

Proof of Entropy Power Inequalities Via MMSE

Dongning Guo

Dept. of Electrical Engineering &
Computer Science, Northwestern University
Evanston, IL 60208, USA
Email: dGuo@Northwestern.edu

Shlomo Shamai (Shitz)

Dept. of Electrical Engineering
Technion-Israel Inst. of Tech.
32000 Haifa, Israel
Email: sshlomo@ee.technion.ac.il

Sergio Verdú

Dept. of Electrical Engineering
Princeton University
Princeton, NJ 08544, USA
Email: Verdu@Princeton.edu

Abstract—The differential entropy of a random variable (or vector) can be expressed as the integral over signal-to-noise ratio (SNR) of the minimum mean-square error (MMSE) of estimating the variable (or vector) when observed in additive Gaussian noise. This representation sidesteps Fisher’s information to provide simple and insightful proofs for Shannon’s entropy power inequality (EPI) and two of its variations: Costa’s strengthened EPI in the case in which one of the variables is Gaussian, and a generalized EPI for linear transformations of a random vector due to Zamir and Feder.

I. INTRODUCTION

The mean squared error is one of the most studied criterion in statistical sciences, signal processing, optimization, and detection and estimation. Given any jointly distributed random variables (U, V) , the minimum mean-square error (MMSE) of estimating U given the observation V is achieved by the conditional mean. For convenience, we denote this MMSE by

$$\text{mmse}(U|V) = \mathbb{E} \{ (U - \mathbb{E} \{ U | V \})^2 \}.$$

Consider further the following observation contaminated by additive Gaussian noise:

$$\mathbf{Y} = \sqrt{\gamma} \mathbf{X} + \mathbf{N} \quad (1)$$

where \mathbf{X} and \mathbf{N} are the independent input and noise vectors of the same dimension n , and $\gamma > 0$ is the signal-to-noise ratio. Throughout this paper, $\mathbf{N} \sim \mathcal{N}(0, \mathbf{I})$ denotes a vector of appropriate dimension with independent standard Gaussian random entries. We denote the MMSE of estimating \mathbf{X} given \mathbf{Y} by

$$\text{mmse}(\mathbf{X}, \gamma) = \text{mmse}(\mathbf{X} | \sqrt{\gamma} \mathbf{X} + \mathbf{N}) \quad (2)$$

which is a decreasing function of γ , and is dependent on the input distribution $P_{\mathbf{X}}$.

There has been a renaissance of interest in the MMSE in the form of (2) since the discovery of its fundamental relationship with the input–output mutual information of the underlying Gaussian channel:¹

$$\frac{d}{d\gamma} I(\mathbf{X}; \sqrt{\gamma} \mathbf{X} + \mathbf{N}) = \frac{1}{2} \text{mmse}(\mathbf{X} | \sqrt{\gamma} \mathbf{X} + \mathbf{N}) \quad (3)$$

which holds regardless of the distribution of \mathbf{X} [1]. The new relationship has proven to be not only convenient for proving

classical results such as the de Bruijn identity [2] and Duncan’s relation for mutual information and causal MMSE [3], but also pivotal for establishing new results, e.g., the relationship between the causal and noncausal MMSEs in discrete- and continuous-time Gaussian channels [1].

As a direct consequence of (3), several information measures have been expressed as integrals of the MMSE [1], [4], [5]. In particular, the differential entropy of an n dimensional random vector \mathbf{X} is [6] (see also [1], [5])

$$h(\mathbf{X}) = \frac{n}{2} \log(2\pi e) - \frac{1}{2} \int_0^\infty \frac{n}{1+\gamma} - \text{mmse}(\mathbf{X}, \gamma) d\gamma. \quad (4)$$

In light of the entropy-MMSE relationship, we questioned in [1] whether there exists an estimation-theoretic proof of Shannon’s celebrated entropy power inequality (EPI). In the vector form, the EPI is given as the following.

Theorem 1 (Shannon): For any independent n -vectors \mathbf{X} and \mathbf{Y} ,

$$\exp \left[\frac{2}{n} h(\mathbf{X} + \mathbf{Y}) \right] \geq \exp \left[\frac{2}{n} h(\mathbf{X}) \right] + \exp \left[\frac{2}{n} h(\mathbf{Y}) \right]. \quad (5)$$

A simple proof of Theorem 1 was given [6] using the information–estimation relationship (4). This is possible because, in view of the representation of the differential entropy in the MMSE, the EPI (5) can be seen as a relationship between the (integrals over SNR of the) MMSE of the sum and individual random vectors. In this paper we apply the relationships (3) and (4) to prove two variations of Shannon’s EPI:

- Costa’s strengthened EPI in which one of the variables Gaussian is Gaussian [7];
- The generalized EPI for linear transformations of random vector due to Zamir and Feder [8].

