})^+$ is the water-filling power allocation.

1 Update the precoder with the gradient (after updating $\boldsymbol{E}(k)$)

$$\boldsymbol{B}(k+1) \leftarrow \boldsymbol{B}(k) + \mu \boldsymbol{H}^\dagger \boldsymbol{\Sigma}_n^{-1} \boldsymbol{H}\boldsymbol{B}(k)\boldsymbol{E}(k). \quad (76)$$

2 If $\|\boldsymbol{B}(k+1)\|^2 > P_T$, then normalize

$$\boldsymbol{B}(k+1) \leftarrow \boldsymbol{B}(k+1)\sqrt{P_T/\|\boldsymbol{B}(k+1)\|^2}. \quad (77)$$

3 Set $k \leftarrow k+1$ and go to Step 1 (until some termination criterion is activated).

Observe that, in principle, at each iteration of the algorithm one has to estimate the MMSE matrix for the given linear precoder at the current iteration. This can be too time-consuming and a convenient solution is to update the MMSE matrix recursively. The rationale underneath this practical trick is that if the stepsize μ is sufficiently small, then the linear precoder will change very slowly, which implies that we can take advantage of the previously computed MMSE matrix. In particular, we have updated the MMSE matrix as follows:

$$\boldsymbol{E}(k+1) = \alpha \boldsymbol{E}(k) + (1-\alpha)\boldsymbol{E}^{\mathrm{new}}(k+1) \quad (78)$$

where α is the forgetting factor which measures how fast old data are forgotten (with a typical value around 0.9–0.99) and $\boldsymbol{E}^{\mathrm{new}}(k+1)$ is the MMSE matrix estimated with new samples (typically just a few samples) based on the new precoder $\boldsymbol{B}(k+1)$.

Fig. 2 illustrates the convergence of the algorithm for the 2×2 channel matrix

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

over which two independent and equiprobable quaternary phase-shift keying (QPSK) symbols (drawn from $\{\pm 1, \pm j\}$) are transmitted. The precoder is first initialized to a uniform power allocation $\boldsymbol{B}(0) = \boldsymbol{I}$ and then to the channel-diagonalizing plus water-filling solution, which happens to be a bad initialization in this case because it only allocates power to one channel eigenmode which translates into a maximum of 2 bits/transmission. Each new MMSE matrix $\boldsymbol{E}^{\mathrm{new}}$ is computed with just $S = 10$ random samples and the averaged matrix $\bar{\boldsymbol{E}}$ is obtained with a forgetting factor of $\alpha = 0.99$.

Fig. 2 also illustrates the convergence for a 2×3 channel matrix obtained by appending a zero column on the left of the previous channel matrix. The 3×2 precoder matrix is again initialized to both a uniform power allocation $\boldsymbol{B}(0) = [\boldsymbol{I}\ \boldsymbol{0}]^T$ and to the channel-diagonalizing plus water-filling solution. As expected, the precoder matrix evolves to have a zero top row (to avoid the zero column of the channel matrix) and the two bottom rows are as in the previous 2×2 case, obtaining therefore the same final value of mutual information.

VI. Conclusion

Building upon the setting in [6], this paper has obtained the gradient of the mutual information of a general linear vector

Gaussian channel with arbitrary signaling with respect to several key system parameters. The basic expression of the gradient with respect to the channel matrix happens to involve the MMSE matrix, connecting the two key quantities from information theory and estimation theory. Gradients and derivatives with respect to arbitrary parameters are then easily obtained with the application of the chain rule for differentiation. Several interesting formulas have been obtained by applying the main result: first-order approximations of the mutual information, matrix generalization of De Bruijn's identity, gradient of the divergence between the conditional and unconditional outputs, gradient of the divergence between signal-plus-noise and noise, and gradient of the non-Gaussianness. We have illustrated the practical application of our results with an optimization problem that arises in the design of a linear precoder that maximizes mutual information.

APPENDIX

A. A Note on Complex Derivatives and Gradients

In general, commonly encountered functions of complex variables are not analytic and a more generalized definition of complex derivative needs to be used [29]. In particular, we use the well-known definition of the complex derivative of a real-valued scalar function f

$$\frac{df}{dx^*} \triangleq \frac{1}{2}\left(\frac{\partial f}{\partial \mathrm{Re}\{x\}} + j\frac{\partial f}{\partial \mathrm{Im}\{x\}}\right) \tag{79}$$

as in [29], so that, for example

$$\frac{\partial}{\partial \boldsymbol{x}^*}(\boldsymbol{x}^\dagger \boldsymbol{R}\boldsymbol{x}) = \boldsymbol{R}\boldsymbol{x} \tag{80}$$

(for $\boldsymbol{R}$ independent of $\boldsymbol{x}$). The derivative of a complex-valued scalar function is defined componentwise on the real and imaginary part, i.e., $\partial f/\partial \boldsymbol{x}^* = \partial \mathrm{Re}\{f\}/\partial \boldsymbol{x}^* + j\partial \mathrm{Im}\{f\}/\partial \boldsymbol{x}^*$.

For the sake of notation, we define the complex gradient vector as $\nabla_{\boldsymbol{x}} f \triangleq \partial f/\partial \boldsymbol{x}^*$ as in [27].[8] The complex gradient matrix is similarly defined as $\nabla_{\boldsymbol{X}} f \triangleq \partial f/\partial \boldsymbol{X}^*$, where $[\nabla_{\boldsymbol{X}} f]_{ij} = \partial f/\partial [\boldsymbol{X}^*]_{ij}$, so that, for example, [15]

$$\nabla_{\boldsymbol{X}} \mathrm{Tr}(\boldsymbol{X}^\dagger \bar{\boldsymbol{R}}) = \bar{\boldsymbol{R}} \tag{81}$$

and

$$\nabla_{\boldsymbol{X}} \log\det(\boldsymbol{I} + \boldsymbol{X}^\dagger \boldsymbol{R}\boldsymbol{X}) = \boldsymbol{R}\boldsymbol{X}(\boldsymbol{I} + \boldsymbol{X}^\dagger \boldsymbol{R}\boldsymbol{X})^{-1}. \tag{82}$$

The chain rule for a nonanalytic real-valued scalar function f of a real vector $\boldsymbol{r}$ and a complex vector $\boldsymbol{c}$, denoted explicitly by $f(\boldsymbol{r}, \boldsymbol{c}, \boldsymbol{c}^*)$, is

$$\frac{\partial f(\boldsymbol{r},\boldsymbol{c},\boldsymbol{c}^*)}{\partial x} = \left(\frac{\partial f(\boldsymbol{r},\boldsymbol{c},\boldsymbol{c}^*)}{\partial \boldsymbol{r}}\right)^T \frac{\partial \boldsymbol{r}}{\partial x} + \left(\frac{\partial f(\boldsymbol{r},\boldsymbol{c},\boldsymbol{c}^*)}{\partial \boldsymbol{c}}\right)^T \frac{\partial \boldsymbol{c}}{\partial x} + \left(\frac{\partial f(\boldsymbol{r},\boldsymbol{c},\boldsymbol{c}^*)}{\partial \boldsymbol{c}^*}\right)^T \frac{\partial \boldsymbol{c}^*}{\partial x}$$

where the last term vanishes for analytic functions, obtaining then the classical chain rule.

[8]Note, however, that in [27] there is an additional factor of 2.

B. Proof of Theorem 1

Observe that the means of the transmitted signal and noise are completely arbitrary in the statement of the theorem. In fact, the result is not affected by the means since both the mutual information and the MMSE matrix are insensitive to the means. For simplicity, we will assume the noise to be zero mean without loss of generality. The justification of the interchange of the order of integrals and derivatives is left for the end of the proof. In the scalar case, this proof reduces to the proof in [6, Appendix III].

Since the noise is Gaussian with zero mean and normalized covariance matrix, the conditional output pdf is

$$p_{\boldsymbol{y}|\boldsymbol{x}}(\boldsymbol{y}) = \frac{1}{\pi^{n_R}} \exp(-\|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}\|^2) \tag{83}$$

and the unconditional output pdf is $p_{\boldsymbol{y}}(\boldsymbol{y}) = \mathbb{E}_{\boldsymbol{x}}[p_{\boldsymbol{y}|\boldsymbol{x}}(\boldsymbol{y})]$. The mutual information can be written as

$$\begin{aligned} I(\boldsymbol{x};\boldsymbol{y}) &= \mathbb{E}\left[\log \frac{p_{\boldsymbol{y}|\boldsymbol{x}}(\boldsymbol{y})}{p_{\boldsymbol{y}}(\boldsymbol{y})}\right] \\ &= -n_R \log(\pi e) - \mathbb{E}[\log p_{\boldsymbol{y}}(\boldsymbol{y})] \\ &= -n_R \log(\pi e) - \int p_{\boldsymbol{y}}(\boldsymbol{y}) \log p_{\boldsymbol{y}}(\boldsymbol{y}) d\boldsymbol{y}. \end{aligned} \tag{84}$$

Then

$$\frac{\partial I}{\partial \boldsymbol{H}^*} = -\int (1 + \log p_{\boldsymbol{y}}(\boldsymbol{y})) \frac{\partial p_{\boldsymbol{y}}(\boldsymbol{y})}{\partial \boldsymbol{H}^*} d\boldsymbol{y} \tag{85}$$

$$= -\int (1 + \log p_{\boldsymbol{y}}(\boldsymbol{y})) \mathbb{E}_{\boldsymbol{x}}\left[\frac{\partial p_{\boldsymbol{y}|\boldsymbol{x}}(\boldsymbol{y})}{\partial \boldsymbol{H}^*}\right] d\boldsymbol{y}. \tag{86}$$

