Yet another entropy power inequality with an application

Miquel Payaró, Maria Gregori
Centre Tecnològic de Telecomunicacions de Catalunya (CTTC),
Parc Mediterrani de la Tecnologia,
Castelldefels, Barcelona, Spain.
{miquel.payaro,maria.gregori}@cttc.es

Daniel Palomar
Department of Electronic and Computer Engineering,
Hong Kong University of Science and Technology,
Clear Water Bay, Kowloon, Hong Kong.
palomar@ust.hk

***Abstract*—In this paper we derive a generalization of the vector entropy power inequality (EPI) recently put forth in [1], which was valid only for diagonal matrices, to the full matrix case. Next, we study the problem of computing the linear precoder that maximizes the mutual information in linear vector Gaussian channels with arbitrary inputs. In particular, we transform the precoder optimization problem into a new form and, capitalizing on the newly unveiled matrix EPI, we show that some particular instances of the optimization problem can be cast in convex form, i.e., we can have an optimality certificate, which, to the best of our knowledge, had never been obtained previously.**

***Index Terms*—Entropy power inequality, linear vector Gaussian channels, precoder optimization, arbitrary inputs.**

I. Introduction

Given a random vector $\boldsymbol{Y} \in \mathbb{R}^n$, Shannon introduced the entropy power function, $N(\boldsymbol{Y})$, in [2] as:

$$N(\boldsymbol{Y}) = \frac{1}{2\pi\mathrm{e}} \exp\left(\frac{2}{n} h(\boldsymbol{Y})\right), \tag{1}$$

where $h(\boldsymbol{Y})$ denotes the differential entropy, which, for continuous random vectors reads as $h(\boldsymbol{Y}) = -\mathsf{E}\left\{\log P_{\boldsymbol{Y}}(\boldsymbol{Y})\right\}$. The previous definition is usually extended with $h(\boldsymbol{Y}) = -\infty$ for the case where the distribution of $\boldsymbol{Y}$ assigns positive mass to one or more singletons in $\mathbb{R}^n$.

Given a standard Gaussian random vector denoted by $\boldsymbol{Y}_G \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \mathbf{I}_n\right)$ such that its differential entropy is identical to that of $\boldsymbol{Y}$, $h(\boldsymbol{Y}) = h(\boldsymbol{Y}_G)$, then the entropy power $N(\boldsymbol{Y})$ is equal to the mean power (or variance) of the components of vector $\boldsymbol{Y}_G$, i.e., $N(\boldsymbol{Y}) = \sigma^2$.

A. Shannon's entropy power inequality (EPI)

For any two independent arbitrary random vectors $\boldsymbol{X} \in \mathbb{R}^n$ and $\boldsymbol{W} \in \mathbb{R}^n$, Shannon gave in [2] the following inequality:

$$N(\boldsymbol{X} + \boldsymbol{W}) \geq N(\boldsymbol{X}) + N(\boldsymbol{W}).$$

The first rigorous proof of Shannon's EPI was given in [3] by Stam, and was simplified by Blachman in [4]. A simple and very elegant proof by Verdú and Guo based on estimation theoretic considerations has recently appeared in [5].

Among many other important results, Bergmans' proof of the converse for the degraded Gaussian broadcast channel [6] and Oohama's partial solution to the rate distortion region problem for Gaussian multiterminal source coding systems [7] follow from Shannon's EPI.

This work was partially funded by the Catalan and Spanish governments under grants 2009SGR 1046 and TEC2008-06327-C03-03/TEC, respectively.

B. Costa's EPI

Under the setting of Shannon's EPI, Costa proved in [8] that, provided that the random vector $\boldsymbol{W}$ is white Gaussian distributed, then Shannon's EPI can be strengthened to

$$N(\boldsymbol{X} + \sqrt{t}\boldsymbol{W}) \geq (1-t)N(\boldsymbol{X}) + tN(\boldsymbol{X} + \boldsymbol{W}), \tag{2}$$

where $t \in [0, 1]$. As Costa noted in [8], the above EPI is equivalent to the concavity of the entropy power function $N(\boldsymbol{X} + \sqrt{t}\boldsymbol{W})$ with respect to the parameter t, which, due to the smoothing caused by the added noise the function, can be formally expressed as the second order condition

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} N\big(\boldsymbol{X} + \sqrt{t}\boldsymbol{W}\big) \leq 0. \tag{3}$$

Due to its inherent interest and to the fact that the proof by Costa was rather involved, simplified proofs of his result have been subsequently given in [9]–[11].

Additionally, in his paper Costa presented two extensions of his main result in (3). Precisely, he showed that the EPI is also valid when the Gaussian vector $\boldsymbol{W}$ is not white, and also for the case where the t variable is multiplying the arbitrarily distributed random vector $\boldsymbol{X}$,

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} N\big(\sqrt{t}\boldsymbol{X} + \boldsymbol{W}\big) \leq 0. \tag{4}$$

Similarly to Shannon's EPI, Costa's EPI has been used successfully to derive important information-theoretic results concerning, e.g., Gaussian interference channels in [12] or multi-antenna flat fading channels with memory in [13].

C. Generalizations of Costa's EPI

Much more recently, the two forms in (2) and (4) of Costa's EPI have been generalized to the multivariate case in [14] and [1], respectively.

