

Gradient of Mutual Information in Linear Vector Gaussian Channels

Daniel P. Palomar and Sergio Verdú

Dept. of Electrical Engineering
Princeton University
Engineering Quadrangle, Princeton, NJ 08544, USA
Email: {danielp,verdu}@princeton.edu

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 [1], 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 estimate of the input given the output.

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 [2]. Another interesting example arises in the concatenation of

subsystems, where various combinations of subsystems can be analyzed in terms of robustness to choose the most appropriate combination.

Recently, a fundamental relation between the mutual information and the minimum mean-square error (MMSE) was unveiled in [1] for the scalar Gaussian channel $y = \sqrt{\text{snr}} x + n$ (complex-valued inputs/outputs are considered throughout this paper) regardless of the input distribution:

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

where mutual information is in nats and $\text{mmse}(\text{snr})$ 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., $\text{mmse}(\text{snr}) = \mathbb{E} \left[|x - \mathbb{E}[x | \sqrt{\text{snr}} x + n]|^2 \right]$. An extension of (1) to the vector case was also given in [1] for the linear vector Gaussian channel $\mathbf{y} = \sqrt{\text{snr}} \mathbf{H}\mathbf{x} + \mathbf{n}$ as:

$$\begin{aligned} \frac{d}{d\text{snr}} I(\mathbf{x}; \sqrt{\text{snr}} \mathbf{H}\mathbf{x} + \mathbf{n}) \\ = \mathbb{E} \left[\left\| \mathbf{H}\mathbf{x} - \mathbb{E}[\mathbf{H}\mathbf{x} | \sqrt{\text{snr}} \mathbf{H}\mathbf{x} + \mathbf{n}] \right\|^2 \right] \end{aligned} \quad (2)$$

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

The derivative of the mutual information with respect to the 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 analyze the sensitivity of mutual information for the linear vector channel with independent Gaussian noise $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}$ is via the gradient with respect to the deterministic matrix \mathbf{H} . The main result of this paper is the following formula for the gradient:

$$\nabla_{\mathbf{H}} I(\mathbf{x}; \mathbf{H}\mathbf{x} + \mathbf{n}) = \mathbf{H}\mathbf{E} \quad (3)$$

where \mathbf{E} is the covariance matrix of the estimation error vector, sometimes referred to as MMSE matrix [3]. The MMSE matrix is the full generalization of the scalar MMSE in (1)

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.

to the vector case. As we will see, applying the calculus chain rule to the basic gradient in (3), 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.

In addition to its intrinsic theoretical interest, the connection between the gradient of the mutual information and the MMSE matrix opens up new computational possibilities. For example, if \mathbf{x} is equally likely to take values on a given codebook, it is notoriously difficult to compute the mutual information. However, for a sparse-graph linear code, the MMSE matrix can be computed with the aid of the belief propagation algorithm that approximates very accurately the posterior marginals and, hence, the MMSE estimates which are given by the expected values of the posterior marginals.

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:

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n} \quad (4)$$

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

The general channel model in (4) describes many different communication systems, e.g., wireless multi-antenna systems, CDMA systems, wireline digital subscriber line systems, or even single-antenna frequency-selective wideband channels.

It is interesting to further generalize the model in (4) to include an additional linear transformation represented by the $n_T \times L$ matrix \mathbf{B} , where L is now the dimension of the transmitted vector \mathbf{x} :

$$\mathbf{y} = \mathbf{H}\mathbf{B}\mathbf{x} + \mathbf{n}. \quad (5)$$

The additional matrix \mathbf{B} can play different roles: i) in a wireless multi-antenna system it may represent a *beamforming matrix* that uses some knowledge about the physical channel \mathbf{H} to properly steer the transmitted signal through the best channel eigenmodes [4], [5]; ii) \mathbf{B} can denote a *linear precoding matrix* or *shaping matrix* that adapts or shapes the transmitted signal to the channel realization [2], [5]; iii) the overall input-output linear transformation may factor into two matrices \mathbf{H} and \mathbf{B} (one of which may be controllable by the designer of the system).