The proofs of Shannon’s EPI [6] and Zamir-Feder’s EPI hinge on generalizations of Lieb’s inequality [9], which are proved through elementary estimation-theoretic reasonings. The relationship between mutual information and MMSE, together with classical rate distortion arguments, leads to a simple proof of the concavity of the entropy power of a vector with additive Gaussian noise as a function of the noise variance, which is equivalent to Costa’s EPI. The new proofs are based on elementary estimation-theoretic reasonings which sidestep invoking the Fisher’s information, in contrast to existing proofs.

This work was partially supported by NSF Grants NCR-0074277 and CCR-0312879, and by the U.S.-Israel Binational Science Foundation.

¹Throughout this paper, we assume the units of information measures to be “nats” and that all logarithms are natural.

Through [6], the new proof of the EPIs in this paper, the new representations of information measures [1], as well as a recent new proof of the monotone increase of the entropy of the sum of independent random variables in [10], the MMSE has invariably proven to be more convenient and insightful than Fisher's information.

The remainder of the paper is organized as follows. The proof of Shannon's EPI given in [6] is included in Section II for the reader's convenience. The Zamir-Feder generalization is proved in Section III. Costa's EPI is treated in Section IV.

II. PROOF OF SHANNON'S EPI [6]

The original scalar version of the EPI was put forth by Shannon in [11]. It has been used in the converse of the capacity region of the degraded Gaussian broadcast channel [12], the rate distortion of multiple Gaussian source [13], and more recently the rate characterization of the CEO problem [14].

We first prove the following inequality (originally due to Lieb [9]): For any independent vectors \mathbf{X}_1 and \mathbf{X}_2 of the same dimension and $\alpha \in [0, 2\pi]$,

$$h(\mathbf{X}_1 \cos \alpha + \mathbf{X}_2 \sin \alpha) \geq h(\mathbf{X}_1) \cos^2 \alpha + h(\mathbf{X}_2) \sin^2 \alpha. \quad (6)$$

According to (4), it suffices to show that for all γ ,

$$\text{mmse}(\mathbf{X}, \gamma) \geq \text{mmse}(\mathbf{X}_1, \gamma) \cos^2 \alpha + \text{mmse}(\mathbf{X}_2, \gamma) \sin^2 \alpha \quad (7)$$

where $\mathbf{X} = \mathbf{X}_1 \cos \alpha + \mathbf{X}_2 \sin \alpha$. Let

$$\begin{aligned} \mathbf{Z}_1 &= \sqrt{\gamma} \mathbf{X}_1 + \mathbf{N}_1 \\ \mathbf{Z}_2 &= \sqrt{\gamma} \mathbf{X}_2 + \mathbf{N}_2 \\ \mathbf{Z} &= \mathbf{Z}_1 \cos \alpha + \mathbf{Z}_2 \sin \alpha \end{aligned}$$

where $\mathbf{N}_1 \sim \mathcal{N}(0, \mathbf{I})$ and $\mathbf{N}_2 \sim \mathcal{N}(0, \mathbf{I})$ are independent. Then, the left side of (7) can be written as

$$\text{mmse}(\mathbf{X}|\mathbf{Z}) \geq \mathbb{E} \left\{ (\mathbf{X} - \mathbb{E} \{ \mathbf{X} | \mathbf{Z}_1, \mathbf{Z}_2 \})^2 \right\} \quad (8)$$

$$\begin{aligned} &= \cos^2 \alpha \mathbb{E} \left\{ (\mathbf{X}_1 - \mathbb{E} \{ \mathbf{X}_1 | \mathbf{Z}_1 \})^2 \right\} \\ &\quad + \sin^2 \alpha \mathbb{E} \left\{ (\mathbf{X}_2 - \mathbb{E} \{ \mathbf{X}_2 | \mathbf{Z}_2 \})^2 \right\} \quad (9) \end{aligned}$$

thereby showing (7). Note that in (9) we have used the independence of $\mathbf{N}_1, \mathbf{N}_2, \mathbf{X}_1, \mathbf{X}_2$.