The derivative of the conditional output can be written as

$$\begin{aligned} \frac{\partial p_{\boldsymbol{y}|\boldsymbol{x}}(\boldsymbol{y})}{\partial \boldsymbol{H}^*} &= -p_{\boldsymbol{y}|\boldsymbol{x}}(\boldsymbol{y}) \frac{\partial}{\partial \boldsymbol{H}^*}((\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x})^\dagger (\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x})) \\ &= p_{\boldsymbol{y}|\boldsymbol{x}}(\boldsymbol{y})(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x})\boldsymbol{x}^\dagger \\ &= -\nabla_{\boldsymbol{y}} p_{\boldsymbol{y}|\boldsymbol{x}}(\boldsymbol{y}) \boldsymbol{x}^\dagger. \end{aligned} \tag{87}$$

Therefore, we have

$$\begin{aligned} \frac{\partial I}{\partial \boldsymbol{H}^*} &= \int (1 + \log p_{\boldsymbol{y}}(\boldsymbol{y})) \mathbb{E}_{\boldsymbol{x}}[\nabla_{\boldsymbol{y}} p_{\boldsymbol{y}|\boldsymbol{x}}(\boldsymbol{y}) \boldsymbol{x}^\dagger]\, d\boldsymbol{y} \\ &= \mathbb{E}\left[\left(\int (1 + \log p_{\boldsymbol{y}}(\boldsymbol{y})) \nabla_{\boldsymbol{y}} p_{\boldsymbol{y}|\boldsymbol{x}}(\boldsymbol{y}) d\boldsymbol{y}\right) \boldsymbol{x}^\dagger\right] \\ &= \mathbb{E}\left[\left(-\int \frac{p_{\boldsymbol{y}|\boldsymbol{x}}(\boldsymbol{y})}{p_{\boldsymbol{y}}(\boldsymbol{y})} \nabla_{\boldsymbol{y}} p_{\boldsymbol{y}}(\boldsymbol{y}) d\boldsymbol{y}\right) \boldsymbol{x}^\dagger\right] \end{aligned} \tag{88}$$

where the last equality follows from integrating by parts [22] [9] and noting that $p_{\boldsymbol{y}|\boldsymbol{x}}(\boldsymbol{y})(1 + \log p_{\boldsymbol{y}}(\boldsymbol{y})) \to 0$ as $\|\boldsymbol{y}\| \to \infty$.

[9]The following real integration by parts should be applied to the real and imaginary parts of each element of $\boldsymbol{y}$ (noting that integration on $\mathbb{C}^n$ can be interpreted as integration on $\mathbb{R}^{2n}$ [7])

$$\int_{-\infty}^{+\infty} (1 + \log_e p_{\boldsymbol{y}})(\partial p_{\boldsymbol{y}|\boldsymbol{x}}/\partial t)\, dt = [(1 + \log_e p_{\boldsymbol{y}}) p_{\boldsymbol{y}|\boldsymbol{x}}]_{t=-\infty}^{+\infty} - \int_{-\infty}^{+\infty} (1/p_{\boldsymbol{y}})(\partial p_{\boldsymbol{y}}/\partial t) p_{\boldsymbol{y}|\boldsymbol{x}}\, dt$$

where t represents a real variable [22].

Noting that, for almost every $\boldsymbol{y}, p_{\boldsymbol{x}\,|\,\boldsymbol{y}} \ll p_{\boldsymbol{x}}$ and the Radon–Nikodym derivative is $\frac{dp_{\boldsymbol{x}\,|\,\boldsymbol{y}}}{dp_{\boldsymbol{x}}}(\boldsymbol{x},\boldsymbol{y}) = p_{\boldsymbol{y}\,|\,\boldsymbol{x}}(\boldsymbol{y}\,|\,\boldsymbol{x})/p_{\boldsymbol{y}}(\boldsymbol{y})$, we can write

$$\begin{aligned}\frac{\partial I}{\partial \boldsymbol{H}^*} &= -\int \nabla_{\boldsymbol{y}} p_{\boldsymbol{y}}(\boldsymbol{y}) \mathbb{E}_{\boldsymbol{x}}\left[\frac{p_{\boldsymbol{y}\,|\,\boldsymbol{x}}(\boldsymbol{y})}{p_{\boldsymbol{y}}(\boldsymbol{y})}\boldsymbol{x}^\dagger\right] d\boldsymbol{y} \\ &= -\int \nabla_{\boldsymbol{y}} p_{\boldsymbol{y}}(\boldsymbol{y}) \mathbb{E}[\boldsymbol{x}^\dagger\,|\,\boldsymbol{y}]\, d\boldsymbol{y}. \end{aligned} \tag{89}$$

Now, using

$$\begin{aligned}\nabla_{\boldsymbol{y}} p_{\boldsymbol{y}}(\boldsymbol{y}) &= \nabla_{\boldsymbol{y}} \mathbb{E}_{\boldsymbol{x}}[p_{\boldsymbol{y}\,|\,\boldsymbol{x}}(\boldsymbol{y})] = \mathbb{E}_{\boldsymbol{x}}[\nabla_{\boldsymbol{y}} p_{\boldsymbol{y}\,|\,\boldsymbol{x}}(\boldsymbol{y})] \\ &= -\mathbb{E}_{\boldsymbol{x}}[p_{\boldsymbol{y}\,|\,\boldsymbol{x}}(\boldsymbol{y})(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x})] \\ &= -\mathbb{E}[p_{\boldsymbol{y}}(\boldsymbol{y})(\boldsymbol{y}-\bar{\boldsymbol{H}}\boldsymbol{x})\,|\,\boldsymbol{y}] \\ &= -p_{\boldsymbol{y}}(\boldsymbol{y})(\boldsymbol{y}-\boldsymbol{H}\mathbb{E}[\boldsymbol{x}\,|\,\boldsymbol{y}]) \end{aligned} \tag{90}$$

we can write

$$\begin{aligned}\frac{\partial I}{\partial \boldsymbol{H}^*} &= \int p_{\boldsymbol{y}}(\boldsymbol{y})(\boldsymbol{y}-\boldsymbol{H}\mathbb{E}[\boldsymbol{x}\,|\,\boldsymbol{y}])\mathbb{E}[\boldsymbol{x}^\dagger\,|\,\boldsymbol{y}]d\boldsymbol{y} \\ &= \mathbb{E}[\boldsymbol{y}\boldsymbol{x}^\dagger] - \mathbb{E}[\bar{\boldsymbol{H}}\mathbb{E}[\boldsymbol{x}\,|\,\boldsymbol{y}]\mathbb{E}[\boldsymbol{x}^\dagger\,|\,\boldsymbol{y}]] \\ &= \boldsymbol{H}(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\dagger] - \mathbb{E}[\mathbb{E}[\boldsymbol{x}\,|\,\boldsymbol{y}]\mathbb{E}[\boldsymbol{x}^\dagger\,|\,\boldsymbol{y}]]). \end{aligned} \tag{91}$$

Finally, noting that

$$\begin{aligned}\boldsymbol{E} &= \mathbb{E}[(\boldsymbol{x}-\mathbb{E}[\boldsymbol{x}\,|\,\boldsymbol{y}])(\boldsymbol{x}-\mathbb{E}[\boldsymbol{x}\,|\,\boldsymbol{y}])^\dagger] \\ &= \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\dagger + \mathbb{E}[\boldsymbol{x}\,|\,\boldsymbol{y}]\mathbb{E}[\boldsymbol{x}^\dagger\,|\,\boldsymbol{y}] - \mathbb{E}[\boldsymbol{x}\,|\,\boldsymbol{y}]\boldsymbol{x}^\dagger - \boldsymbol{x}\mathbb{E}[\boldsymbol{x}^\dagger\,|\,\boldsymbol{y}]] \\ &= \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\dagger] - \mathbb{E}[\mathbb{E}[\boldsymbol{x}\,|\,\boldsymbol{y}]\mathbb{E}[\boldsymbol{x}^\dagger\,|\,\boldsymbol{y}]] \end{aligned} \tag{92}$$

we obtain the desired result $\partial I/\partial \boldsymbol{H}^* = \boldsymbol{H}\boldsymbol{E}$.

Interchange of Order of Integrals and Derivatives: Consider first the interchange of order in (86) given by

$$\partial \mathbb{E}_{\boldsymbol{x}}[p_{\boldsymbol{y}\,|\,\boldsymbol{x}}(\boldsymbol{y})]/\partial[\boldsymbol{H}^*]_{ij} = \mathbb{E}_{\boldsymbol{x}}[\partial p_{\boldsymbol{y}\,|\,\boldsymbol{x}}(\boldsymbol{y})/\partial[\boldsymbol{H}^*]_{ij}].$$

Simply note from (87) that

$$\partial p_{\boldsymbol{y}\,|\,\boldsymbol{x}}(\boldsymbol{y})/\partial[\boldsymbol{H}^*]_{ij} = p_{\boldsymbol{y}\,|\,\boldsymbol{x}}(\boldsymbol{y})[(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x})\boldsymbol{x}^\dagger]_{ij}$$

and then

$$\begin{aligned}\left|\frac{\partial p_{\boldsymbol{y}\,|\,\boldsymbol{x}}(\boldsymbol{y})}{\partial[\boldsymbol{H}^*]_{ij}}\right| &\le |[(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x})\boldsymbol{x}^\dagger]_{ij}| \\ &\le |y_i||x_j| + \sum_k |[\boldsymbol{H}]_{ik}||x_k||x_j| \end{aligned}$$

which is an upper bound valid for the given channel $\boldsymbol{H}$. To make the upper bound valid for a neighborhood around $\bar{\boldsymbol{H}}$, just add some ϵ to each term $|[\bar{\boldsymbol{H}}]_{ik}|$. The expectation over $\boldsymbol{x}$ of the right-hand side is clearly finite since the second-order moments are finite[10] and Lemma 2 below can be invoked.