A special case of (5) by setting $\mathbf{B} = \sqrt{\text{snr}} \mathbf{I}$ yields the model used in (2):

$$\mathbf{y} = \sqrt{\text{snr}} \mathbf{H}\mathbf{x} + \mathbf{n}. \quad (6)$$

Consider the estimation of the input signal \mathbf{x} based on the observation of the output \mathbf{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 $\mathbb{E} \left[|x - \hat{x}(y)|^2 \right]$, 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 minimum MSE (MMSE) and is termed *minimum MSE (MMSE) estimator*. In the more general vector setup, the MMSE estimator $\hat{\mathbf{x}}(\mathbf{y})$ achieves simultaneously the minimum MSE at each component of the estimation error vector and is again given by the conditional mean estimator:

$$\hat{\mathbf{x}}(\mathbf{y}) = \mathbb{E} [\mathbf{x} | \mathbf{y}]. \quad (7)$$

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:

$$\mathbf{E} \triangleq \mathbb{E} \left[(\mathbf{x} - \mathbb{E} [\mathbf{x} | \mathbf{y}]) (\mathbf{x} - \mathbb{E} [\mathbf{x} | \mathbf{y}])^\dagger \right]. \quad (8)$$

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

III. MAIN RESULT

A. Basic Result: Gradient with Respect to \mathbf{H}

To start with, consider the signal model in (4) with *Gaussian signaling*, where the covariance matrix of the transmitted signal \mathbf{x} is denoted by Σ_x and the covariance matrix of the noise is the identity matrix. The mutual information, $I(\mathbf{x}; \mathbf{y})$, is¹ (e.g., [6])

$$I = \log \det \left(\mathbf{I} + \mathbf{H}\Sigma_x\mathbf{H}^\dagger \right). \quad (9)$$

The MMSE estimator is

$$\hat{\mathbf{x}} = \Sigma_x\mathbf{H}^\dagger \left(\mathbf{I} + \mathbf{H}\Sigma_x\mathbf{H}^\dagger \right)^{-1} \mathbf{y} \quad (10)$$

(assuming \mathbf{x} and \mathbf{n} with zero mean), which happens to be linear, and the MMSE matrix is [3, Thm. 11.1]

$$\mathbf{E} = \left(\Sigma_x^{-1} + \mathbf{H}^\dagger\mathbf{H} \right)^{-1}. \quad (11)$$

Taking the gradient of the expression in (9)² we obtain

$$\begin{aligned} \nabla_{\mathbf{H}} I &= \nabla_{\mathbf{H}} \log \det \left(\mathbf{I} + \mathbf{H}^\dagger\mathbf{H}\Sigma_x \right) \\ &= \mathbf{H}\Sigma_x \left(\mathbf{I} + \mathbf{H}^\dagger\mathbf{H}\Sigma_x \right)^{-1} \\ &= \mathbf{H}\mathbf{E}. \end{aligned} \quad (12)$$

The main result of the paper shows that the Gaussian assumption on the input is unnecessary for (3) to hold.

Theorem 1: Consider the signal model in (4), where \mathbf{H} is an arbitrary deterministic matrix, the signaling \mathbf{x} is arbitrarily distributed (with finite second-order moments), and the noise \mathbf{n} is Gaussian, independent of the input \mathbf{x} , with normalized

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

²We use the well-known definition of the complex derivative of a real-valued scalar function f : $\frac{df}{dx^*} \triangleq \frac{1}{2} \left[\frac{\partial f}{\partial \text{Re}\{x\}} + j \frac{\partial f}{\partial \text{Im}\{x\}} \right]$ [7]. For the sake of notation, we define the complex gradient matrix as $\nabla_{\mathbf{X}} f \triangleq \partial f / \partial \mathbf{X}^*$, where $[\partial f / \partial \mathbf{X}^*]_{ij} = \partial f / \partial [\mathbf{X}^*]_{ij}$.

covariance matrix $\Sigma_n = \mathbf{I}$. Then, the mutual information I and the MMSE matrix \mathbf{E} satisfy:

$$\nabla_{\mathbf{H}} I(\mathbf{x}; \mathbf{H}\mathbf{x} + \mathbf{n}) = \mathbf{H}\mathbf{E}. \quad (13)$$

Proof: See Appendix. \blacksquare

B. Gradient with Respect to Arbitrary Parameters

From the basic gradient in Theorem 1 and the gradient chain rule it is possible to obtain gradients with respect to other parameters of the system.