For arbitrary independent n -vectors \mathbf{X} and \mathbf{Y} , let

$$\cos \alpha = \frac{\exp \left[\frac{1}{n} h(\mathbf{X}) \right]}{\sqrt{\exp \left[\frac{2}{n} h(\mathbf{X}) \right] + \exp \left[\frac{2}{n} h(\mathbf{Y}) \right]}}, \quad (10)$$

$$\mathbf{X}_1 = \frac{\mathbf{X}}{\cos \alpha}, \quad \text{and} \quad \mathbf{X}_2 = \frac{\mathbf{Y}}{\sin \alpha}. \quad (11)$$

Then, using (6) and $h(a\mathbf{V}) = h(\mathbf{V}) + n \log |a|$ for any random n -vector \mathbf{V} , the EPI follows:

$$\begin{aligned} \frac{1}{n} h(\mathbf{X} + \mathbf{Y}) &\geq \cos^2 \alpha \left(\frac{1}{n} h(\mathbf{X}) - \log |\cos \alpha| \right) \\ &\quad + \sin^2 \alpha \left(\frac{1}{n} h(\mathbf{Y}) - \log |\sin \alpha| \right) \\ &= \frac{1}{2} \log \left(\exp \left[\frac{2}{n} h(\mathbf{X}) \right] + \exp \left[\frac{2}{n} h(\mathbf{Y}) \right] \right). \end{aligned}$$

Note that the only inequality used in the proof of (5) is (8), namely the fact that for the estimation of the sum of two random variables it is preferable to have access to individual noisy measurements than to the sum of the measurements.

III. A PROOF OF ZAMIR AND FEDER'S GENERALIZED EPI

Variations of the EPI are key to certain information theoretic relations [15]. In [8], Zamir and Feder generalized the EPI to a lower bound on the entropy power of a linear transformation of a random vector. The result has been used to study the rate distortion of bandlimited sources [16] and lower bound the volume of Minkowski set sums [17].

Theorem 2 (Zamir and Feder [8]): Let \mathbf{X} be an n -vector with independent entries. Then for every $m \times n$ matrix \mathbf{A} ,

$$h(\mathbf{A}\mathbf{X}) \geq h(\mathbf{A}\tilde{\mathbf{X}}) \quad (12)$$

where $\tilde{\mathbf{X}}$ consists of independent Gaussian components with $h(\tilde{X}_i) = h(X_i)$ for $i = 1, \dots, n$.

Theorem 2 reduces to Shannon's original EPI if the linear transformation is the following 1×2 vector: $\mathbf{A} = [\cos \alpha, \sin \alpha]$.

Let p_i be the entropy power of X_i , i.e., $h(X_i) = \frac{1}{2} \log(2\pi e p_i)$. Let $\mathbf{P} = \text{diag}(p_1, \dots, p_n)$. Then the entropy of the linear transformed Gaussian vector can be obtained as

$$h(\mathbf{A}\tilde{\mathbf{X}}) = \frac{1}{2} \log \det (2\pi e \mathbf{A} \mathbf{P} \mathbf{A}^T). \quad (13)$$

The original proof of the generalized EPI of Zamir and Feder through a double induction on the rows and columns of the linear transformation is quite involved [8]. So far, the simplest known proof of Theorem 2 involves de Bruijn's inequality and the following matrix version of Lieb's inequality, proved in [18] using a matrix Fisher information inequality.

A. Vector Version of Lieb's Inequality

Theorem 3 (Zamir and Feder [18]): Let \mathbf{X} be an n -vector with independent entries. Then

$$h(\mathbf{U}\mathbf{X}) \geq \text{tr} \{ \mathbf{U} \text{diag}(h(X_1), \dots, h(X_n)) \mathbf{U}^T \} \quad (14)$$

for every $m \times n$ matrix \mathbf{U} with orthonormal rows ($\mathbf{U}\mathbf{U}^T = \mathbf{I}$).

This paper presents an alternative simple proof of Theorem 3 via MMSE following in principle the proof of (6) in Section II. While [18] showed that Theorem 2 is a corollary of Theorem 3, here we actually prove that they are equivalent.