Consider now the interchange of order in (90) given by $\nabla_{\boldsymbol{y}}\mathbb{E}_{\boldsymbol{x}}[p_{\boldsymbol{y}\,|\,\boldsymbol{x}}(\boldsymbol{y})] = \mathbb{E}_{\boldsymbol{x}}[\nabla_{\boldsymbol{y}} p_{\boldsymbol{y}\,|\,\boldsymbol{x}}(\boldsymbol{y})]$. Simply note from (87) that $\partial p_{\boldsymbol{y}\,|\,\boldsymbol{x}}(\boldsymbol{y})/\partial y_i^* = -p_{\boldsymbol{y}\,|\,\boldsymbol{x}}(\boldsymbol{y})|[\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x}]_i|$ and then

$$\left|\frac{\partial p_{\boldsymbol{y}\,|\,\boldsymbol{x}}(\boldsymbol{y})}{\partial y_i^*}\right| = |[\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x}]_i| \le |y_i| + \sum_k |\boldsymbol{H}_{ik}||x_k|.$$

As before, to guarantee that the upper bound holds on some neighborhood of $\boldsymbol{y}$ just add some small ϵ to the term $|y_i|$. The expectation over $\boldsymbol{x}$ of the right-hand side is clearly finite since the second-order moments are finite and Lemma 2 below can be invoked. □

The following result was used in the foregoing proof.

Lemma 2: Consider a function $f(\boldsymbol{x};\theta)$ and a nonnegative function $g(\boldsymbol{x})$. The relation

$$\frac{\partial}{\partial\theta}\int g(\boldsymbol{x}) f(\boldsymbol{x};\theta) d\boldsymbol{x} = \int g(\boldsymbol{x})\frac{\partial}{\partial\theta} f(\boldsymbol{x};\theta)\, d\boldsymbol{x} \tag{93}$$

holds[11] if for each θ_0 there exists a neighborhood of $\theta_0, \mathcal{N}_{\theta_0}$, and a function $M(\boldsymbol{x};\theta_0)$ such that

$$\sup_{\theta\in\mathcal{N}_{\theta_0}}\left|\frac{\partial}{\partial\theta} f(\boldsymbol{x};\theta)\right| \le M(\boldsymbol{x};\theta_0) \quad \text{a.e.}$$

with $\int g(\boldsymbol{x})M(\boldsymbol{x};\theta)d\boldsymbol{x} < \infty$.

Proof: For any θ_0, the mean value theorem [22] says that

$$f_\delta(\boldsymbol{x};\theta_0) \triangleq \frac{f(\boldsymbol{x};\theta_0+\delta) - f(\boldsymbol{x};\theta_0)}{\delta} = \frac{\partial}{\partial\theta} f(\boldsymbol{x};\tilde{\theta})$$

where $\tilde{\theta}$ is in a neighborhood of θ_0. Therefore, $f_\delta(\boldsymbol{x};\theta_0)$ can be upper-bounded as

$$|f_\delta(\boldsymbol{x};\theta_0)| \le \sup_{\theta\in\mathcal{N}_{\theta_0}}\left|\frac{\partial}{\partial\theta} f(\boldsymbol{x};\theta)\right| \le M(\boldsymbol{x};\theta_0) \quad \text{a.e.}$$

Now, if $\int g(\boldsymbol{x})M(\boldsymbol{x};\theta_0)d\boldsymbol{x} < \infty$, we can invoke the Lebesgue Dominated Convergence Theorem [22], [30] to show that

$$\lim_{\delta\to 0}\int g(\boldsymbol{x}) f_\delta(\boldsymbol{x};\theta_0) d\boldsymbol{x} = \int \lim_{\delta\to 0} g(\boldsymbol{x}) f_\delta(\boldsymbol{x};\theta_0)\, d\boldsymbol{x}. \quad \square$$

C. Proof of Theorem 2

We first give the following two chain rules which will be instrumental in proving Theorem 2.

Lemma 3:

a) Let f be a scalar real-valued function which depends on $\boldsymbol{B}$ through $\boldsymbol{H} = \boldsymbol{A}\boldsymbol{B}\boldsymbol{C}$, where $\boldsymbol{A}$ and $\boldsymbol{C}$ are arbitrary fixed matrices. Then

$$\nabla_{\boldsymbol{B}} f = \boldsymbol{A}^\dagger \nabla_{\boldsymbol{H}} f \boldsymbol{C}^\dagger. \tag{94}$$

b) Let f be a scalar real-valued function which depends on $\boldsymbol{B}$ through $\boldsymbol{R} \triangleq \boldsymbol{H}\boldsymbol{H}^\dagger = \boldsymbol{A}\boldsymbol{B}\boldsymbol{C}\boldsymbol{C}^\dagger\boldsymbol{B}^\dagger\boldsymbol{A}^\dagger$, where $\boldsymbol{A}$ and $\boldsymbol{C}$ are arbitrary fixed matrices. Then

$$\nabla_{\boldsymbol{B}} f = \boldsymbol{A}^\dagger \nabla_{\boldsymbol{R}} f \boldsymbol{A}\boldsymbol{B}\boldsymbol{C}\boldsymbol{C}^\dagger. \tag{95}$$

Proof: Consider the first chain rule. Given $[\boldsymbol{H}^\dagger]_{ij} = \mathrm{Tr}(\boldsymbol{C}^\dagger\boldsymbol{B}^\dagger\boldsymbol{A}^\dagger\boldsymbol{e}_j\boldsymbol{e}_i^\dagger)$, it follows that

$$\begin{aligned}\partial[\boldsymbol{H}^\dagger]_{ij}/\partial\boldsymbol{B}^* &= \boldsymbol{A}^\dagger\boldsymbol{e}_j\boldsymbol{e}_i^\dagger\boldsymbol{C}^\dagger, \partial[\boldsymbol{H}^\dagger]_{ij}/\partial[\boldsymbol{B}^*]_{kl} \\ &= \boldsymbol{e}_k^\dagger\boldsymbol{A}^\dagger\boldsymbol{e}_j\boldsymbol{e}_i^\dagger\boldsymbol{C}^\dagger\boldsymbol{e}_l \\ &= \boldsymbol{e}_i^\dagger\boldsymbol{C}^\dagger\boldsymbol{e}_l\boldsymbol{e}_k^\dagger\boldsymbol{A}^\dagger\boldsymbol{e}_j \end{aligned}$$

[10] Observe that if $\int |x|^2 f(x)\,dx$ is finite, so is $\int |x| f(x) dx$ [30]. Also, by Hölder's inequality (for $p = q = 2$) $\mathbb{E}[|x||y|] \le \mathbb{E}^{1/2}[|x|^2]\mathbb{E}^{1/2}[|y|^2]$ [30], if $\mathbb{E}[|x|^2]$ and $\mathbb{E}[|y|^2]$ are finite, so is $\mathbb{E}[|x||y|]$.

[11] It is implicitly assumed in (93) that $\int g(\boldsymbol{x}) f(\boldsymbol{x};\theta)d\boldsymbol{x}$ and $\partial f(\boldsymbol{x};\theta)/\partial\theta$ exist.

and, more compactly, $\partial \boldsymbol{H}^\dagger / \partial [\boldsymbol{B}^*]_{kl} = \boldsymbol{C}^\dagger \boldsymbol{e}_l \boldsymbol{e}_k^\dagger \boldsymbol{A}^\dagger$ (note also that $\partial \bar{\boldsymbol{H}} / \partial [\boldsymbol{B}^*]_{kl} = \boldsymbol{0}$). Now, we can use the chain rule

$$\begin{aligned} \frac{\partial f}{\partial [\boldsymbol{B}^*]_{kl}} &= \mathrm{Tr}\left(\left(\frac{\partial f}{\partial \boldsymbol{H}^\dagger} \right)^T \frac{\partial \boldsymbol{H}^\dagger}{\partial [\boldsymbol{B}^*]_{kl}} \right) \\ &= \mathrm{Tr}\left(\boldsymbol{e}_k^\dagger \boldsymbol{A}^\dagger \frac{\partial f}{\partial \bar{\boldsymbol{H}}^*} \boldsymbol{C}^\dagger \boldsymbol{e}_l \right) \end{aligned} \tag{96}$$

to obtain the desired result $\partial f / \partial \boldsymbol{B}^* = \boldsymbol{A}^\dagger \partial f / \partial \boldsymbol{H}^* \boldsymbol{C}^\dagger$.