Theorem 2: Consider the general signal model in (5) including a linear precoder \mathbf{B} at the transmitter, where \mathbf{H} is an arbitrary deterministic matrix, the signaling \mathbf{x} is arbitrarily distributed with covariance matrix Σ_x , the noise \mathbf{n} is Gaussian, independent of the input \mathbf{x} , and has positive definite covariance matrix Σ_n , the transmit covariance matrix is $\mathbf{Q} = \mathbf{B}\Sigma_x\mathbf{B}^\dagger$ (which includes the precoding as well), and the squared linear precoder is $\mathbf{Q}_B = \mathbf{B}\mathbf{B}^\dagger$. Then, the mutual information I and the MMSE matrix \mathbf{E} satisfy:

$$\nabla_{\mathbf{H}} I(\mathbf{x}; \mathbf{H}\mathbf{B}\mathbf{x} + \mathbf{n}) = \Sigma_n^{-1} \mathbf{H}\mathbf{B}\mathbf{E}\mathbf{B}^\dagger, \quad (14)$$

$$\nabla_{\mathbf{B}} I(\mathbf{x}; \mathbf{H}\mathbf{B}\mathbf{x} + \mathbf{n}) = \mathbf{H}^\dagger \Sigma_n^{-1} \mathbf{H}\mathbf{B}\mathbf{E}, \quad (15)$$

$$\nabla_{\mathbf{Q}} I(\mathbf{x}; \mathbf{H}\mathbf{B}\mathbf{x} + \mathbf{n}) \mathbf{B}\Sigma_x = \mathbf{H}^\dagger \Sigma_n^{-1} \mathbf{H}\mathbf{B}\mathbf{E}, \quad (16)$$

$$\nabla_{\mathbf{Q}_B} I(\mathbf{x}; \mathbf{H}\mathbf{B}\mathbf{x} + \mathbf{n}) \mathbf{B} = \mathbf{H}^\dagger \Sigma_n^{-1} \mathbf{H}\mathbf{B}\mathbf{E}, \quad (17)$$

$$\nabla_{\Sigma_x} I(\mathbf{x}; \mathbf{H}\mathbf{B}\mathbf{x} + \mathbf{n}) \Sigma_x = \mathbf{B}^\dagger \mathbf{H}^\dagger \Sigma_n^{-1} \mathbf{H}\mathbf{B}\mathbf{E}, \quad (18)$$

$$\nabla_{\Sigma_n^{-1}} I(\mathbf{x}; \mathbf{H}\mathbf{B}\mathbf{x} + \mathbf{n}) = \mathbf{H}\mathbf{B}\mathbf{E}\mathbf{B}^\dagger \mathbf{H}^\dagger, \quad (19)$$

$$\nabla_{\Sigma_n} I(\mathbf{x}; \mathbf{H}\mathbf{B}\mathbf{x} + \mathbf{n}) = -\Sigma_n^{-1} \mathbf{H}\mathbf{B}\mathbf{E}\mathbf{B}^\dagger \mathbf{H}^\dagger \Sigma_n^{-1}. \quad (20)$$

C. First-Order Approximation of Mutual Information

Using the gradients of the mutual information, we can write first-order approximations of the mutual information as a function of different parameters.