Proof: [Theorem 3] For each $i = 1, \dots, n$, the differential entropy $h(X_i)$ can be obtained as an integral of

$$d_i = \text{mmse}(X_i | \sqrt{\gamma} X_i + N) \quad (15)$$

according to the scalar version of (4). Furthermore, by (4),

$$\begin{aligned} h(\mathbf{U}\mathbf{X}) &= \frac{m}{2} \log(2\pi e) \\ &\quad - \frac{1}{2} \int_0^\infty \frac{m}{1+\gamma} - \text{mmse}(\mathbf{U}\mathbf{X} | \sqrt{\gamma} \mathbf{U}\mathbf{X} + N) \, d\gamma \quad (16) \end{aligned}$$

where $\mathbf{N} \sim \mathcal{N}(0, \mathbf{I}_{m \times m})$. Since $\mathbf{U}\mathbf{U}^\top = \mathbf{I}$, conditioned on \mathbf{X} , $\sqrt{\gamma}\mathbf{U}\mathbf{X} + \mathbf{N}$ has the same statistics as $\mathbf{U}(\sqrt{\gamma}\mathbf{X} + \mathbf{N}')$ where $\mathbf{N}' \sim \mathcal{N}(0, \mathbf{I}_{n \times n})$. Consequently,

$$\begin{aligned} \text{mmse}(\mathbf{U}\mathbf{X}|\sqrt{\gamma}\mathbf{U}\mathbf{X} + \mathbf{N}) &= \text{mmse}(\mathbf{U}\mathbf{X}|\mathbf{U}(\sqrt{\gamma}\mathbf{X} + \mathbf{N}')) \\ &\geq \text{mmse}(\mathbf{U}\mathbf{X}|\sqrt{\gamma}\mathbf{X} + \mathbf{N}') \\ &= \text{tr}\{\mathbf{U}\mathbf{D}\mathbf{U}^\top\} \end{aligned} \quad (17)$$

where \mathbf{D} is an $n \times n$ diagonal matrix with diagonal elements (d_1, \dots, d_n) . Therefore, by (15) and (16),

$$\begin{aligned} h(\mathbf{U}\mathbf{X}) - \text{tr}\{\mathbf{U} \text{diag}(h(X_1), \dots, h(X_n)) \mathbf{U}^\top\} \\ = \frac{1}{2} \int_0^\infty \text{mmse}(\mathbf{U}\mathbf{X}|\sqrt{\gamma}\mathbf{U}\mathbf{X} + \mathbf{N}) - \text{tr}\{\mathbf{U}\mathbf{D}\mathbf{U}^\top\} d\gamma \end{aligned}$$

which is nonnegative since the integrand is nonnegative for every γ according to (17). ■

Following an observation in [6], note that Theorem 3 has a counterpart for vectors whose entries are finitely or countably valued random variables:

$$H(\mathbf{U}\mathbf{X}) \geq \text{tr}\{\mathbf{U} \text{diag}(H(X_1), \dots, H(X_n)) \mathbf{U}^\top\}.$$

This is due to the MMSE-integral representation of the entropy [1, (176)], and that the argument we gave above for (17) holds also for discrete random variables.

B. Proof of Theorem 2 Using Theorem 3

Theorem 3 is boot-strapped to a proof of Theorem 2.

Proof: [Theorem 3 \Rightarrow Theorem 2] First, (12) holds if $\mathbf{A} = \mathbf{U}$ contains m orthonormal rows and that all the components of \mathbf{X} have the same differential entropy, i.e., $h(X_i) = h(X_j)$ for all $i, j = 1, \dots, n$. In this case,

$$h(\mathbf{U}\mathbf{X}) \geq \text{tr}\{\mathbf{U}\mathbf{U}^\top\}h(X_1) \quad (18)$$

$$\begin{aligned} &= m \cdot h(X_1) \\ &= h(\mathbf{U}\tilde{\mathbf{X}}) \end{aligned} \quad (19)$$

where (18) is by (14) and (19) is by (13) because $\tilde{\mathbf{X}}$ consists of independent identically distributed Gaussian entries.

Consider now an arbitrary $m \times n$ matrix \mathbf{A} with full row rank. By the QR factorization of \mathbf{A}^\top [19, Theorem 2.6.1], there exist an invertible matrix \mathbf{R} and an $m \times n$ matrix \mathbf{U} with orthonormal rows such that $\mathbf{A} = \mathbf{R}\mathbf{U}$. In fact, \mathbf{R} is lower triangular and can be regarded as the composition of a sequence of row operators that orthogonalizes and normalizes the rows of \mathbf{A} . Clearly

$$\begin{aligned} h(\mathbf{A}\mathbf{X}) &= h(\mathbf{U}\mathbf{X}) + \log |\det \mathbf{R}| \\ &\geq h(\mathbf{U}\tilde{\mathbf{X}}) + \log |\det \mathbf{R}| \\ &= h(\mathbf{A}\tilde{\mathbf{X}}). \end{aligned}$$

Therefore, (12) holds for every \mathbf{A} with full row rank as long as the components of \mathbf{X} have the same differential entropy.