Consider now the second chain rule. Given

$$[\boldsymbol{R}]_{ij} = \mathrm{Tr}(\boldsymbol{A}\boldsymbol{B}\boldsymbol{C}\boldsymbol{C}^\dagger \boldsymbol{B}^\dagger \boldsymbol{A}^\dagger \boldsymbol{e}_j \boldsymbol{e}_i^\dagger)$$

it follows that

$$\begin{aligned} \partial [\boldsymbol{R}]_{ij} / \partial \boldsymbol{B}^* &= \boldsymbol{A}^\dagger \boldsymbol{e}_j \boldsymbol{e}_i^\dagger \boldsymbol{A}\boldsymbol{B}\boldsymbol{C}\boldsymbol{C}^\dagger \\ \partial [\boldsymbol{R}]_{ij} / \partial [\boldsymbol{B}^*]_{kl} &= \boldsymbol{e}_k^\dagger \boldsymbol{A}^\dagger \boldsymbol{e}_j \boldsymbol{e}_i^\dagger \boldsymbol{A}\boldsymbol{B}\boldsymbol{C}\boldsymbol{C}^\dagger \boldsymbol{e}_l \\ &= \boldsymbol{e}_i^\dagger \boldsymbol{A}\boldsymbol{B}\boldsymbol{C}\boldsymbol{C}^\dagger \boldsymbol{e}_l \boldsymbol{e}_k^\dagger \boldsymbol{A}^\dagger \boldsymbol{e}_j \end{aligned}$$

and, more compactly, $\partial \boldsymbol{R} / \partial [\boldsymbol{B}^*]_{kl} = \boldsymbol{A}\boldsymbol{B}\boldsymbol{C}\boldsymbol{C}^\dagger \boldsymbol{e}_l \boldsymbol{e}_k^\dagger \boldsymbol{A}^\dagger$. Now, we can use the chain rule (noticing that $\bar{\boldsymbol{R}}^\dagger = \bar{\boldsymbol{R}}$):

$$\begin{aligned} \frac{\partial f}{\partial [\boldsymbol{B}^*]_{kl}} &= \mathrm{Tr}\left(\left(\frac{\partial f}{\partial \boldsymbol{R}} \right)^T \frac{\partial \boldsymbol{R}}{\partial [\boldsymbol{B}^*]_{kl}} \right) \\ &= \mathrm{Tr}\left(\frac{\partial f}{\partial \boldsymbol{R}^*} \boldsymbol{A}\boldsymbol{B}\boldsymbol{C}\boldsymbol{C}^\dagger \boldsymbol{e}_l \boldsymbol{e}_k^\dagger \boldsymbol{A}^\dagger \right) \\ &= \boldsymbol{e}_k^\dagger \boldsymbol{A}^\dagger \frac{\partial f}{\partial \boldsymbol{R}^*} \boldsymbol{A}\boldsymbol{B}\boldsymbol{C}\boldsymbol{C}^\dagger \boldsymbol{e}_l, \end{aligned} \tag{97}$$

to obtain the desired result

$$\frac{\partial f}{\partial \boldsymbol{B}^*} = \boldsymbol{A}^\dagger \frac{\partial f}{\partial \boldsymbol{R}^*} \boldsymbol{A}\boldsymbol{B}\boldsymbol{C}\boldsymbol{C}^\dagger. \tag{98}$$

□

Proof of Theorem 2: First, write the equivalent whitened received signal $\tilde{\boldsymbol{y}} = \boldsymbol{\Sigma}_n^{-1/2} \boldsymbol{y}$ (note that both the mutual information and the MMSE remain unchanged by this invertible transformation) as

$$\tilde{\boldsymbol{y}} = \boldsymbol{\Sigma}_n^{-1/2} \boldsymbol{H}\boldsymbol{B}\boldsymbol{x} + \tilde{\boldsymbol{n}} \tag{99}$$

where $\tilde{\boldsymbol{n}} = \boldsymbol{\Sigma}_n^{-1/2} \boldsymbol{n}$ is the normalized version of $\boldsymbol{n}$ (assuming that $\boldsymbol{\Sigma}_n = \boldsymbol{\Sigma}_n^{1/2} \boldsymbol{\Sigma}_n^{\dagger/2}$). Now, we can define $\tilde{\boldsymbol{H}} = \boldsymbol{\Sigma}_n^{-1/2} \boldsymbol{H}\boldsymbol{B}$ and invoke Theorem 1 to obtain

$$\nabla_{\tilde{\boldsymbol{H}}} I = \tilde{\boldsymbol{H}} \boldsymbol{E} = \boldsymbol{\Sigma}_n^{-1/2} \boldsymbol{H}\boldsymbol{B}\boldsymbol{E}. \tag{100}$$

By Lemma 3 a), we have $\nabla_{\boldsymbol{H}} I = \boldsymbol{\Sigma}_n^{-1/2} \nabla_{\tilde{\boldsymbol{H}}} I \boldsymbol{B}^\dagger$ and $\nabla_{\boldsymbol{B}} I = \boldsymbol{H}^\dagger \boldsymbol{\Sigma}_n^{-1/2} \nabla_{\tilde{\boldsymbol{H}}} I$, from which (21) and (22) are obtained, respectively.

Also, by Lemma 3 b), we have $\nabla_{\boldsymbol{B}} f = \nabla_{\boldsymbol{Q}} f \boldsymbol{B} \boldsymbol{\Sigma}_x$ and $\nabla_{\boldsymbol{B}} f = \nabla_{\boldsymbol{Q}_B} f \boldsymbol{B}$, from which (23) and (24) are obtained, respectively.

To obtain (25), simply invoke Lemma 3 a) from which $\nabla_{\boldsymbol{\Sigma}_x} I = \boldsymbol{B}^\dagger \nabla_{\boldsymbol{Q}} I \boldsymbol{B}$.

To derive (26), first invoke Lemma 3 a) to obtain

$$\nabla_{\boldsymbol{\Sigma}_n^{-1/2}} I = \nabla_{\tilde{\boldsymbol{H}}} I \, \boldsymbol{B}^\dagger \boldsymbol{H}^\dagger = \boldsymbol{\Sigma}_n^{-1/2} \boldsymbol{H}\boldsymbol{B}\boldsymbol{E}\boldsymbol{B}^\dagger \boldsymbol{H}^\dagger$$

and then invoke Lemma 3 b) (noting that $\boldsymbol{\Sigma}_n = \boldsymbol{\Sigma}_n^{1/2} \boldsymbol{\Sigma}_n^{\dagger/2}$ and $\boldsymbol{\Sigma}_n^{-1} = \boldsymbol{\Sigma}_n^{-\dagger/2} \boldsymbol{\Sigma}_n^{-1/2}$) to obtain

$$\nabla_{\boldsymbol{\Sigma}_n^{-1/2}} I = \boldsymbol{\Sigma}_n^{-1/2} \nabla_{\boldsymbol{\Sigma}_n^{-1}} I$$

which gives $\nabla_{\boldsymbol{\Sigma}_n^{-1}} I = \boldsymbol{H}\boldsymbol{B}\boldsymbol{E}\boldsymbol{B}^\dagger \boldsymbol{H}^\dagger$.

Finally, to obtain (27), simply use

$$\partial f / \partial \boldsymbol{X}^* = -\boldsymbol{X}^{-1} (\partial f / \partial \boldsymbol{X}^{-*}) \boldsymbol{X}^{-1}$$

for a positive definite (Hermitian) matrix $\boldsymbol{X}$ [15] to get $\nabla_{\boldsymbol{\Sigma}_n} I = -\boldsymbol{\Sigma}_n^{-1} \nabla_{\boldsymbol{\Sigma}_n^{-1}} I \boldsymbol{\Sigma}_n^{-1}$.

D. Gradient of Mutual Information for Arbitrary $\boldsymbol{\Sigma}_x$

Define $f_n(\boldsymbol{\Sigma}_x) = I(\boldsymbol{\Sigma}_x + \epsilon_n \boldsymbol{I})$ where $\epsilon_n \to 0$ as $n \to \infty$. Clearly, $f_n(\boldsymbol{\Sigma}_x) \to I(\boldsymbol{\Sigma}_x)$ and (from Theorem 2)

$$\nabla_{\boldsymbol{\Sigma}_x} f_n(\boldsymbol{\Sigma}_x) = \nabla_{\boldsymbol{\Sigma}_x} I(\boldsymbol{\Sigma}_x + \epsilon_n \boldsymbol{I}) = \boldsymbol{H}^\dagger \boldsymbol{H} \boldsymbol{E} (\boldsymbol{\Sigma}_x + \epsilon_n \boldsymbol{I})^{-1} \tag{101}$$

where $\boldsymbol{E}$ is the MMSE matrix when the transmit covariance matrix is $\boldsymbol{\Sigma}_x + \epsilon_n \boldsymbol{I}$. To prove (38), it suffices to show that $\nabla_{\boldsymbol{\Sigma}_x} \lim_{n\to\infty} f_n(\boldsymbol{\Sigma}_x) = \lim_{n\to\infty} \nabla_{\boldsymbol{\Sigma}_x} f_n(\boldsymbol{\Sigma}_x)$, which holds if $\boldsymbol{E}(\boldsymbol{\Sigma}_x + \epsilon_n \boldsymbol{I})^{-1}$ is uniformly continuous on a set containing $\boldsymbol{\Sigma}_x$ [22]. Since we are only interested in the point $\boldsymbol{\Sigma}_x$, we can choose a convenient compact set containing $\boldsymbol{\Sigma}_x$ so that uniform continuity is implied by pointwise continuity [22]. Therefore, we just need to show that $\boldsymbol{E}(\boldsymbol{\Sigma}_x + \epsilon_n \boldsymbol{I})^{-1}$ has a limit.