Precisely, in [14], the authors proved the multivariate EPI:

$$N(\boldsymbol{X} + \mathbf{A}^{1/2}\boldsymbol{W}) \geq |\mathbf{I}_n - \mathbf{A}|^{1/n} N(\boldsymbol{X}) + |\mathbf{A}|^{1/n} N(\boldsymbol{X} + \boldsymbol{W}), \tag{5}$$

which is valid under the same setting as in the previous section and with the additional contraint that $\mathbf{0} \leq \mathbf{A} \leq \mathbf{I}_n$. The authors of [14] used (5) to establish the secrecy capacity regions of the degraded vector Gaussian broadcast channel with layered confidential messages.

The multivariate counterpart of (4) was given in [1] as

$$\mathsf{H}_{\boldsymbol{\mu}} N(\mathbf{Diag}(\boldsymbol{\mu})^{1/2}\boldsymbol{X} + \boldsymbol{W}) \leq \mathbf{0}, \tag{6}$$

where H represents the Hessian operator, $\boldsymbol{\mu} \in \mathbb{R}^n$ is an n-dimensional vector with non-negative entries, and $\mathbf{Diag}(\boldsymbol{\mu})$ represents a diagonal matrix, whose non-zero entries are given by the elements of the vector $\boldsymbol{\mu}$. Observe that (6) implies that the entropy power function $N(\mathbf{Diag}(\boldsymbol{\mu})^{1/2}\boldsymbol{X} + \boldsymbol{W})$ is concave with respect to $\boldsymbol{\mu}$. The multivariate EPI in (6) was used in [15] to derive outer bounds on the capacity region in multiuser channels with feedback and in [16] to compute the optimal singular values of the linear precoder that maximizes the mutual information in linear vector Gaussian channels with arbitrary inputs. In the following subsection, we present a brief state-of-the-art regarding the design of the linear precoder that maximizes the mutual information in linear vector Gaussian channels because this problem will be further studied in this paper.

D. Precoder design for linear vector Gaussian channels

It has been long well known from [17], [18] that, in linear vector Gaussian channels with an average power constraint, capacity is achieved by zero-mean Gaussian inputs, whose covariance is aligned with the channel eigenmodes and where the power is distributed among the covariance eigenvalues according to the waterfilling policy. Despite the information theoretic optimality of Gaussian inputs, they are seldom used in practice due to their implementation complexity. Rather, system designers often resort to simple discrete constellations, such as BPSK or QAM.

In this context, the scalar relationship between mutual information and minimum mean square error (MMSE) for linear vector Gaussian channels put forth recently in [19], and extended to the vector case in [20], has become a fundamental tool in transmitter design beyond the Gaussian signaling case.

In [21], the authors derived the optimum diagonal precoder, or power allocation, in quasi-closed form, coining the term mercury/waterfilling. Their results were found for the particular case of a diagonal channel corrupted with AWGN and imposing independence on the components of the input vector.

The linear transmitter design (or linear precoding) problem was recently studied in [22] with a wider scope by considering full (non-diagonal) precoder and channel matrices and arbitrary inputs with possibly dependent components. In their work, the authors gave necessary conditions for the optimal precoder and optimal transmit covariance matrix and proposed numerical iterative methods to compute a (in general suboptimal) solution.

More recently, in [23], an alternative approach was taken where, instead of taking the usual strategy of optimizing the transmitted signal covariance matrix, a different covariance matrix, which was also related to the precoder, was optimized. Despite the fact that the method proposed in [23] also provided a suboptimal solution, the author showed that, in general, her method yielded better convergence properties than the strategy proposed in [22].

In the light of the fact that, despite all the research efforts mentioned above, a general solution for the design of the precoder that maximizes the mutual information with arbitrary inputs is still missing, the authors of [16] assessed the computational complexity of this problem and argued that it was likely to be NP-hard. Precisely, the authors showed that the dependence of the mutual information on the right singular vector matrix of the precoder is a key element in the intractability of computing the precoder that maximizes the mutual information. In this paper, we will show that some particular instances of the precoder design problem can be cast in convex form and, thus, efficiently solved.[1]

II. System model and definitions

Throughout this paper, we consider that the random vector $\boldsymbol{Y}$ is the sum of a linearly-transformed arbitrarily distributed random vector $\boldsymbol{S} \in \mathbb{R}^m$ and a Gaussian distributed random vector $\boldsymbol{W}$ whose covariance matrix is the identity matrix, i.e.,

$$\boldsymbol{Y} = \mathbf{G}\boldsymbol{S} + \boldsymbol{W}, \tag{7}$$

where $\mathbf{G} \in \mathbb{R}^{n \times m}$ is the matrix with the coefficients of the linear transformation undergone by the random vector $\boldsymbol{S}$.