Since the above is true for every \mathbf{A} , the requirement of equal entropy on the components of \mathbf{X} is unnecessary due to the following. Suppose the entries of \mathbf{X} may take different

differential entropies. Let $\mathbf{D} = \text{diag}(d_1, \dots, d_n)$ where $d_i = \exp[-h(X_i)]$ so that $h(d_i X_i) = 0$. Then

$$\begin{aligned} h(\mathbf{A}\mathbf{X}) &= h\left(\left(\mathbf{A}\mathbf{D}^{-1}\right)\left(\mathbf{D}\mathbf{X}\right)\right) \\ &\geq h\left(\left(\mathbf{A}\mathbf{D}^{-1}\right)\left(\mathbf{D}\tilde{\mathbf{X}}\right)\right) \\ &= h(\mathbf{A}\tilde{\mathbf{X}}) \end{aligned} \quad (20)$$

where (20) is because the components of $\mathbf{D}\mathbf{X}$ are independent and have identical differential entropy (zero). Finally, the condition of full row rank on \mathbf{A} is unnecessary because if \mathbf{A} is row-rank-deficient, $h(\mathbf{A}\tilde{\mathbf{X}}) = -\infty$ and (12) is trivial. ■

C. Proof of Theorem 3 Using Theorem 2

In order to show that the generalized EPI implies the vector version of Lieb's inequality, we make use of the following result from [18], [20].

Lemma 1: Let \mathbf{U} be $m \times n$ with orthonormal rows. Then

$$\begin{aligned} \log \det (\mathbf{U} \text{diag}(p_1, \dots, p_n) \mathbf{U}^\top) \\ \geq \text{tr}\{\mathbf{U} \text{diag}(\log p_1, \dots, \log p_n) \mathbf{U}^\top\} \end{aligned}$$

if $p_1, \dots, p_n > 0$.

Lemma 1 is stated in [18] as a consequence of Theorem 3. In the Appendix we give a direct proof of the lemma.

Proof: [Theorem 2 \Rightarrow Theorem 3] Suppose (12) holds. Let \mathbf{U} be $m \times n$ with orthonormal rows. It is clear that

$$h(\mathbf{U}\mathbf{X}) \geq h(\mathbf{U}\tilde{\mathbf{X}}) = \frac{1}{2} \log \det (\mathbf{U}\mathbf{P}\mathbf{U}^\top)$$

where $\mathbf{P} = \text{diag}(p_1, \dots, p_n)$ with $p_i = \exp[2h(X_i)]$. Hence

$$h(\mathbf{U}\mathbf{X}) \geq \frac{1}{2} \text{tr}\{\mathbf{U} \text{diag}(\log p_1, \dots, \log p_n) \mathbf{U}^\top\}$$

by Lemma 1, which gives (14). ■

IV. A PROOF OF COSTA'S EPI

In the case where one of the random vectors in the sum is Gaussian, Shannon's EPI can be strengthened.

Theorem 4 (Costa [7]): For every random vector \mathbf{X} and Gaussian vector \mathbf{N} of identical dimension n , and $\alpha \in [0, 1]$,

$$\begin{aligned} \exp\left[\frac{2}{n} h(\mathbf{X} + \alpha\mathbf{N})\right] &\geq (1 - \alpha^2) \exp\left[\frac{2}{n} h(\mathbf{X})\right] \\ &\quad + \alpha^2 \exp\left[\frac{2}{n} h(\mathbf{X} + \mathbf{N})\right] \end{aligned} \quad (21)$$

Theorem 4 states that the entropy power of a random vector plus additive Gaussian noise is concave in the strength of the noise. The result finds its application in interference channels [7] and fading channels [21].

Costa's original proof in [7] shows that (21) is equivalent to that the function

$$f(a) = \exp\left[\frac{2}{n} h(a\mathbf{X} + \mathbf{N})\right] - \exp\left[\frac{2}{n} h(a\mathbf{X})\right] \quad (22)$$

being monotonic increasing for $a \geq 0$, which is then proved using rather involved analysis of its derivatives. It is clear that the monotonicity of (22) implies $f(1) \geq f(0)$, which is the classical EPI in this case, i.e., (5) with \mathbf{Y} replaced by \mathbf{N} .

The original proof of Costa's EPI is simplified in [22] and [20] based on a Fisher information inequality. In the following, we provide an alternative proof using a new information-estimation relationship in arbitrary additive noise channels with Gaussian input. The proof also applies a simple observation of the rate distortion function in such a channel.