Indeed, if $\boldsymbol{\Sigma}_x$ is nonsingular, then $\boldsymbol{E}(\boldsymbol{\Sigma}_x + \epsilon_n \boldsymbol{I})^{-1} \to \boldsymbol{E} \boldsymbol{\Sigma}_x^{-1}$; if $\boldsymbol{\Sigma}_x = \boldsymbol{0}$, then $\boldsymbol{E}(\boldsymbol{\Sigma}_x + \epsilon_n \boldsymbol{I})^{-1} \to \boldsymbol{I}$ (as shown next); and in general we can work with the subspaces of $\boldsymbol{\Sigma}_x$ associated to null eigenvalues and positive eigenvalues to obtain a limit. To deal with the case $\boldsymbol{\Sigma}_x = \boldsymbol{0}$, define the normalized vector $\bar{\boldsymbol{x}} = \boldsymbol{x} / \sqrt{\epsilon_n}$, such that its covariance matrix is equal to the identity matrix, and define the MMSE of $\bar{\boldsymbol{x}}$ as $\bar{\boldsymbol{E}}$, which satisfies $\bar{\boldsymbol{E}} = \boldsymbol{E} / \epsilon_n$ (observe that $\bar{\boldsymbol{E}}$ can be interpreted as the MMSE matrix of a communication with noise $\boldsymbol{n} / \sqrt{\epsilon_n}$). Then, $\boldsymbol{E}(\epsilon_n \boldsymbol{I})^{-1} = \bar{\boldsymbol{E}}$ and the desired results follows from $\bar{\boldsymbol{E}} \to \boldsymbol{I}$, since $\mathbb{E}[\bar{\boldsymbol{x}} \,|\, \boldsymbol{y}] \to \mathbb{E}[\bar{\boldsymbol{x}}]$.

E. Proof of Theorem 4

Consider the signal model $\boldsymbol{y} = \boldsymbol{z} + \boldsymbol{T}^{1/2} \boldsymbol{n}$, where $\boldsymbol{T} = \boldsymbol{T}^{1/2} \boldsymbol{T}^{\dagger/2} (T^{\dagger/2} \triangleq (T^{1/2})^\dagger$. Equivalently, we can write $\boldsymbol{y} = \boldsymbol{z} + \tilde{\boldsymbol{n}}$, where we have defined the equivalent noise vector as $\tilde{\boldsymbol{n}} = \boldsymbol{T}^{1/2} \boldsymbol{n}$ with covariance matrix $\tilde{\boldsymbol{\Sigma}}_n = \boldsymbol{T}^{1/2} \boldsymbol{\Sigma}_n \boldsymbol{T}^{\dagger/2}$. Noting that [13]

$$\begin{aligned} h(\boldsymbol{z} + \tilde{\boldsymbol{n}}) &= I(\boldsymbol{z}; \boldsymbol{y}) + h(\boldsymbol{z} + \tilde{\boldsymbol{n}} \,|\, \boldsymbol{z}) \\ &= I(\boldsymbol{z}; \boldsymbol{y}) + \log \det(\pi e \tilde{\boldsymbol{\Sigma}}_n) \end{aligned} \tag{102}$$

we have that

$$\begin{aligned} \frac{\partial}{\partial \tilde{\boldsymbol{\Sigma}}_n^*} h(\boldsymbol{z} + \tilde{\boldsymbol{n}}) &= \frac{\partial}{\partial \tilde{\boldsymbol{\Sigma}}_n^*} I(\boldsymbol{z}; \boldsymbol{y}) + \frac{\partial}{\partial \tilde{\boldsymbol{\Sigma}}_n^*} \log \det(\pi e \tilde{\boldsymbol{\Sigma}}_n) \\ &= -\tilde{\boldsymbol{\Sigma}}_n^{-1} \boldsymbol{E} \tilde{\boldsymbol{\Sigma}}_n^{-1} + \tilde{\boldsymbol{\Sigma}}_n^{-1} \\ &= \tilde{\boldsymbol{\Sigma}}_n^{-1} \left(\boldsymbol{I} - \boldsymbol{E} \tilde{\boldsymbol{\Sigma}}_n^{-1} \right) \end{aligned} \tag{103}$$

where we have used (27). Now, invoking Lemma 3 b) twice we get $\nabla_{\boldsymbol{T}^{1/2}} f = \nabla_{\tilde{\boldsymbol{\Sigma}}_n} f \boldsymbol{T}^{1/2} \boldsymbol{\Sigma}_n$ and $\nabla_{\boldsymbol{T}^{1/2}} f = \nabla_{\boldsymbol{T}} f \boldsymbol{T}^{1/2}$, which implies $\nabla_{\tilde{\boldsymbol{\Sigma}}_n} f \tilde{\boldsymbol{\Sigma}}_n = \nabla_{\boldsymbol{T}} f \boldsymbol{T}$. Hence, we can finally write

$$\begin{aligned} \nabla_{\boldsymbol{T}} h(\boldsymbol{z} + \tilde{\boldsymbol{n}}) &= \nabla_{\tilde{\boldsymbol{\Sigma}}_n} h(\boldsymbol{z} + \tilde{\boldsymbol{n}}) \tilde{\boldsymbol{\Sigma}}_n \boldsymbol{T}^{-1} \\ &= \tilde{\boldsymbol{\Sigma}}_n^{-1} (\tilde{\boldsymbol{\Sigma}}_n - \boldsymbol{E}) \boldsymbol{T}^{-1}. \end{aligned} \tag{104}$$

Now, we compute $\boldsymbol{J}(\boldsymbol{y}) = \mathbb{E}_{\boldsymbol{y}}[\nabla \log p_{\boldsymbol{y}}(\boldsymbol{y}) \nabla^\dagger \log p_{\boldsymbol{y}}(\boldsymbol{y})]$ as follows (for simplicity of notation and without affecting the result we now consider that the noise is zero mean). First, obtain the inner gradient (from (90)) as

$$\nabla \log p_{\boldsymbol{y}}(\boldsymbol{y}) = \frac{1}{p_{\boldsymbol{y}}(\boldsymbol{y})} \nabla p_{\boldsymbol{y}}(\boldsymbol{y}) = -\tilde{\boldsymbol{\Sigma}}_n^{-1}(\boldsymbol{y} - \mathbb{E}[\boldsymbol{z} \,|\, \boldsymbol{y}]) \quad (105)$$

and then

$$\begin{aligned}
\boldsymbol{J}(\boldsymbol{y}) &= \tilde{\boldsymbol{\Sigma}}_n^{-1} \mathbb{E}[(\boldsymbol{y} - \mathbb{E}[\boldsymbol{z} \,|\, \boldsymbol{y}])(\boldsymbol{y} - \mathbb{E}[\boldsymbol{z} \,|\, \boldsymbol{y}])^\dagger] \tilde{\boldsymbol{\Sigma}}_n^{-1} \\
&= \tilde{\boldsymbol{\Sigma}}_n^{-1} \mathbb{E}[((\boldsymbol{z} - \mathbb{E}[\boldsymbol{z} \,|\, \boldsymbol{y}]) + \tilde{\boldsymbol{n}})((\boldsymbol{z} - \mathbb{E}[\boldsymbol{z} \,|\, \boldsymbol{y}]) + \tilde{\boldsymbol{n}})^\dagger] \tilde{\boldsymbol{\Sigma}}_n^{-1} \\
&= \tilde{\boldsymbol{\Sigma}}_n^{-1} \Big(\mathbb{E}[(\boldsymbol{z} - \mathbb{E}[\boldsymbol{z} \,|\, \boldsymbol{y}])(\boldsymbol{z} - \mathbb{E}[\boldsymbol{z} \,|\, \boldsymbol{y}])^\dagger] + \tilde{\boldsymbol{\Sigma}}_n \\
&\quad + \mathbb{E}[(\boldsymbol{z} - \mathbb{E}[\boldsymbol{z} \,|\, \boldsymbol{y}]) \tilde{\boldsymbol{n}}^\dagger] + \mathbb{E}[\tilde{\boldsymbol{n}} (\boldsymbol{z} - \mathbb{E}[\boldsymbol{z} \,|\, \boldsymbol{y}])^\dagger] \Big) \tilde{\boldsymbol{\Sigma}}_n^{-1} \\
&= \tilde{\boldsymbol{\Sigma}}_n^{-1} (\tilde{\boldsymbol{\Sigma}}_n - \boldsymbol{E}) \tilde{\boldsymbol{\Sigma}}_n^{-1} \qquad (106)
\end{aligned}$$

where we have used

$$\begin{aligned}
\mathbb{E}[\tilde{\boldsymbol{n}} (\boldsymbol{z} - \mathbb{E}[\boldsymbol{z} \,|\, \boldsymbol{y}])^\dagger] &= \mathbb{E}[(\boldsymbol{y} - \boldsymbol{z})(\boldsymbol{z} - \mathbb{E}[\boldsymbol{z} \,|\, \boldsymbol{y}])^\dagger] \\
&= -\mathbb{E}[\boldsymbol{z} (\boldsymbol{z} - \mathbb{E}[\boldsymbol{z} \,|\, \boldsymbol{y}])^\dagger] \\
&= -\mathbb{E}[(\boldsymbol{z} - \mathbb{E}[\boldsymbol{z} \,|\, \boldsymbol{y}])(\boldsymbol{z} - \mathbb{E}[\boldsymbol{z} \,|\, \boldsymbol{y}])^\dagger] \\
&= -\boldsymbol{E}. \qquad (107)
\end{aligned}$$

Combining both results, we can finally write

$$\frac{\partial}{\partial \boldsymbol{T}^*} h(\boldsymbol{z} + \boldsymbol{T}^{1/2} \boldsymbol{n}) = \boldsymbol{J}(\boldsymbol{y}) \tilde{\boldsymbol{\Sigma}}_n \boldsymbol{T}^{-1} = \boldsymbol{J}(\boldsymbol{y}) \boldsymbol{T}^{1/2} \boldsymbol{\Sigma}_n \boldsymbol{T}^{-1/2} \quad (108)$$

which is the desired result in (56). Now, (57) is readily obtained particularizing for $\boldsymbol{\Sigma}_n = \boldsymbol{I}$, which is equivalent to the expression in (55).