Consider the signal model $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}$ (with $\Sigma_n = \mathbf{I}$). The first-order expansion of the mutual information as a function of the transmit (Hermitian) covariance matrix Σ_x , denoted by $I(\Sigma_x)$, is obtained from the gradient $\nabla_{\Sigma_x} I = \mathbf{H}^\dagger \mathbf{H}\mathbf{E}\Sigma_x^{-1}$ (Theorem 2) as

$$I(\Sigma_{x,0} + \Delta) = I(\Sigma_{x,0}) + \text{Tr} \left(\Sigma_{x,0}^{-1} \mathbf{E}\mathbf{H}^\dagger \mathbf{H}\Delta \right) + o(\|\Delta\|). \quad (21)$$

This expansion can be particularized around $\Sigma_{x,0} = \mathbf{0}$ (low-power regime) as

$$I(\Sigma_x) = \text{Tr}(\mathbf{H}\Sigma_x\mathbf{H}^\dagger) + o(\|\Sigma_x\|). \quad (22)$$

IV. SOME CONNECTIONS AND APPLICATIONS

This section contains some particularizations and theoretical connections derived from Theorem 1. For more theoretical results and practical applications the reader is referred to [8].

A. Particularizations to SNR-Gradients

The fundamental relation between the mutual information and the MMSE was thoroughly explored in [1] for the scalar Gaussian channel $y = \sqrt{\text{snr}}x + n$ and, to some extent, also for the vector Gaussian channel $\mathbf{y} = \sqrt{\text{snr}}\mathbf{x} + \mathbf{n}$ 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 [1].

Corollary 1: Consider the signal model in (6), where all the terms are defined as in Theorem 2. Then,

$$\frac{dI}{d\text{snr}} = \text{Tr}(\mathbf{H}^\dagger \Sigma_n^{-1} \mathbf{H}\mathbf{E}) \quad (23)$$

$$= \mathbb{E} \left[\left\| \Sigma_n^{-1/2} \mathbf{H}\mathbf{x} - \mathbb{E} \left[\Sigma_n^{-1/2} \mathbf{H}\mathbf{x} \mid \mathbf{y} \right] \right\|^2 \right]. \quad (24)$$

Corollary 2: Consider the following signal model:

$$\mathbf{y} = \mathbf{H}\mathbf{\Gamma}\mathbf{x} + \mathbf{n} \quad (25)$$

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

$$\nabla_{\mathbf{\Gamma}\mathbf{\Gamma}^\dagger} I = \mathbf{H}^\dagger \Sigma_n^{-1} \mathbf{H}\mathbf{\Gamma}\mathbf{E}\mathbf{\Gamma}^{-1}. \quad (26)$$

Observe that, when (25) models a multiuser system with the symbols transmitted by all K users stacked in \mathbf{x} and $\mathbf{\Gamma} = \text{diag}(\{\sqrt{\text{snr}_k}\}_{k=1}^K)$, then (26) is particularly useful since $\partial I / \partial \text{snr}_k = [\nabla_{\mathbf{\Gamma}\mathbf{\Gamma}^\dagger} I]_{kk}$.

B. Matrix Generalization of De Bruijn's Identity

For a multivariate density function $p_{\mathbf{y}}(\mathbf{y})$, De Bruijn's identity [9], [10], [11], [12] relates the derivative of the differential entropy $h(\cdot)$ with the Fisher information matrix defined as

$$\mathbf{J}(\mathbf{y}) \triangleq \mathbb{E}_{\mathbf{y}} \left[\nabla_{\mathbf{y}} \log p_{\mathbf{y}}(\mathbf{y}) \nabla_{\mathbf{y}}^\dagger \log p_{\mathbf{y}}(\mathbf{y}) \right]. \quad (27)$$

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 is usually the case [12].³ The scalar version is similarly defined as

$$J(\mathbf{y}) \triangleq \text{Tr}(\mathbf{J}(\mathbf{y})). \quad (28)$$

The original De Bruijn's identity was obtained for the scalar case [9], [10] as

$$\frac{d}{dt} h(z + \sqrt{tn}) = J(z + \sqrt{tn}) \quad (29)$$

where the noise n has a normalized variance $\sigma_n^2 = 1$. The scalar version of the multivariate De Bruijn's identity (assuming a normalized covariance matrix for the noise $\Sigma_n = \mathbf{I}$) is [11][12, Thm. 14][1, eq. (51)]

$$\frac{d}{dt} h(\mathbf{z} + \sqrt{t}\mathbf{n}) = J(\mathbf{z} + \sqrt{t}\mathbf{n}). \quad (30)$$