We consider the estimation of the input signal $\boldsymbol{S}$ based on the observation of a realization of the output $\boldsymbol{Y} = \boldsymbol{y}$. The estimator that simultaneously achieves the minimum mean square error (MSE) for all the components of the estimation error vector is given by the conditional mean estimator $\widehat{\boldsymbol{S}}(\boldsymbol{y}) = \mathsf{E}\{\boldsymbol{S} \mid \boldsymbol{y}\}$ and the corresponding MSE matrix, referred to as the MMSE matrix, is

$$\mathbf{E}_{\boldsymbol{S}} = \mathsf{E}\big\{(\boldsymbol{S} - \mathsf{E}\{\boldsymbol{S} \mid \boldsymbol{Y}\})(\boldsymbol{S} - \mathsf{E}\{\boldsymbol{S} \mid \boldsymbol{Y}\})^{\mathsf{T}}\big\}. \tag{8}$$

An alternative and useful expression for the MMSE matrix can be obtained by considering first the MMSE matrix conditioned on a specific realization of the output $\boldsymbol{Y} = \boldsymbol{y}$, which is denoted by $\boldsymbol{\Phi}_{\boldsymbol{S}}(\boldsymbol{y})$ and defined as:

$$\boldsymbol{\Phi}_{\boldsymbol{S}}(\boldsymbol{y}) = \mathsf{E}\big\{(\boldsymbol{S} - \mathsf{E}\{\boldsymbol{S}|\boldsymbol{y}\})(\boldsymbol{S} - \mathsf{E}\{\boldsymbol{S}|\boldsymbol{y}\})^{\mathsf{T}} \bigm| \boldsymbol{y}\big\}. \tag{9}$$

Observe from (9) that $\boldsymbol{\Phi}_{\boldsymbol{S}}(\boldsymbol{y})$ is a positive semidefinite matrix. Finally, the MMSE matrix in (8) can be obtained by taking the expectation of $\boldsymbol{\Phi}_{\boldsymbol{S}}(\boldsymbol{y})$ with respect to the distribution of the output, $\mathbf{E}_{\boldsymbol{S}} = \mathsf{E}\{\boldsymbol{\Phi}_{\boldsymbol{S}}(\boldsymbol{Y})\}$.

In the following, we will prove that the entropy, $h(\boldsymbol{Y})$, and the entropy power, $N(\boldsymbol{Y})$, are concave functions with respect to the matrix $\mathbf{Q} \triangleq \mathbf{G}^{\mathsf{T}}\mathbf{G}$ generalizing thus the result in [1, Theorem 6], where we studied the particular case where the matrix $\mathbf{G}$ is diagonal, i.e., $\mathbf{G} = \mathbf{Diag}(\sqrt{\boldsymbol{\mu}})$. Observe that the matrix $\mathbf{Q}$ is not the transmit covariance matrix, which, for the case where $\mathbf{R}_{\boldsymbol{S}} = \mathbf{I}_m$, would be given by $\mathbf{G}\mathbf{G}^{\mathsf{T}}$.

[1] Observe that the fact that there exist particular instances efficiently solvable does not necessarily imply that the problem is not NP-hard, as NP-hardness refers to worst-case performance.

III. Derivation of the new EPI

In order to derive the matrix extension of the EPI in (6) we first need some preliminary results, which are stated in the following lemmas.

Lemma 1: Under the setting described in Section II, the functions $N(\boldsymbol{Y})$ and $h(\boldsymbol{Y})$ are well defined functions of $\mathbf{Q}$.

Proof: From the definition of the entropy power in (1) it is clear that we only need to prove that the entropy function $h(\boldsymbol{Y})$ is well defined in terms of $\mathbf{Q}$. To do so, we first note that $\mathbf{G}^\mathsf{T}\boldsymbol{Y}$ is a sufficient statistic of $\boldsymbol{Y}$, [18, Section 2.10]. Again from [18], we can thus infer that

$$I(\boldsymbol{S};\boldsymbol{Y}) = I(\boldsymbol{S};\mathbf{G}^\mathsf{T}\boldsymbol{Y}) = I(\boldsymbol{S};\mathbf{Q}\boldsymbol{S}+\mathbf{G}^\mathsf{T}\boldsymbol{W}), \tag{10}$$

which, from the definition of mutual information in terms of entropy, immediately yields that

$$h(\boldsymbol{Y}) = h(\mathbf{Q}\boldsymbol{S}+\mathbf{G}^\mathsf{T}\boldsymbol{W}) - h(\mathbf{G}^\mathsf{T}\boldsymbol{W}) + h(\boldsymbol{W}). \tag{11}$$

Noting now that $\mathbf{G}^\mathsf{T}\boldsymbol{W}$ is a zero mean Gaussian random vector, which is fully characterized by its covariance matrix, $\mathbf{Q}$, it is easy to see that all the terms in (11) depend on $\mathbf{G}$ exclusively through $\mathbf{Q}$. ■

Remark 1: Similar results to the one presented in Lemma 1 were already given in [16], [23].

Remark 2: We recall that, as discussed in [1, Section IV], differently from the case stated in Lemma 1, the entropy and entropy power are, in general, not properly defined functions with respect to $\mathbf{G}\mathbf{G}^\mathsf{T}$.

Lemma 2: Under the setting described in Section II, the Hessian matrix of the entropy function $h(\boldsymbol{Y})$ with respect to the (non-repeated) entries of the matrix $\mathbf{Q}$ is given by

$$\mathsf{H}_\mathbf{Q} h(\boldsymbol{Y}) = -\frac{1}{2}\mathbf{D}_m^\mathsf{T}\mathsf{E}\left\{\boldsymbol{\Phi}_{\boldsymbol{S}}(\boldsymbol{Y})\otimes\boldsymbol{\Phi}_{\boldsymbol{S}}(\boldsymbol{Y})\right\}\mathbf{D}_m, \tag{12}$$

where $\mathbf{D}_m$ is the duplication matrix defined in [24].