A. Additive Channels with Gaussian Input

An alternative perspective of the channel model (1) is to exchange the roles of \mathbf{X} and \mathbf{N} , i.e., it is regarded as an arbitrary additive noise channel with Gaussian input. A corollary of (3) is the following relationship of the input-output mutual information of such a channel and its corresponding MMSE, which is an interesting observation on its own.

Theorem 5: As long as $P_{\mathbf{X}}$ is continuous,

$$\frac{d}{d\gamma} I(\mathbf{N}; \sqrt{\gamma}\mathbf{X} + \mathbf{N}) = \frac{1}{2\gamma} (\text{mmse}(\mathbf{N}|\sqrt{\gamma}\mathbf{X} + \mathbf{N}) - n) \quad (23)$$

where \mathbf{N} consists of n independent standard Gaussian entries.

Proof: First note that

$$\mathbf{Y} = \sqrt{\gamma}\mathbf{X} + \mathbf{N} = \sqrt{\gamma}\mathbb{E}\{\mathbf{X}|\mathbf{Y}\} + \mathbb{E}\{\mathbf{N}|\mathbf{Y}\}$$

and hence

$$\mathbf{N} - \mathbb{E}\{\mathbf{N}|\mathbf{Y}\} = \sqrt{\gamma}(\mathbb{E}\{\mathbf{X}|\mathbf{Y}\} - \mathbf{X}).$$

Also note that for every (\mathbf{X}, \mathbf{Y}) ,

$$\text{mmse}(\mathbf{X}|\mathbf{Y}) = \text{tr}\{\text{cov}(\mathbf{X}|\mathbf{Y})\} \quad (24)$$

where the error covariance matrix is defined by

$$\text{cov}(\mathbf{X}|\mathbf{Y}) = \mathbb{E}\{(\mathbf{X} - \mathbb{E}\{\mathbf{X}|\mathbf{Y}\})(\mathbf{X} - \mathbb{E}\{\mathbf{X}|\mathbf{Y}\})^T\}.$$

Therefore,

$$\text{mmse}(\mathbf{N}|\sqrt{\gamma}\mathbf{X} + \mathbf{N}) = \gamma \text{mmse}(\mathbf{X}|\sqrt{\gamma}\mathbf{X} + \mathbf{N}). \quad (25)$$

Furthermore,

$$\begin{aligned} I(\mathbf{N}; \sqrt{\gamma}\mathbf{X} + \mathbf{N}) &= h(\sqrt{\gamma}\mathbf{X} + \mathbf{N}) - h(\sqrt{\gamma}\mathbf{X}) \\ &= I(\mathbf{X}; \sqrt{\gamma}\mathbf{X} + \mathbf{N}) + h(\mathbf{N}) - h(\mathbf{X}) - \frac{n}{2} \log \gamma. \end{aligned} \quad (26)$$

Here we assume $h(\mathbf{X}) > -\infty$, since otherwise $I(\mathbf{N}; \sqrt{\gamma}\mathbf{X} + \mathbf{N})$ is infinite. Taking the derivative on both sides of (26) with respect to γ and invoking (3) and (25) yields (23). ■

B. A Proof of Costa's EPI

We apply Theorem 5 to show that $f(a) > 0$.

Proof: Since

$$I(\mathbf{N}; a\mathbf{X} + \mathbf{N}) = h(a\mathbf{X} + \mathbf{N}) - h(a\mathbf{X}),$$

we can rewrite (22) as

$$\begin{aligned} f(a) &= \exp\left[\frac{2}{n}h(a\mathbf{X})\right] \left(\exp\left[\frac{2}{n}I(\mathbf{N}; a\mathbf{X} + \mathbf{N})\right] - 1\right) \\ &= ca^2 \left(\exp\left[\frac{2}{n}I(\mathbf{N}; a\mathbf{X} + \mathbf{N})\right] - 1\right) \end{aligned}$$

where $c = \exp[2h(\mathbf{X})/n]$ is a constant independent of a . Taking the derivative and invoking Theorem 5 yields