A more general multivariate De Bruijn's identity is

$$\frac{d}{dt} h(\mathbf{z} + \sqrt{t}\mathbf{A}\mathbf{n}) = \text{Tr}(\mathbf{A}^\dagger \mathbf{J}(\mathbf{z} + \sqrt{t}\mathbf{A}\mathbf{n}) \mathbf{A}) \quad (31)$$

where again the noise has a normalized noise covariance matrix and \mathbf{A} is an arbitrary square invertible matrix. Note that

³The Fisher information matrix with respect to a parameter θ is defined as $\mathbf{J}(\mathbf{y}) \triangleq \mathbb{E}_{\mathbf{y}} \left[\nabla_{\theta} \log p_{\mathbf{y}}(\mathbf{y}; \theta) \nabla_{\theta}^\dagger \log p_{\mathbf{y}}(\mathbf{y}; \theta) \right]$ [3].

in the context of the signal model in (4) we can set $\mathbf{z} = \mathbf{H}\mathbf{x}$ to particularize the results.

As can be observed from (29)-(31), the existing versions of De Bruijn's identity take a derivative with respect to a scalar parameter which takes the role of 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 $\mathbf{T}^{1/2}\mathbf{n}$, where the linear transformation $\mathbf{T}^{1/2}$ plays the role of \sqrt{t} in the scalar version of the identity.

Theorem 3: [Matrix version of the multivariate De Bruijn's identity] Consider an arbitrary random variable (with finite second-order moments) \mathbf{z} contaminated with a Gaussian noise \mathbf{n} independent of \mathbf{z} and with positive definite covariance matrix Σ_n . Then,

$$\nabla_{\Sigma_n} h(\mathbf{z} + \mathbf{n}) = \mathbf{J}(\mathbf{z} + \mathbf{n}). \quad (32)$$

If \mathbf{z} is contaminated with a linearly transformed version of the Gaussian noise $\mathbf{T}^{1/2}\mathbf{n}$, where $\mathbf{T}^{1/2}$ is an arbitrary square invertible matrix, then,

$$\nabla_{\mathbf{T}} h(\mathbf{z} + \mathbf{T}^{1/2}\mathbf{n}) = \mathbf{J}(\mathbf{z} + \mathbf{T}^{1/2}\mathbf{n}) \times \mathbf{T}^{1/2}\Sigma_n\mathbf{T}^{-1/2}. \quad (33)$$

Observe that when the noise has a normalized covariance matrix $\Sigma_n = \mathbf{I}$, (33) simplifies to

$$\nabla_{\mathbf{T}} h(\mathbf{z} + \mathbf{T}^{1/2}\mathbf{n}) = \mathbf{J}(\mathbf{z} + \mathbf{T}^{1/2}\mathbf{n}) \quad (34)$$

which is the natural generalization of the scalar identity (29).

The scalar expressions of the multivariate De Bruijn's identity in (29)-(31) can be obtained from a simple application of the chain rule to Theorem 3.

APPENDIX PROOF OF THEOREM 1

Observe that the mean 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 the sake of notation, we will assume the noise to be zero mean without loss of generality. Note that the interchange of the order of integrals and derivatives can be proved by the Lebesgue Dominated Convergence Theorem. In the scalar case, this proof reduces to the proof in [1].

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

$$p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}) = \frac{1}{\pi^{n_R}} \exp\left(-\|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2\right)$$

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

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

Then,

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

where we have used

$$\frac{\partial p_{\mathbf{y}|\mathbf{x}}(\mathbf{y})}{\partial \mathbf{H}^*} = -\nabla_{\mathbf{y}} p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}) \mathbf{x}^\dagger$$

and the last equality follows from integrating by parts and noting that $p_{\mathbf{y}|\mathbf{x}}(\mathbf{y})(1 + \log p_{\mathbf{y}}(\mathbf{y})) \rightarrow 0$ as $\|\mathbf{y}\| \rightarrow \infty$.