Proof: Let us start by assuming that $\mathbf{Q}$ is non-singular and compute the Jacobian matrix $\mathsf{D}_\mathbf{Q} h(\boldsymbol{Y})$ through the chain rule $\mathsf{D}_\mathbf{G} h(\boldsymbol{Y}) = \mathsf{D}_\mathbf{Q} h(\boldsymbol{Y})\mathsf{D}_\mathbf{G}\mathbf{Q}$. Now, from [20] and [24, Table 6, Sec. 9.13] we have that $\mathsf{D}_\mathbf{G} h(\boldsymbol{Y}) = \mathsf{vec}^\mathsf{T}(\mathbf{G}\mathbf{E}_{\boldsymbol{S}})$ and $\mathsf{D}_\mathbf{G}\mathbf{Q} = 2\mathbf{D}_m^+(\mathbf{I}_m\otimes\mathbf{G}^\mathsf{T})$, respectively, which yields $\mathsf{D}_\mathbf{Q} h(\boldsymbol{Y}) = \mathsf{vec}^\mathsf{T}(\mathbf{E}_{\boldsymbol{S}})\mathbf{D}_m/2$. Following [24], from this Jacobian matrix, the desired Hessian can be computed as $\mathsf{H}_\mathbf{Q} h(\boldsymbol{Y}) = \mathsf{D}_\mathbf{Q}(\mathsf{D}_\mathbf{Q} h(\boldsymbol{Y}))^\mathsf{T} = \mathbf{D}_m^\mathsf{T}\mathbf{D}_m\mathsf{D}_\mathbf{Q}\mathbf{E}_{\boldsymbol{S}}/2$. Consequently, we need to compute $\mathsf{D}_\mathbf{Q}\mathbf{E}_{\boldsymbol{S}}$, which can be obtained, similarly as we have done before, from $\mathsf{D}_\mathbf{G}\mathbf{E}_{\boldsymbol{S}}$, which was computed in [1] and given by $\mathsf{D}_\mathbf{G}\mathbf{E}_{\boldsymbol{S}} = -2\mathbf{D}_m^+\mathsf{E}\big\{\boldsymbol{\Phi}_{\boldsymbol{S}}(\boldsymbol{Y})\otimes\boldsymbol{\Phi}_{\boldsymbol{S}}(\boldsymbol{Y})\big\}(\mathbf{I}_m\otimes\mathbf{G}^\mathsf{T})$. Recalling from [24] that $\mathbf{D}_m\mathbf{D}_m^+ = \mathbf{N}_m$ and that $\mathbf{D}_m^\mathsf{T}\mathbf{N}_m = \mathbf{D}_m^\mathsf{T}$, the desired result follows.

Now, the derivation for singular $\mathbf{Q}$ follows by first considering $\mathbf{Q}+\epsilon\mathbf{I}_m$, which is non-singular and, thus, all the derivations above apply directly and, then, taking the limit as $\epsilon\to 0$. ■

Remark 3: The expression in (12) holds verbatim if we replace the entropy in the left hand side of the equation with the mutual information $I(\boldsymbol{S};\boldsymbol{Y})$.

Remark 4: From (12), the concavity of the entropy or, equivalently, $\mathsf{H}_\mathbf{Q} h(\boldsymbol{Y}) \leq \mathbf{0}$, follows directly from the facts that $\boldsymbol{\Phi}_{\boldsymbol{S}}(\boldsymbol{y})$ is positive semidefinite $\forall\boldsymbol{y}$ and that the Kronecker product and the expectation operator are closed for positive semidefinite matrices.

Remark 5: Observe that the expression in (12) was already shown by Lamarca in [23]. Although her result is essentially correct, in her expression the duplication matrix $\mathbf{D}_m$ is replaced by the symmetrization matrix $\mathbf{N}_m$ (meaning that she implicitly assumed that the matrix $\mathbf{Q}$ could be nonsymmetric, in which case it is important to highlight that the entropy and mutual information are not properly defined).

Finally, before stating one of the main results in the paper, a matrix result is needed:

Lemma 3: Let $\mathbf{A}\in\mathbb{S}_+^m$ be a symmetric positive definite m-dimensional matrix. Then, it follows that

$$\mathbf{A}\otimes\mathbf{A} \geq \frac{\mathsf{vec}(\mathbf{A})\mathsf{vec}^\mathsf{T}(\mathbf{A})}{m}, \tag{13}$$

where $\mathsf{vec}(\mathbf{A})$ represents a column vector obtained by stacking the columns in matrix $\mathbf{A}$.

Proof: Let $\mathbf{A}\in\mathbb{R}^{m\times m}$ and $\mathbf{B}\in\mathbb{R}^{m\times m}$ be two square matrices and consider their inner product [25, Section 5.1]:

$$\langle\mathbf{A},\mathbf{B}\rangle = \mathsf{Tr}\big(\mathbf{B}^\mathsf{T}\mathbf{A}\big). \tag{14}$$

Let's also consider the square matrix $\mathbf{Z}\in\mathbb{R}^{m\times m}$ and define $\mathbf{z} = \mathsf{vec}(\mathbf{Z})\in\mathbb{R}^{m^2}$.