$$\begin{aligned} \frac{f'(a)}{2ca} &= \exp\left[\frac{2}{n}I(\mathbf{N}; a\mathbf{X} + \mathbf{N})\right] - 1 \\ &\quad + \frac{a}{n} \exp\left[\frac{2}{n}I(\mathbf{N}; a\mathbf{X} + \mathbf{N})\right] \frac{d}{da} I(\mathbf{N}; a\mathbf{X} + \mathbf{N}) \\ &= \exp\left[\frac{2}{n}I(\mathbf{N}; a\mathbf{X} + \mathbf{N})\right] - 1 \\ &\quad + \exp\left[\frac{2}{n}I(\mathbf{N}; a\mathbf{X} + \mathbf{N})\right] \\ &\quad \times \left(\frac{1}{n} \text{mmse}(\mathbf{N}|a\mathbf{X} + \mathbf{N}) - 1\right) \\ &= \frac{1}{n} \text{mmse}(\mathbf{N}|a\mathbf{X} + \mathbf{N}) \\ &\quad \times \exp\left[\frac{2}{n}I(\mathbf{N}; a\mathbf{X} + \mathbf{N})\right] - 1. \end{aligned} \quad (27)$$

Recall the rate distortion function of an independent standard Gaussian source, which is given by [23]

$$R_D(d) = \frac{1}{2} \log \frac{1}{d}, \quad d \leq 1.$$

Consider also the following test channel

$$\mathbf{N} \longrightarrow \hat{\mathbf{N}} = \mathbb{E}\{\mathbf{N} | a\mathbf{X} + \mathbf{N}\}$$

the performance of which is inferior to the optimal rate distortion function. Thus,

$$\begin{aligned} \frac{1}{n} \text{mmse}(\mathbf{N}|a\mathbf{X} + \mathbf{N}) &\geq R_D^{-1}\left(\frac{1}{n} I(\mathbf{N}; \hat{\mathbf{N}})\right) \\ &\geq R_D^{-1}\left(\frac{1}{n} I(\mathbf{N}; a\mathbf{X} + \mathbf{N})\right) \\ &= \exp\left[-\frac{2}{n} I(\mathbf{N}; a\mathbf{X} + \mathbf{N})\right]. \end{aligned} \quad (28)$$

The monotonicity of $f(a)$ is established by plugging (28) into (27), which completes the proof. ■

Note that the two sides of (28) become equal if \mathbf{X} is replaced by a Gaussian vector \mathbf{X}_G of the same covariance. Thus the saddle-point property of mutual information [23]

$$I(\mathbf{N}; \mathbf{X} + \mathbf{N}) \geq I(\mathbf{N}; \mathbf{X}_G + \mathbf{N})$$

is obtained in light of

$$\text{mmse}(\mathbf{N}|\mathbf{X} + \mathbf{N}) \leq \text{mmse}(\mathbf{N}|\mathbf{X}_G + \mathbf{N}), \quad (29)$$

which holds because the right hand side of (29) is achievable by a (suboptimal) linear estimator of \mathbf{N} given $\mathbf{X} + \mathbf{N}$.

APPENDIX

We prove the inequality in Lemma 1 which is rewritten as

$$\log \det(\mathbf{U}\mathbf{P}\mathbf{U}^T) \geq \text{tr}\{\mathbf{U}(\log \mathbf{P})\mathbf{U}^T\} \quad (30)$$

where \mathbf{U} is $m \times n$ with orthonormal rows, $p_1, \dots, p_n > 0$, and for convenience we use shorthand notation

$$\mathbf{P} = \text{diag}(p_1, \dots, p_n)$$

and

$$\log \mathbf{P} = \text{diag}(\log p_1, \dots, \log p_n).$$

Proof: Let $\mathbf{R} = \mathbf{U}\mathbf{P}\mathbf{U}^\top$. Note that $\mathbf{U}\mathbf{U}^\top = \mathbf{I}$ but in general $\mathbf{U}^\top\mathbf{U} \neq \mathbf{I}$. However, since \mathbf{R} is positive definite, there exists an $m \times m$ orthogonal matrix \mathbf{V} (i.e., $\mathbf{V}\mathbf{V}^\top = \mathbf{V}^\top\mathbf{V} = \mathbf{I}$) and a diagonal matrix \mathbf{D} such that $\mathbf{R} = \mathbf{V}^\top\mathbf{D}\mathbf{V}$. In fact $\mathbf{D} = \mathbf{V}\mathbf{U}\mathbf{P}\mathbf{U}^\top\mathbf{V}^\top$, and its diagonal elements are expressed as

$$d_j = \sum_{i=1}^n w_{ji}^2 p_i,$$

where w_{ji} represents the (j, i) -th entry of the matrix product $\mathbf{V}\mathbf{U}$, which contains orthonormal row vectors because $\mathbf{V}\mathbf{U}\mathbf{U}^\top\mathbf{V}^\top = \mathbf{I}$. Therefore

$$\begin{aligned} \log \det \mathbf{R} &= \log \det \mathbf{D} \\ &= \sum_{j=1}^m \log d_j \\ &= \sum_{j=1}^m \log \sum_{i=1}^n w_{ji}^2 p_i. \end{aligned}$$