F. Proof of Theorem 5

This proof is a refinement of the proof of Theorem 1 in Appendix B, so we will skip many details. We will assume the noise to be zero mean without loss of generality.

The divergence of interest is

$$D(p_{\boldsymbol{y} \,|\, \boldsymbol{x} = \boldsymbol{x}_0} \,\|\, p_{\boldsymbol{y}}) = -n_R \log(\pi e) - \mathbb{E}[\log p_{\boldsymbol{y}}(\boldsymbol{y}) \,|\, \boldsymbol{x} = \boldsymbol{x}_0]. \quad (109)$$

Then

$$\begin{aligned}
&\nabla_{\boldsymbol{H}} D(p_{\boldsymbol{y} \,|\, \boldsymbol{x} = \boldsymbol{x}_0} \,\|\, p_{\boldsymbol{y}}) \\
&\quad = -\frac{\partial}{\partial \boldsymbol{H}^*} \int p_{\boldsymbol{y} \,|\, \boldsymbol{x} = \boldsymbol{x}_0}(\boldsymbol{y}) \log p_{\boldsymbol{y}}(\boldsymbol{y}) \, d\boldsymbol{y} \\
&\quad = -\int \left(\frac{\partial p_{\boldsymbol{y} \,|\, \boldsymbol{x} = \boldsymbol{x}_0}(\boldsymbol{y})}{\partial \boldsymbol{H}^*} \log p_{\boldsymbol{y}}(\boldsymbol{y}) \right. \\
&\qquad \left. + p_{\boldsymbol{y} \,|\, \boldsymbol{x} = \boldsymbol{x}_0}(\boldsymbol{y}) \frac{\partial \log p_{\boldsymbol{y}}(\boldsymbol{y})}{\partial \boldsymbol{H}^*} \right) d\boldsymbol{y} \qquad (110)
\end{aligned}$$

where the interchange of the derivative and the integral follows from Lebesgue Dominated Convergence Theorem and the finiteness of $\boldsymbol{x}_0$.

Now, expanding the derivatives as in Appendix B by using

$$\nabla_{\boldsymbol{H}} p_{\boldsymbol{y} \,|\, \boldsymbol{x} = \boldsymbol{x}_0}(\boldsymbol{y}) = p_{\boldsymbol{y} \,|\, \boldsymbol{x} = \boldsymbol{x}_0}(\boldsymbol{y})(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}_0) \boldsymbol{x}_0^\dagger \quad (111)$$

and

$$\nabla_{\boldsymbol{y}} p_{\boldsymbol{y} \,|\, \boldsymbol{x} = \boldsymbol{x}_0}(\boldsymbol{y}) = -p_{\boldsymbol{y} \,|\, \boldsymbol{x} = \boldsymbol{x}_0}(\boldsymbol{y})(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}_0) \quad (112)$$

we obtain

$$\begin{aligned}
&\nabla_{\boldsymbol{H}} D(p_{\boldsymbol{y} \,|\, \boldsymbol{x} = \boldsymbol{x}_0} \,\|\, p_{\boldsymbol{y}}) \\
&\quad = \int \nabla_{\boldsymbol{y}} p_{\boldsymbol{y} \,|\, \boldsymbol{x} = \boldsymbol{x}_0}(\boldsymbol{y}) \boldsymbol{x}_0^\dagger \log p_{\boldsymbol{y}}(\boldsymbol{y}) d\boldsymbol{y} \\
&\qquad + \int \frac{p_{\boldsymbol{y} \,|\, \boldsymbol{x} = \boldsymbol{x}_0}(\boldsymbol{y})}{p_{\boldsymbol{y}}(\boldsymbol{y})} \mathbb{E}_{\boldsymbol{x}}[\nabla_{\boldsymbol{y}} p_{\boldsymbol{y} \,|\, \boldsymbol{x} = \boldsymbol{x}}(\boldsymbol{y}) \boldsymbol{x}^\dagger] \, d\boldsymbol{y} \\
&\quad = \int \log p_{\boldsymbol{y}}(\boldsymbol{y}) \nabla_{\boldsymbol{y}} p_{\boldsymbol{y} \,|\, \boldsymbol{x} = \boldsymbol{x}_0}(\boldsymbol{y}) d\boldsymbol{y} \boldsymbol{x}_0^\dagger \\
&\qquad + \mathbb{E}_{\boldsymbol{x}} \left[\int \frac{p_{\boldsymbol{y} \,|\, \boldsymbol{x} = \boldsymbol{x}_0}(\boldsymbol{y})}{p_{\boldsymbol{y}}(\boldsymbol{y})} \nabla_{\boldsymbol{y}} p_{\boldsymbol{y} \,|\, \boldsymbol{x} = \boldsymbol{x}}(\boldsymbol{y}) \, d\boldsymbol{y} \boldsymbol{x}^\dagger \right] \\
&\quad = -\int p_{\boldsymbol{y} \,|\, \boldsymbol{x} = \boldsymbol{x}_0}(\boldsymbol{y}) \nabla_{\boldsymbol{y}} \log p_{\boldsymbol{y}}(\boldsymbol{y}) d\boldsymbol{y} \boldsymbol{x}_0^\dagger \\
&\qquad - \mathbb{E}_{\boldsymbol{x}} \left[\int p_{\boldsymbol{y} \,|\, \boldsymbol{x} = \boldsymbol{x}}(\boldsymbol{y}) \nabla_{\boldsymbol{y}} \left(\frac{p_{\boldsymbol{y} \,|\, \boldsymbol{x} = \boldsymbol{x}_0}(\boldsymbol{y})}{p_{\boldsymbol{y}}(\boldsymbol{y})} \right) d\boldsymbol{y} \boldsymbol{x}^\dagger \right] \qquad (113)
\end{aligned}$$

where the last equality follows from integrating by parts [22] as in Appendix B.

Proceeding, we get

$$\begin{aligned}
&\nabla_{\boldsymbol{H}} D(p_{\boldsymbol{y} \,|\, \boldsymbol{x} = \boldsymbol{x}_0} \,\|\, p_{\boldsymbol{y}}) \\
&\quad = -\int \frac{p_{\boldsymbol{y} \,|\, \boldsymbol{x} = \boldsymbol{x}_0}(\boldsymbol{y})}{p_{\boldsymbol{y}}(\boldsymbol{y})} \nabla_{\boldsymbol{y}} p_{\boldsymbol{y}}(\boldsymbol{y}) \boldsymbol{x}_0^\dagger \, d\boldsymbol{y} \\
&\qquad - \int \nabla_{\boldsymbol{y}} p_{\boldsymbol{y} \,|\, \boldsymbol{x} = \boldsymbol{x}_0}(\boldsymbol{y}) \mathbb{E}_{\boldsymbol{x}} \left[\frac{p_{\boldsymbol{y} \,|\, \boldsymbol{x} = \boldsymbol{x}}(\boldsymbol{y})}{p_{\boldsymbol{y}}(\boldsymbol{y})} \boldsymbol{x}^\dagger \right] d\boldsymbol{y} \\
&\qquad + \int \frac{\nabla_{\boldsymbol{y}} p_{\boldsymbol{y}}(\boldsymbol{y})}{p_{\boldsymbol{y}}(\boldsymbol{y})} p_{\boldsymbol{y} \,|\, \boldsymbol{x} = \boldsymbol{x}_0}(\boldsymbol{y}) \mathbb{E}_{\boldsymbol{x}} \left[\frac{p_{\boldsymbol{y} \,|\, \boldsymbol{x} = \boldsymbol{x}}(\boldsymbol{y})}{p_{\boldsymbol{y}}(\boldsymbol{y})} \boldsymbol{x}^\dagger \right] d\boldsymbol{y} \\
&\quad = -\int \frac{\nabla_{\boldsymbol{y}} p_{\boldsymbol{y}}(\boldsymbol{y})}{p_{\boldsymbol{y}}(\boldsymbol{y})} p_{\boldsymbol{y} \,|\, \boldsymbol{x} = \boldsymbol{x}_0}(\boldsymbol{y}) \boldsymbol{x}_0^\dagger d\boldsymbol{y} \\
&\qquad - \int \nabla_{\boldsymbol{y}} p_{\boldsymbol{y} \,|\, \boldsymbol{x} = \boldsymbol{x}_0}(\boldsymbol{y}) \mathbb{E}[\boldsymbol{x}^\dagger \,|\, \boldsymbol{y}] d\boldsymbol{y} \\
&\qquad + \int \frac{\nabla_{\boldsymbol{y}} p_{\boldsymbol{y}}(\boldsymbol{y})}{p_{\boldsymbol{y}}(\boldsymbol{y})} p_{\boldsymbol{y} \,|\, \boldsymbol{x} = \boldsymbol{x}_0}(\boldsymbol{y}) \mathbb{E}[\boldsymbol{x}^\dagger \,|\, \boldsymbol{y}] d\boldsymbol{y} \\
&\quad = -\int \frac{\nabla_{\boldsymbol{y}} p_{\boldsymbol{y}}(\boldsymbol{y})}{p_{\boldsymbol{y}}(\boldsymbol{y})} p_{\boldsymbol{y} \,|\, \boldsymbol{x} = \boldsymbol{x}_0}(\boldsymbol{y}) (\boldsymbol{x}_0 - \mathbb{E}[\boldsymbol{x} \,|\, \boldsymbol{y}])^\dagger \, d\boldsymbol{y} \\
&\qquad - \int \nabla_{\boldsymbol{y}} p_{\boldsymbol{y} \,|\, \boldsymbol{x} = \boldsymbol{x}_0}(\boldsymbol{y}) \mathbb{E}[\boldsymbol{x}^\dagger \,|\, \boldsymbol{y}] \, d\boldsymbol{y} \\
&\quad = \int \mathbb{E}[\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x} \,|\, \boldsymbol{y}] p_{\boldsymbol{y} \,|\, \boldsymbol{x} = \boldsymbol{x}_0}(\boldsymbol{y}) (\boldsymbol{x}_0 - \mathbb{E}[\boldsymbol{x} \,|\, \boldsymbol{y}])^\dagger d\boldsymbol{y} \\
&\qquad + \int p_{\boldsymbol{y} \,|\, \boldsymbol{x} = \boldsymbol{x}_0}(\boldsymbol{y}) (\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}_0) \mathbb{E}[\boldsymbol{x}^\dagger \,|\, \boldsymbol{y}] d\boldsymbol{y} \qquad (114)
\end{aligned}$$