Now, we use the fact that the matrix $\mathbf{A}$ is positive semidefinite, which implies that $\mathbf{A}^{1/2}$ exists and is well defined, and compute the following inner products:

$$\begin{aligned}
\langle\mathbf{I}_m,\mathbf{I}_m\rangle &= m\\
\langle\mathbf{A}^{1/2}\mathbf{Z}\mathbf{A}^{1/2},\mathbf{I}_m\rangle &= \mathsf{Tr}(\mathbf{A}\mathbf{Z})\\
&= \mathsf{vec}^\mathsf{T}(\mathbf{A})\mathbf{z}\\
\langle\mathbf{A}^{1/2}\mathbf{Z}\mathbf{A}^{1/2},\mathbf{A}^{1/2}\mathbf{Z}\mathbf{A}^{1/2}\rangle &= \mathsf{Tr}\big(\mathbf{Z}^\mathsf{T}\mathbf{A}\mathbf{Z}\mathbf{A}\big)\\
&= \mathbf{z}^\mathsf{T}\mathsf{vec}(\mathbf{A}\mathbf{Z}\mathbf{A})\\
&= \mathbf{z}^\mathsf{T}(\mathbf{A}\otimes\mathbf{A})\mathbf{z},
\end{aligned} \tag{15}$$

where we have used $\mathsf{vec}^\mathsf{T}(\mathbf{A})\mathsf{vec}(\mathbf{B}) = \mathsf{Tr}\big(\mathbf{A}^\mathsf{T}\mathbf{B}\big)$ and $\mathsf{vec}(\mathbf{A}\mathbf{Z}\mathbf{B}) = (\mathbf{B}^\mathsf{T}\otimes\mathbf{A})\mathsf{vec}(\mathbf{Z})$.

Applying the Cauchy-Schwarz inequality we have

$$\langle\mathbf{I}_m,\mathbf{I}_m\rangle\langle\mathbf{A}^{1/2}\mathbf{Z}\mathbf{A}^{1/2},\mathbf{A}^{1/2}\mathbf{Z}\mathbf{A}^{1/2}\rangle \geq \langle\mathbf{A}^{1/2}\mathbf{Z}\mathbf{A}^{1/2},\mathbf{I}_m\rangle^2.$$

From the inequality above and (15) we obtain

$$m\cdot\mathbf{z}^\mathsf{T}(\mathbf{A}\otimes\mathbf{A})\mathbf{z} \geq \mathbf{z}^\mathsf{T}\mathsf{vec}(\mathbf{A})\mathsf{vec}^\mathsf{T}(\mathbf{A})\mathbf{z}, \tag{16}$$

which, since the choice of $\mathbf{z}$ is arbitrary, implies the desired result in (13). ■

Remark 6: The result in Lemma 3 is a generalization of the result proved in [1, Proposition H.9] for $\mathbf{A}\in\mathbb{R}^{m\times m}$:

$$\mathbf{A}\circ\mathbf{A} \geq \frac{\mathbf{diag}(\mathbf{A})\mathbf{diag}(\mathbf{A})^\mathsf{T}}{m}. \tag{17}$$

Observe that (17) follows from (13) by recalling the definition of the reduction matrix, $\mathbf{S}_m$, in [1, Appendix A] such that we have $\mathbf{S}_m^\mathsf{T}(\mathbf{A}\otimes\mathbf{A})\mathbf{S}_m = \mathbf{A}\circ\mathbf{A}$ and $\mathsf{vec}^\mathsf{T}(\mathbf{A})\mathbf{S}_m = \mathbf{diag}(\mathbf{A})^\mathsf{T}$.

We are now ready to state the following result.

Theorem 1: Under the setting described in Section II, the Hessian matrix of the entropy power function $N(\boldsymbol{Y})$ with respect to the (non-repeated) entries of matrix $\mathbf{Q}$ is given by

$$\mathsf{H}_{\mathbf{Q}} N(\boldsymbol{Y}) = \frac{N(\boldsymbol{Y})}{m} \mathbf{D}_m^{\mathsf{T}} \Big(\frac{\mathsf{vec}(\mathbf{E}_{\boldsymbol{S}}) \mathsf{vec}^{\mathsf{T}}(\mathbf{E}_{\boldsymbol{S}})}{m} - \mathsf{E}\left\{ \boldsymbol{\Phi}_{\boldsymbol{S}}(\boldsymbol{Y}) \otimes \boldsymbol{\Phi}_{\boldsymbol{S}}(\boldsymbol{Y}) \right\} \Big) \mathbf{D}_m. \quad (18)$$

Moreover, we have that $\mathsf{H}_{\mathbf{Q}} N(\boldsymbol{Y}) \leq \mathbf{0}$.