By concavity of the logarithm,

$$\begin{aligned} \log \det \mathbf{R} &\geq \sum_{j=1}^m \sum_{i=1}^n w_{ji}^2 \log p_i \\ &= \text{tr}\{\mathbf{V}\mathbf{U}(\log \mathbf{P})\mathbf{U}^\top\mathbf{V}^\top\} \\ &= \text{tr}\{\mathbf{U}(\log \mathbf{P})\mathbf{U}^\top\} \end{aligned}$$

since $[w_{j1}, \dots, w_{jn}]$ is unitary for every j . ■

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, pp. 1261–1282, Apr. 2005.
- [2] A. J. Stam, "Some inequalities satisfied by the quantities of information of Fisher and Shannon," *Information and Control*, vol. 2, pp. 101–112, 1959.
- [3] T. E. Duncan, "Likelihood functions for stochastic signals in white noise," *Information and Control*, vol. 16, pp. 303–310, 1970.
- [4] J. Binia, "Divergence and minimum mean-square error in continuous-time additive white Gaussian noise channels," *IEEE Trans. Inform. Theory*, vol. 52, pp. 1160–1163, Mar. 2006.
- [5] 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, Jan. 2006.
- [6] S. Verdú and D. Guo, "A simple proof of the entropy power inequality," *IEEE Trans. Inform. Theory*, vol. 52, pp. 2165–2166, May 2006.
- [7] M. H. M. Costa, "A new entropy power inequality," *IEEE Trans. Inform. Theory*, vol. 31, pp. 751–760, Nov. 1985.
- [8] R. Zamir and M. Feder, "A generalization of the entropy power inequality with applications," *IEEE Trans. Inform. Theory*, vol. 39, pp. 1723–1728, Sept. 1993.
- [9] E. H. Lieb, "Proof of an entropy conjecture of Wehrl," *Commun. Math. Phys.*, vol. 62, no. 1, pp. 35–41, 1978.
- [10] A. M. Tulino and S. Verdú, "Monotonic decrease of the non-Gaussianness of the sum of independent random variables: A simple proof," *IEEE Trans. Inform. Theory*, 2006, to appear.
- [11] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423 and 623–656, July and Oct. 1948.
- [12] P. P. Bergmans, "A simple converse for broadcast channels with additive white Gaussian noise," *IEEE Trans. Inform. Theory*, vol. 20, pp. 279–280, Mar. 1974.
- [13] Y. Oohama, "Gaussian multiterminal source coding," *IEEE Trans. Inform. Theory*, vol. 43, pp. 1912–1923, Nov. 1997.
- [14] Y. Oohama, "The rate-distortion function for the quadratic Gaussian CEO problem," *IEEE Trans. Inform. Theory*, vol. 44, pp. 1057–1070, May 1998.
- [15] A. Dembo, T. M. Cover, and J. A. Thomas, "Information theoretic inequalities," *IEEE Trans. Inform. Theory*, vol. 37, pp. 1501–1518, Nov. 1991.
- [16] R. Zamir and M. Feder, "Rate-distortion performance in coding bandlimited sources by sampling and dithered quantization," *IEEE Trans. Inform. Theory*, vol. 41, pp. 141–154, Jan. 1995.
- [17] R. Zamir, "A proof of the Fisher information inequality via a data processing theorem," *IEEE Trans. Inform. Theory*, vol. 44, pp. 1246–1250, May 1998.
- [18] R. Zamir and M. Feder, "A generalization of information theoretic inequalities to linear transformations of independent vector," in *Proceedings of the 6-th Joint Swedish-Russian International Workshop on Information Theory*, pp. 254–258, Mölre, Sweden, Aug. 1993.
- [19] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 1985.
- [20] C. Villani, "A short proof of the 'concavity of entropy power'," *IEEE Trans. Inform. Theory*, vol. 46, pp. 1695–1696, July 2000.
- [21] A. Lapidoth and S. M. Moser, "Capacity bounds via duality with applications to multiple-antenna systems on flat-fading channels," *IEEE Trans. Inform. Theory*, vol. Oct., pp. 2426–2467, Oct. 2003.
- [22] A. Dembo, "Simple proof on the concavity of the entropy power with respect to added Gaussian noise," *IEEE Trans. Inform. Theory*, vol. 35, pp. 887–888, July 1989.
- [23] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.