where we have used $\nabla_{\boldsymbol{y}} p_{\boldsymbol{y}}(\boldsymbol{y}) = -p_{\boldsymbol{y}}(\boldsymbol{y}) \mathbb{E}[\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x} \,|\, \boldsymbol{y}]$ in the last equality.

Finally, expanding the products, we obtain the desired result

$$\begin{aligned}
&\nabla_{\boldsymbol{H}} D(p_{\boldsymbol{y} \,|\, \boldsymbol{x} = \boldsymbol{x}_0} \,\|\, p_{\boldsymbol{y}}) \\
&\quad = \mathbb{E}[(\boldsymbol{y} - \boldsymbol{H}\mathbb{E}[\boldsymbol{x} \,|\, \boldsymbol{y}])(\boldsymbol{x}_0 - \mathbb{E}[\boldsymbol{x} \,|\, \boldsymbol{y}])^\dagger \\
&\qquad + (\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}_0) \mathbb{E}[\boldsymbol{x}^\dagger \,|\, \boldsymbol{y}] \,|\, \boldsymbol{x} = \boldsymbol{x}_0] \\
&\quad = \mathbb{E}[\boldsymbol{y}\boldsymbol{x}_0^\dagger + \boldsymbol{H}\mathbb{E}[\boldsymbol{x} \,|\, \boldsymbol{y}] \mathbb{E}[\boldsymbol{x}^\dagger \,|\, \boldsymbol{y}] \\
&\qquad - \boldsymbol{H}\mathbb{E}[\boldsymbol{x} \,|\, \boldsymbol{y}] \boldsymbol{x}_0^\dagger - \boldsymbol{H}\boldsymbol{x}_0 \mathbb{E}[\boldsymbol{x}^\dagger \,|\, \boldsymbol{y}] \,|\, \boldsymbol{x} = \boldsymbol{x}_0] \\
&\quad = \boldsymbol{H}\mathbb{E}[(\boldsymbol{x}_0 - \mathbb{E}[\boldsymbol{x} \,|\, \boldsymbol{y}])(\boldsymbol{x}_0 - \mathbb{E}[\boldsymbol{x} \,|\, \boldsymbol{y}])^\dagger \,|\, \boldsymbol{x} = \boldsymbol{x}_0]. \qquad (115)
\end{aligned}$$

□

G. Proof of Theorem 6

The proof is a direct consequence of Theorem 1 as we now show. Alternatively, it can also be directly proved without resorting to Theorem 1 (as in [6, Theorem 5]), but the proof becomes tedious and is omitted for lack of additional interest.

Since for an arbitrary $p_{\boldsymbol{y}'}$ such that $p_{\boldsymbol{y}} \ll p_{\boldsymbol{y}'}$

$$\begin{aligned} I(\boldsymbol{x};\boldsymbol{y}) &= D(p_{\boldsymbol{y}\,|\,\boldsymbol{x}} \,\|\, p_{\boldsymbol{y}} \,|\, p_{\boldsymbol{x}}) \\ &= D(p_{\boldsymbol{y}\,|\,\boldsymbol{x}} \,\|\, p_{\boldsymbol{y}'} \,|\, p_{\boldsymbol{x}}) - D(p_{\boldsymbol{y}} \,\|\, p_{\boldsymbol{y}'}) \end{aligned}$$

we can choose $p_{\boldsymbol{y}'} = p_{\boldsymbol{y}\,|\,\boldsymbol{x}=\boldsymbol{x}_0}$ to express

$$\begin{aligned} D(p_{\boldsymbol{y}} \,\|\, p_{\boldsymbol{y}\,|\,\boldsymbol{x}=\boldsymbol{x}_0}) &= D(p_{\boldsymbol{y}\,|\,\boldsymbol{x}} \,\|\, p_{\boldsymbol{y}\,|\,\boldsymbol{x}=\boldsymbol{x}_0} \,|\, p_{\boldsymbol{x}}) - I(\boldsymbol{x};\boldsymbol{y}) \\ &= \mathbb{E}[(\boldsymbol{x}-\boldsymbol{x}_0)^\dagger \boldsymbol{H}^\dagger \boldsymbol{H}(\boldsymbol{x}-\boldsymbol{x}_0)] - I(\boldsymbol{x};\boldsymbol{y}) \end{aligned} \tag{116}$$

where the second expression follows simply from the explicit evaluation of the divergence between two proper complex Gaussian distributions (with arbitrary units) [16, eq. (59)]

$$\begin{aligned} &D(\mathcal{CN}(\boldsymbol{m}_1,\boldsymbol{\Sigma}_1),\mathcal{CN}(\boldsymbol{m}_0,\boldsymbol{\Sigma}_0)) \\ &\quad = \log\det(\boldsymbol{\Sigma}_0) - \log\det(\boldsymbol{\Sigma}_1) \\ &\qquad + (\boldsymbol{m}_1-\boldsymbol{m}_0)^\dagger \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{m}_1-\boldsymbol{m}_0)\log e \\ &\qquad + \mathrm{Tr}(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Sigma}_1 - \boldsymbol{I})\log e. \end{aligned} \tag{117}$$

Therefore, invoking Theorem 1, we obtain the desired result

$$\begin{aligned} &\nabla_{\boldsymbol{H}} D(p_{\boldsymbol{y}} \,\|\, p_{\boldsymbol{y}\,|\,\boldsymbol{x}=\boldsymbol{x}_0}) \\ &\quad = \nabla_{\boldsymbol{H}} \mathbb{E}[(\boldsymbol{x}-\boldsymbol{x}_0)^\dagger \boldsymbol{H}^\dagger \boldsymbol{H}(\boldsymbol{x}-\boldsymbol{x}_0)] - \nabla_{\boldsymbol{H}} I(\boldsymbol{x};\boldsymbol{y}) \\ &\quad = \boldsymbol{H}(\mathbb{E}[(\boldsymbol{x}-\boldsymbol{x}_0)(\boldsymbol{x}-\boldsymbol{x}_0)^\dagger] - \boldsymbol{E}). \end{aligned} \tag{118}$$

The second part of the result follows from the definition of the MMSE matrix $\boldsymbol{E}$ in (12).

H. Proof of Derivatives for Random $\boldsymbol{H}$

For the case of the channel known at the receiver, note that

$$I(\boldsymbol{x};\boldsymbol{C}\boldsymbol{H}\boldsymbol{x}+\boldsymbol{n}\,|\,\boldsymbol{H}) = \mathbb{E}[I(\boldsymbol{x};\boldsymbol{C}\boldsymbol{H}\boldsymbol{x}+\boldsymbol{n}\,|\,\boldsymbol{H}=\boldsymbol{H})]. \tag{119}$$

To prove (30), it suffices to show that

$$\begin{aligned} &\nabla_{\boldsymbol{C}} \mathbb{E}[I(\boldsymbol{x};\boldsymbol{C}\boldsymbol{H}\boldsymbol{x}+\boldsymbol{n}\,|\,\boldsymbol{H}=\boldsymbol{H})] \\ &\qquad = \mathbb{E}[\nabla_{\boldsymbol{C}} I(\boldsymbol{x};\boldsymbol{C}\boldsymbol{H}\boldsymbol{x}+\boldsymbol{n}\,|\,\boldsymbol{H}=\boldsymbol{H})] \end{aligned} \tag{120}$$

and then use (29). The validity of the interchange of the order of the derivative and expectation can be shown by the Lebesgue Dominated Convergence Theorem: simply invoke Lemma 2 observing that

$$\begin{aligned} \left|\left[\boldsymbol{\Sigma}_n^{-1}\boldsymbol{C}\boldsymbol{H}\boldsymbol{E}\boldsymbol{H}^\dagger\right]_{ij}\right| &= |\boldsymbol{a}^\dagger \boldsymbol{H}\boldsymbol{E}\boldsymbol{H}^\dagger \boldsymbol{b}| \\ &\le \|\boldsymbol{a}\|\|\boldsymbol{b}\|\lambda_{\max}(\boldsymbol{E})\lambda_{\max}(\boldsymbol{H}\boldsymbol{H}^\dagger) \\ &\le \|\boldsymbol{a}\|\|\boldsymbol{b}\|\mathrm{Tr}(\boldsymbol{\Sigma}_x)\mathrm{Tr}(\boldsymbol{H}\boldsymbol{H}^\dagger) \end{aligned} \tag{121}$$

where $\boldsymbol{a}^\dagger = \boldsymbol{e}_i^\dagger \boldsymbol{\Sigma}_n^{-1}\boldsymbol{C}$ ($\boldsymbol{e}_i$ is the canonical ith vector, i.e., the all-zero vector except a one in the ith position), $\boldsymbol{b} = \boldsymbol{e}_j$, and we have used the Cauchy–Schwarz inequality in the first inequality and the fact that the maximum eigenvalue of a matrix is upper-bounded by the trace. Observe now that

$$\mathbb{E}[\mathrm{Tr}(\boldsymbol{\Sigma}_x)\mathrm{Tr}(\boldsymbol{H}\boldsymbol{H}^\dagger)] = \mathrm{Tr}(\boldsymbol{\Sigma}_x)\mathrm{Tr}(\mathbb{E}[\boldsymbol{H}\boldsymbol{H}^\dagger]) < \infty$$

due to the finite second-order moments of $\boldsymbol{H}$ and $\boldsymbol{x}$.