Proof: From (12) and [1, Equation (152)], the expression for the Hessian of the entropy power can be straightforwardly computed as given in (18). Now, recalling that, for any arbitrary random vector $\boldsymbol{Z}$, we have that

$$\mathsf{E}\{\boldsymbol{Z}\boldsymbol{Z}^{\mathsf{T}}\} \geq \mathsf{E}\{\boldsymbol{Z}\}\mathsf{E}\{\boldsymbol{Z}^{\mathsf{T}}\} \quad (19)$$

and using the result in Lemma 3 together with $\mathbf{E}_{\boldsymbol{S}} = \mathsf{E}\left\{\boldsymbol{\Phi}_{\boldsymbol{S}}(\boldsymbol{Y})\right\}$, it can be seen that $\mathsf{H}_{\mathbf{Q}} N(\boldsymbol{Y})$ is negative semidefinite following similar steps as in [1, Proof Th. 6]. ■

IV. Application to an optimization problem

One possible application of the results presented in the section above is to consider the problem of computing the linear precoder that maximizes the mutual information in linear vector Gaussian channels with arbitrarily distributed inputs. To that purpose, let's consider the following variation on the model in (7):

$$\boldsymbol{Y} = \mathbf{H}\mathbf{P}\boldsymbol{S} + \boldsymbol{W}. \quad (20)$$

Observe that we have decomposed the linear transformation, $\mathbf{G}$, undergone by the input $\boldsymbol{S}$ in (7) into two factors: $\mathbf{H} \in \mathbb{R}^{n \times p}$, which represents the channel transformation (given by the nature) and $\mathbf{P} \in \mathbb{R}^{p \times m}$, which is the precoding linear transformation (to be determined by the system designer). For a given input distribution on $\boldsymbol{S}$, our goal is to solve

$$\underset{\mathbf{P}}{\text{maximize}} \quad I(\boldsymbol{S}; \mathbf{H}\mathbf{P}\boldsymbol{S} + \boldsymbol{W}) \quad (21)$$

$$\text{subject to} \quad \mathsf{Tr}\big(\mathbf{P}\mathbf{R}_{\boldsymbol{S}}\mathbf{P}^{\mathsf{T}}\big) \leq P_T, \quad (22)$$

where P_T is the transmitted power constraint and $\mathbf{R}_{\boldsymbol{S}} = \mathsf{E}\{\boldsymbol{S}\boldsymbol{S}^{\mathsf{T}}\}$. Without loss of generality, from this point we assume $\mathbf{R}_{\boldsymbol{S}} = \mathbf{I}_m$.

The problem in (21)-(22) was studied in [16] and it was shown that, given the eigendecomposition $\mathbf{H}^{\mathsf{T}}\mathbf{H} = \mathbf{U}_{\mathbf{H}}\boldsymbol{\Lambda}_{\mathbf{H}}\mathbf{U}_{\mathbf{H}}^{\mathsf{T}}$, the optimal precoder structure is given by

$$\mathbf{P}^{\star} = \mathbf{U}_{\mathbf{H}}\mathbf{Diag}(\sqrt{\boldsymbol{\lambda}})\mathbf{V}^{\mathsf{T}}, \quad (23)$$

where $\mathbf{V}$ is constrained to be unitary. Consequently, only $\boldsymbol{\lambda}$ and $\mathbf{V}$ remain to be determined.

From this point, we define the quantity $\mathbf{Q}_{\mathbf{H}} = \mathbf{P}^{\star\mathsf{T}}\mathbf{H}^{\mathsf{T}}\mathbf{H}\mathbf{P}^{\star} = \mathbf{V}\mathbf{Diag}(\boldsymbol{\lambda})\boldsymbol{\Lambda}_{\mathbf{H}}\mathbf{V}^{\mathsf{T}}$. Observe that the following equality holds

$$I(\boldsymbol{S}; \mathbf{H}\mathbf{P}^{\star}\boldsymbol{S} + \boldsymbol{W}) = I(\boldsymbol{S}; \mathbf{Q}_{\mathbf{H}}^{1/2}\boldsymbol{S} + \boldsymbol{W}), \quad (24)$$

which highlights that the mutual information can be expressed as a function of $\mathbf{Q}_{\mathbf{H}}$ because $\mathbf{Q}_{\mathbf{H}}^{1/2}$ is uniquely defined up to a left orthogonal transformation, which can be removed because the noise distribution is also invariant to left orthogonal transformations.

Consequently, from the results derived in the previous section, we now know that the mutual information $I(\boldsymbol{S}; \mathbf{Q}_{\mathbf{H}}^{1/2}\boldsymbol{S} + \boldsymbol{W})$ is a concave function with respect to the matrix $\mathbf{Q}_{\mathbf{H}}$, which yields interesting consequences in the study of the optimization problem in (21)-(22).

A. Problem approach in [23]

The method proposed by Lamarca to solve (21)-(22) consists in iteratively performing the following two steps until convergence:

1) For a fixed $\mathbf{V}$, compute the $\boldsymbol{\lambda}$ that maximizes the mutual information and fulfills the power constraint, in an analogous way as our proposed scheme in [16]. As it was shown in [1], the mutual information is concave with respect to $\boldsymbol{\lambda}$ and, thus, for a given $\mathbf{V}$, convergence to the optimal $\boldsymbol{\lambda}$ is guaranteed.