Now, focusing on the (i, j) th element,

$$\begin{aligned} \left[\frac{\partial I}{\partial \mathbf{H}^*} \right]_{ij} &= \mathbb{E}_{\mathbf{x}} \left[\mathbf{e}_i^\dagger \left(- \int \frac{p_{\mathbf{y}|\mathbf{x}}(\mathbf{y})}{p_{\mathbf{y}}(\mathbf{y})} \nabla_{\mathbf{y}} p_{\mathbf{y}}(\mathbf{y}) d\mathbf{y} \right) \mathbf{x}^\dagger \mathbf{e}_j \right] \\ &= - \int \mathbb{E} \left[\mathbf{x}^\dagger \mathbf{e}_j \mathbf{e}_i^\dagger | \mathbf{y} \right] \nabla_{\mathbf{y}} p_{\mathbf{y}}(\mathbf{y}) d\mathbf{y} \\ &= \mathbb{E} \left[\mathbb{E} \left[\mathbf{x}^\dagger \mathbf{e}_j \mathbf{e}_i^\dagger | \mathbf{y} \right] (\mathbf{y} - \mathbf{H}\mathbb{E}[\mathbf{x} | \mathbf{y}]) \right] \\ &= \mathbb{E} \left[\mathbf{x}^\dagger \mathbf{e}_j \mathbf{e}_i^\dagger \mathbf{y} \right] - \mathbb{E} \left[\mathbb{E}[\mathbf{x}^\dagger | \mathbf{y}] \mathbf{e}_j \mathbf{e}_i^\dagger \mathbf{H}\mathbb{E}[\mathbf{x} | \mathbf{y}] \right] \end{aligned}$$

where we have used $\nabla_{\mathbf{y}} p_{\mathbf{y}}(\mathbf{y}) = -p_{\mathbf{y}}(\mathbf{y})(\mathbf{y} - \mathbf{H}\mathbb{E}[\mathbf{x} | \mathbf{y}])$. More compactly,

$$\frac{\partial I}{\partial \mathbf{H}^*} = \mathbf{H} \left(\mathbb{E}[\mathbf{x}\mathbf{x}^\dagger] - \mathbb{E}[\mathbb{E}[\mathbf{x} | \mathbf{y}] \mathbb{E}[\mathbf{x}^\dagger | \mathbf{y}]] \right).$$

■

REFERENCES

- [1] D. Guo, S. Shamai, and S. Verdú, "Mutual information and minimum mean-square error in Gaussian channels," *IEEE Trans. Inform. Theory*, vol. 51, no. 4, pp. 1261–1282, April 2005.
- [2] 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.
- [3] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ, USA: Prentice Hall, 1993.
- [4] G. Jöngren, M. Skoglund, and B. Ottersen, "Combining beamforming and orthogonal space-time block coding," *IEEE Trans. Inform. Theory*, vol. 48, no. 3, pp. 611–627, March 2002.
- [5] D. P. Palomar, J. M. Cioffi, and M. A. Lagunas, "Joint Tx-Rx beamforming design for multicarrier MIMO channels: A unified framework for convex optimization," *IEEE Trans. Signal Processing*, vol. 51, no. 9, pp. 2381–2401, Sept. 2003.
- [6] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley, 1991.
- [7] D. H. Brandwood, "A complex gradient operator and its application in adaptive array theory," *IEE Proc.*, vol. 130, no. 1, pp. 11–16, Feb. 1983.
- [8] D. P. Palomar and S. Verdú, "Gradient of mutual information in linear vector Gaussian channels," *IEEE Trans. Inform. Theory*, submitted 2005.
- [9] A. J. Stam, "Some inequalities satisfied by the quantities of information of Fisher and Shannon," *Inform. Contr.*, vol. 2, pp. 101–112, June 1959.
- [10] N. M. Blachman, "The convolution inequality for entropy powers," *IEEE Trans. Inform. Theory*, vol. IT-11, pp. 267–271, April 1965.
- [11] M. H. M. Costa, "A new entropy power inequality," *IEEE Trans. Inform. Theory*, vol. IT-31, no. 6, pp. 751–760, Nov. 1985.
- [12] A. Dembo, T. M. Cover, and J. A. Thomas, "Information theoretic inequalities," *IEEE Trans. Inform. Theory*, vol. 37, no. 6, pp. 1501–1518, Nov. 1991.