For the case of a channel unknown at the receiver, rewrite the signal model as $\boldsymbol{y} = \boldsymbol{C}\boldsymbol{z} + \boldsymbol{n}$, where $\boldsymbol{z} = \boldsymbol{H}\boldsymbol{x}$ is a random variable. Since $\boldsymbol{x} \to \boldsymbol{z} \to \boldsymbol{y}$ is a Markov chain, then $I(\boldsymbol{x};\boldsymbol{y}) = I(\boldsymbol{z};\boldsymbol{y}) - I(\boldsymbol{z};\boldsymbol{y}\,|\,\boldsymbol{x})$ [13] and using (29), we get (35) provided that

$$\nabla_{\boldsymbol{C}} \mathbb{E}[I(\boldsymbol{z};\boldsymbol{y}\,|\,\boldsymbol{x}=\boldsymbol{x})] = \mathbb{E}[\nabla_{\boldsymbol{C}} I(\boldsymbol{z};\boldsymbol{y}\,|\,\boldsymbol{x}=\boldsymbol{x})] \tag{122}$$

where the validity of the interchange of the order of the derivative and expectation follows again from the Lebesgue Dominated Convergence Theorem by invoking Lemma 2 and observing that

$$\begin{aligned} &\left|\left[\boldsymbol{\Sigma}_n^{-1}\boldsymbol{C}\mathbb{E}[(\boldsymbol{H}\boldsymbol{x}-\mathbb{E}[\boldsymbol{H}\boldsymbol{x}\,|\,\boldsymbol{y},\boldsymbol{x}])(\boldsymbol{H}\boldsymbol{x}-\mathbb{E}[\boldsymbol{H}\boldsymbol{x}\,|\,\boldsymbol{y},\boldsymbol{x}])^\dagger\,|\,\boldsymbol{x}]\right]_{ij}\right| \\ &\quad = |\boldsymbol{a}^\dagger \mathbb{E}[(\boldsymbol{H}\boldsymbol{x}-\mathbb{E}[\boldsymbol{H}\boldsymbol{x}\,|\,\boldsymbol{y},\boldsymbol{x}])(\boldsymbol{H}\boldsymbol{x}-\mathbb{E}[\boldsymbol{H}\boldsymbol{x}\,|\,\boldsymbol{y},\boldsymbol{x}])^\dagger\,|\,\boldsymbol{x}]\boldsymbol{b}| \\ &\quad \le \|\boldsymbol{a}\|\|\boldsymbol{b}\|\lambda_{\max}(\mathbb{E}[(\boldsymbol{H}\boldsymbol{x}-\mathbb{E}[\boldsymbol{H}\boldsymbol{x}\,|\,\boldsymbol{y},\boldsymbol{x}])(\boldsymbol{H}\boldsymbol{x} \\ &\qquad -\mathbb{E}[\boldsymbol{H}\boldsymbol{x}\,|\,\boldsymbol{y},\boldsymbol{x}])^\dagger\,|\,\boldsymbol{x}]) \\ &\quad \le \|\boldsymbol{a}\|\|\boldsymbol{b}\|\mathrm{Tr}(\mathbb{E}[(\boldsymbol{H}-\mathbb{E}[\boldsymbol{H}])\boldsymbol{x}\boldsymbol{x}^\dagger(\boldsymbol{H}-\mathbb{E}[\boldsymbol{H}])^\dagger]). \end{aligned} \tag{123}$$

Observe now that

$$\begin{aligned} &\mathbb{E}[\mathrm{Tr}(\mathbb{E}[(\boldsymbol{H}-\mathbb{E}[\boldsymbol{H}])\boldsymbol{x}\boldsymbol{x}^\dagger(\boldsymbol{H}-\mathbb{E}[\boldsymbol{H}])^\dagger])] \\ &\qquad = \mathrm{Tr}(\mathbb{E}[(\boldsymbol{H}-\mathbb{E}[\boldsymbol{H}])^\dagger(\boldsymbol{H}-\mathbb{E}[\boldsymbol{H}])\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\dagger]]) < \infty \end{aligned} \tag{124}$$

due to the finite second-order moments of $\boldsymbol{H}$ and $\boldsymbol{x}$.

REFERENCES

[1] K. H. Lee and D. P. Petersen, "Optimal linear coding for vector channels," *IEEE Trans. Commun.*, vol. COM-24, no. 12, pp. 1283–1290, Dec. 1976.

[2] A. J. Stam, "Some inequalities satisfied by the quantities of information of Fisher and Shannon," *Inf. Contr.*, vol. 2, pp. 101–112, Jun. 1959.

[3] T. E. Duncan, "On the calculation of mutual information," *SIAM J. Applied Math.*, vol. 19, pp. 215–220, Jul. 1970.

[4] T. T. Kadota, M. Zakai, and J. Ziv, "Mutual information of the white Gaussian channel with and without feedback," *IEEE Trans. Inf. Theory*, vol. IT-17, no. 4, pp. 368–371, Jul. 1971.

[5] J. Ziv and M. Zakai, "Some lower bounds on signal parameter estimation," *IEEE Trans. Inf. Theory*, vol. IT-15, no. 3, pp. 386–391, May 1969.

[6] D. Guo, S. Shamai (Shitz), and S. Verdú, "Mutual information and minimum mean-square error in Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1261–1282, Apr. 2005.

[7] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[8] G. Jöngren, M. Skoglund, and B. Ottersen, "Combining beamforming and orthogonal space-time block coding," *IEEE Trans. Inf. Theory*, vol. 48, no. 3, pp. 611–627, Mar. 2002.

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

[10] J. Salz, "Digital transmission over cross-coupled linear channels," *AT&T Tech. J.*, vol. 64, no. 6, pp. 1147–1159, Jul.–Aug. 1985.

[11] J. Yang and S. Roy, "On joint transmitter and receiver optimization for multiple-input-multiple-output (MIMO) transmission systems," *IEEE Trans. Commun.*, vol. 42, no. 12, pp. 3221–3231, Dec. 1994.

[12] A. Scaglione, G. B. Giannakis, and S. Barbarossa, "Redundant filterbank precoders and equalizers. Part I: Unification and optimal designs," *IEEE Trans. Signal Process.*, vol. 47, no. 7, pp. 1988–2006, Jul. 1999.

[13] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[14] V. V. Prelov and S. Verdú, "Second-order asymptotics of mutual information," *IEEE Trans. Inf. Theory*, vol. 50, no. 8, pp. 1567–1580, Aug. 2004.

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

[16] S. Verdú, "Spectral efficiency in the wideband regime," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1319–1343, Jun. 2002.

[17] N. M. Blachman, "The convolution inequality for entropy powers," *IEEE Trans. Inf. Theory*, vol. IT-11, no. 2, pp. 267–211, Apr. 1965.

[18] M. H. M. Costa, "A new entropy power inequality," *IEEE Trans. Inf. Theory*, vol. IT-31, no. 6, pp. 751–760, Nov. 1985.

[19] A. Dembo, T. M. Cover, and J. A. Thomas, "Information theoretic inequalities," *IEEE Trans. Inf. Theory*, vol. 37, no. 6, pp. 1501–1518, Nov. 1991.

[20] A. Dembo, "Information Inequalities and Uncertainty Principles ," Dept. Statistics, Stanford Univ., Stanford, CA, Tech. Rep. 75, Jul. 1990.

[21] O. Johnson and Y. Suhov, "Entropy and random vectors," *J. Statist. Phys.*, vol. 104, no. 1–2, pp. 145–165, Jul. 2001.

[22] W. Rudin, *Principles of Mathematical Analysis*, 3rd ed. New York: McGraw-Hill, 1976.

[23] H. V. Poor, *An Introduction to Signal Detection and Estimation*, 2nd ed. New York: Springer-Verlag, 1994.

[24] M. S. Pinsker, V. V. Prelov, and S. Verdú, "Sensitivity of channel capacity," *IEEE Trans. Inf. Theory*, vol. 41, no. 6, pp. 1877–1888, Nov. 1995.

[25] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*. Englewood Cliffs, NJ: Prentice-Hall, 1989.

[26] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed. Belmont, MA: Athena Scientific, 1999.

[27] S. Haykin, *Adaptive Filter Theory*, 4th ed. Upper Saddle River, NJ: Prentice-Hall, 2002.

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

[29] D. H. Brandwood, "A compex gradient operator and its application in adaptive array theory," *Proc. Inst. Elec. Eng.*, vol. 130, no. 1, pp. 11–16, Feb. 1983.

[30] P. Billingsley, *Probability and Measure*, 3rd ed. New York: Wiley, 1995.