2) For a fixed $\boldsymbol{\lambda}$, solve the following optimization problem:

$$\underset{\mathbf{Q}_{\mathbf{H}}}{\text{maximize}} \quad I(\boldsymbol{S}; \mathbf{Q}_{\mathbf{H}}^{1/2}\boldsymbol{S} + \boldsymbol{W}) \quad (25)$$

$$\text{subject to} \quad [\mathbf{eigs}(\mathbf{Q}_{\mathbf{H}})]_i = [\boldsymbol{\lambda}]_i [\boldsymbol{\Lambda}_{\mathbf{H}}]_{ii}, \quad \forall i, \quad (26)$$

and, then, compute $\mathbf{V}$ from $\boldsymbol{\Lambda}_{\mathbf{H}}$, $\boldsymbol{\lambda}$, and the obtained $\mathbf{Q}_{\mathbf{H}}$, which fulfills $\mathbf{Q}_{\mathbf{H}} = \mathbf{V}\mathbf{Diag}(\boldsymbol{\lambda})\boldsymbol{\Lambda}_{\mathbf{H}}\mathbf{V}^{\mathsf{T}}$. Observe that in (26) we have used $\mathbf{eigs}(\cdot)$ to denote a vector containing the eigenvalues of its matrix argument.

Remark 7: Although the objective function in (25) is concave with respect to $\mathbf{Q}_{\mathbf{H}}$, the constraint in (26) is not convex. Thus, the resulting problem is non-convex and no optimality of the obtained solution can be guaranteed. The lack of optimality guarantee can be intuitively related to the computational hardness of the problem which was studied in [16].

B. Our approach to the problem

Capitalizing on the problem in (25)-(26), we propose to reformulate it by rewriting the set of constraints in (26) as a unique constraint by adding them together. Thus, following our approach, the general (non-convex) problem is

$$\underset{\mathbf{Q}_{\mathbf{H}}}{\text{maximize}} \quad I(\boldsymbol{S}; \mathbf{Q}_{\mathbf{H}}^{1/2}\boldsymbol{S} + \boldsymbol{W}) \quad (27)$$

$$\text{subject to} \quad \sum_i [\mathbf{eigs}(\mathbf{Q}_{\mathbf{H}})]_i / [\boldsymbol{\Lambda}_{\mathbf{H}}]_{ii} \leq P_T, \quad (28)$$

$$\mathbf{rank}(\mathbf{Q}_{\mathbf{H}}) \leq m, \quad (29)$$

which, similarly as the problem in (25)-(26), is nonconvex. Observe that the rank constraint on $\mathbf{Q}_{\mathbf{H}}$, which was missing in (25)-(26), has been added to ensure that the rank of $\mathbf{Q}_{\mathbf{H}}$ is smaller than the dimension of the input signal $\boldsymbol{S}$ as, otherwise, the result of the optimization problem would be lacking practical applicability. Note also that for the case where $m \geq p$ the rank constraint is automatically satisfied and can be removed.

Despite the fact that our problem formulation in (27)-(29) does not present any a priori advantage with respect to the formulation in (25)-(26), in the following we will present a particular case, which yields a surprising result.

For the case where $\mathbf{\Lambda_H} = \alpha \mathbf{I}$ and $m \geq p$, it is straightforward to see that the problem in (27)-(29) simplifies to

$$\underset{\mathbf{Q}}{\text{maximize}} \quad I(\boldsymbol{S}; \sqrt{\alpha}\mathbf{Q}^{1/2}\boldsymbol{S} + \boldsymbol{W}) \tag{30}$$

$$\text{subject to} \quad \mathsf{Tr}(\mathbf{Q}) \leq P_T, \tag{31}$$

which, according to the results in Secion III, is a convex optimization problem. We recall that $\mathbf{Q} = \mathbf{G}^{\mathsf{T}}\mathbf{G}$ and, thus, when $\mathbf{\Lambda_H} = \alpha\mathbf{I}$ we have $\mathbf{Q_H} = \alpha\mathbf{Q}$. This is a quite surprising and interesting result because, although its application to practical systems is rather limited (requiring that all the channel singular values are equal $\mathbf{\Lambda_H} = \alpha\mathbf{I}$), we have an instance of the problem of maximization of mutual information such that the optimality of the obtained solution can be guaranteed.

C. Algorithms comparison

For the particular case where $\mathbf{\Lambda_H} = \alpha\mathbf{I}$, we have performed a comparison of the precoder optimization algorithms in [22] and [23] with a simple implementation of the Newton method to solve our proposed optimization problem in (30)-(31).

In Fig.1 we have depicted the evolution of the mutual information as a function of the index of the iteration for the three different algorithms mentioned in the previous paragraph. For the sake of simplicity we have taken $\alpha = P_T = 1$, $n = m = p = 4$, and the distribution of $\boldsymbol{S}$ has been chosen as $P_{\boldsymbol{S}}(\boldsymbol{s}) = \sum_{i=1}^{64} \delta(\boldsymbol{s} - \mathbf{s}_i)$, where the values of $\mathbf{s}_i$ have been assigned randomly and are not reproduced here for the sake of space.

From the figure it can be seen that, not only is our proposed algorithm the one yielding a faster convergence, but it is also the only one that can obtain the global solution as the others can get trapped in local maxima.

V. Conclusion

In this paper, we have derived a generalization of the vector EPI presented in [1] to the full matrix case. Capitalizing on this new EPI, we have proved that the problem of computing the linear precoder that maximizes the mutual information in linear vector Gaussian channels with arbitrary inputs has particular instances where the optimal precoder can be computed efficiently. We see the presence of particular instances that can be efficiently solved as a starting point to find other efficiently solvable instances of the problem with higher practical interest.

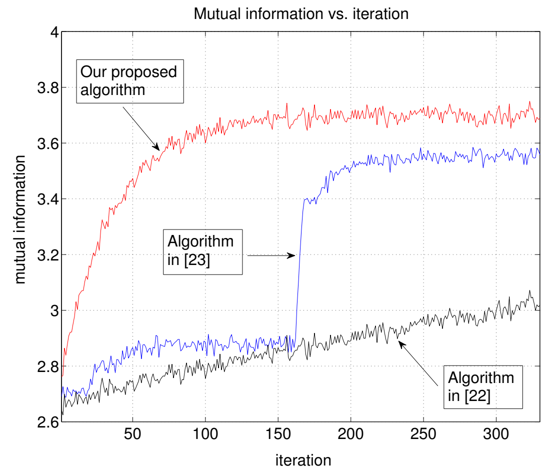

Fig. 1. Comparison of three different algorithms that optimize the precoder so that the mutual information is maximized.

References

[1] M. Payaró and D. P. Palomar, "Hessian and concavity of mutual information, differential entropy, and entropy power in linear vector Gaussian channels," *IEEE Trans. on Information Theory*, vol. 55, no. 8, pp. 3613 – 3628, Aug. 2009.

[2] C. E. Shannon, "A mathematical theory of communication," *Bell Syst Tech. J*, vol. 27, no. 3, pp. 379–423, 1948.

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

[4] N. Blachman, "The convolution inequality for entropy powers," *IEEE Trans. on Information Theory*, vol. 11, no. 2, pp. 267–271, 1965.

[5] S. Verdú and D. Guo, "A simple proof of the entropy-power inequality," *IEEE Trans. on Information Theory*, vol. 52, no. 5, pp. 2165–2166, 2006.

[6] P. Bergmans, "A simple converse for broadcast channels with additive white Gaussian noise," *IEEE Trans. on Information Theory*, vol. 20, no. 2, pp. 279–280, 1974.

[7] Y. Oohama, "Rate-distortion theory for Gaussian multiterminal source coding systems with several side informations at the decoder," *IEEE Trans. on Information Theory*, vol. 51, no. 7, pp. 2577–2593, 2005.

[8] M. H. M. Costa, "A new entropy power inequality," *IEEE Trans. on Information Theory*, vol. 31, no. 6, pp. 751–760, 1985.

[9] A. Dembo, "Simple proof of the concavity of the entropy power with respect to added Gaussian noise," *IEEE Trans. on Information Theory*, vol. 35, no. 4, pp. 887–888, 1989.

[10] C. Villani, "A short proof of the "concavity of entropy power"," *IEEE Trans. on Information Theory*, vol. 46, no. 4, pp. 1695–1696, 2000.

[11] O. Rioul, "Information Theoretic Proofs of Entropy Power Inequalities," *IEEE Trans. on Information Theory*, vol. 57, no. 1, pp. 33–55, Jan. 2011.

[12] M. H. M. Costa, "On the Gaussian interference channel," *IEEE Trans. on Information Theory*, vol. 31, no. 5, pp. 607–615, 1985.

[13] A. Lapidoth and S. M. Moser, "Capacity bounds via duality with applications to multiple-antenna systems on flat-fading channels," *IEEE Trans. on Information Theory*, vol. 49, no. 10, pp. 2426–2467, 2003.

[14] R. Liu, T. Liu, H. Poor, and S. Shamai, "A vector generalization of Costa's entropy-power inequality with applications," *IEEE Trans. on Information Theory*, vol. 56, no. 4, pp. 1865–1879, Apr. 2010.

[15] R. Tandon and S. Ulukus, "Dependence balance based outer bounds for Gaussian networks with cooperation and feedback," *Submitted to IEEE Trans. on Information Theory*, 2008.

[16] M. Payaró and D. P. Palomar, "On optimal precoding in linear vector Gaussian channels with arbitrary input distribution," in *Proc. IEEE Intl. Symp. on Information Theory (ISIT'09)*, Seoul, Korea, July 2009.

[17] E. Telatar, "Capacity of multi-antenna Gaussian channels," *European Trans. on Telecomm.*, vol. 10, no. 6, pp. 585–595, Nov.-Dec. 1999.

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

[19] D. Guo, S. Shamai, and S. Verdú, "Mutual information and minimum mean-square error in Gaussian channels," *IEEE Trans. on Information Theory*, vol. 51, no. 4, pp. 1261–1282, 2005.

[20] D. P. Palomar and S. Verdú, "Gradient of mutual information in linear vector Gaussian channels," *IEEE Trans. on Information Theory*, vol. 52, no. 1, pp. 141–154, 2006.

[21] A. Lozano, A. Tulino, and S. Verdú, "Optimum power allocation for parallel Gaussian channels with arbitrary input distributions," *IEEE Trans. on Information Theory*, vol. 52, no. 7, pp. 3033–3051, July 2006.

[22] F. Pérez-Cruz, M. Rodrigues, and S. Verdú, "MIMO Gaussian channels with arbitrary inputs: optimal precoding and power allocation," *IEEE Trans. on Information Theory*, vol. 56, no. 3, pp. 1070–1084, Mar. 2010.

[23] M. Lamarca, "Linear precoding for mutual information maximization in MIMO systems," in *Proc. IEEE Intl. Symp. Wireless Communication Systems*, Sep. 2009, pp. 26–30.

[24] J. Magnus and H. Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics*, 3rd ed. New York: Wiley, 2007.

[25] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 1985.